IntechOpen

Artificial Intelligence
Emerging Trends and Applications

Edited by Marco Antonio Aceves-Fernandez

ARTIFICIAL INTELLIGENCE - EMERGING TRENDS AND APPLICATIONS

Edited by **Marco Antonio Aceves-Fernandez**

Artificial Intelligence - Emerging Trends and Applications

Contributors

Vo Ngoc Phu, Tran Vo Thi Ngoc, Yasir Ali, Spyros Makridakis, Antonis Polemitis, George Giaglis, Soulla Louca, Josep Llorca, Matej Rojc, Zdravko Kačič, Izidor Mlakar, Arturo Lezama-León, José Angel Montes-Olguín, Evangelina Lezama-León, José Juan Zarate-Corona, Juan Ángel Rosales-Alba, Ma. De La Luz Carrillo González, Noureddine Bouhmala, Saejoon Kim, Soong Kim, Chaomin Luo, Gene Eu (Ching Yuh) Eu Jan, Zhenzhong Chu, Li, Francisco Lopez-Orozco, Rogelio Florencia-Juarez, David Kaufman, Dov Schafer, Jun Dong, Lin-Peng Jin, Mohd Najib Mohd Salleh, Noureen Talpur, Lj Gong, Sugiono Sugiono, Juana Canul-Reich, Rafael Rivera-Lopez, Ahmadi Azqhandi, Tony McCaffrey, Pingzhao Hu, Yang Wang, Md. Mohaiminul Islam, Zeinab Hajjar, Shokoufe Tayyebi, Mohammad Hosein Eghbal Ahmadi, Masakazu Sato, Kaori Koga, Tomoyuki Fujii, Yutaka Osuga, Evon M.O. Abu-Taieh, Auhood Abd. Alfaries, Nabeel M. Zanoon, Issam H. Al Hadid, Alia M. Abu-Tayeh

Notice

We are IntechOpen,
the world's leading publisher of
Open Access books
Built by scientists, for scientists

3,550+
Open access books available

112,000+
International authors and editors

115M+
Downloads

151
Countries delivered to

Our authors are among the
Top 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Meet the editor

Dr. Marco A. Aceves-Fernandez obtained his BSc degree (Eng) in Telematics from the Universidad de Colima, Mexico. He obtained both his MSc and PhD degrees from the University of Liverpool, England, in the field of Intelligent Systems. He is a full professor at the Universidad Autonoma de Queretaro, Mexico. He is being recognized as a member of the National System of Researchers (SNI) since 2009. He has published more than 80 research papers as well as numerous book chapters and congress papers. He has contributed to more than 20 funded research projects both academic and industrial in the area of artificial intelligence ranging from environmental, biomedical, automotive, aviation, consumer, and robotics fields, among other applications. He is also an honorary president at the Mexican Association of Embedded Systems (AMESE), a senior member of the IEEE, and a board member of many institutions and associations. His research interests include intelligent and embedded systems.

Contents

Preface

From the earliest humans to populate the earth, we have gradually tried to understand and control the world around us. From trying to understand such phenomena, humans started to make predictions to various extents. For instance, humans started to make predictions about motions of the planets, eclipses, cycles of rainfall, or periodicity of certain diseases. However, in the last few decades, the complexity of the predictions needed to be carried out has reached the limits of our abilities to predict.

Luckily, the dawn or electronic computers are deeply increasing its abilities to predict the nature, although the problems we are facing now are far more complex than the problems we faced a century ago.

The ability of these machines to demonstrate advanced cognitive skills in taking decisions, learn and perceive the environment, predict certain behavior, and process written or spoken languages, among other skills, makes this discipline of paramount importance in today's world.

The wide range of applications in artificial intelligence nowadays is shown in a variety of applications compiled in this book. This book is organized as follows:

1. Artificial Intelligence Fundamentals
2. Engineering and Technology Applications
3. Life and Health Sciences

I hope that this work is of interest to students and researchers alike as I did my best to comprise quality research contributions with a number of different applications.

Dr. Marco A. Aceves-Fernandez, PhD
Universidad Autonoma de Queretaro, Mexico

Artificial Intelligence Fundamentals

Biologically Inspired Intelligence with Applications on Robot Navigation

Chaomin Luo, Gene Eu Jan, Zhenzhong Chu and
Xinde Li

Additional information is available at the end of the chapter

Abstract

Biologically inspired intelligence technique, an important embranchment of series on computational intelligence, plays a crucial role for robotics. The autonomous robot and vehicle industry has had an immense impact on our economy and society and this trend will continue with biologically inspired neural network techniques. In this chapter, multiple robots cooperate to achieve a common coverage goal efficiently, which can improve the work capacity, share the coverage tasks, and reduce the completion time by a biologically inspired intelligence technique, is addressed. In many real-world applications, the coverage task has to be completed without any prior knowledge of the environment. In this chapter, a neural dynamics approach is proposed for complete area coverage by multiple robots. A bio-inspired neural network is designed to model the dynamic environment and to guide a team of robots for the coverage task. The dynamics of each neuron in the topologically organized neural network is characterized by a shunting neural equation. Each mobile robot treats the other robots as moving obstacles. Each robot path is autonomously generated from the dynamic activity landscape of the neural network and the previous robot position. The proposed model algorithm is computationally simple. The feasibility is validated by four simulation studies.

Keywords: biologically inspired intelligence, real-time motion planning, navigation and mapping

1. Introduction

Biologically inspired intelligence technique, an important embranchment of series on computational intelligence, plays a crucial role for robotics. The autonomous robot and vehicle

industry has had an immense impact on our economy and society and this trend will continue with biologically inspired neural network techniques. Biologically inspired intelligence, such as biologically inspired neural networks (BNNs), is about learning from nature, which can be applied to the real world robot and vehicle systems. Recently, the research and development of bio-inspired systems for robotic applications is increasingly expanding worldwide. Biologically inspired algorithms contain emerging subtopics such as bio-inspired neural network algorithms, brain-inspired neural networks, swam intelligence with BNN, ant colony optimization algorithms (ACO) with BNN, bee colony optimization algorithms (BCO), particle swarm optimization with BNN, immune systems with BNN, and biologically inspired evolutionary optimization and algorithms. Additionally, it is decomposed of computational aspects of bio-inspired systems such as machine vision, pattern recognition for robot and vehicle systems, motion control, motion planning, movement control, sensor-motor coordination, and learning in biological systems for robot and vehicle systems.

One of the applications of biologically inspired intelligence techniques on the robot navigation is complete area coverage navigation of autonomous mobile robots. Complete area coverage (CAC) is an essential issue in mobile robots, which requires the robot path to pass through every area in the workspace. Many robotic applications require CAC, e.g., cleaning robots [1–6], vacuum robots [7], painter robots, autonomous underwater covering vehicles, de-mining robots [8], land mine detectors [9], lawn mowers, automated harvesters [10], agricultural crop harvesting equipment [11], and window cleaners [12]. CAC can be completed by a single robot or multiple robots.

Nowadays, cooperative coverage by a multiple robot system is becoming increasingly important. The cooperative area coverage by multiple robots can improve the efficiency and complete the work more quickly than a single robot. These robots may share the coverage tasks and thus reduce the time to complete the coverage task. Additionally, if one of the robots fails, the rest will fulfill the missions, therefore, the coverage by robots is able to improve reliability and robustness. For instance, in de-mining applications, coverage reliability, an important factor, is enhanced by using cooperative multirobots. In some cleaning applications, the workspace (e.g., a stadium) needs to be cleaned in a limited amount of time. Thus, it requires multiple robots to work in a cooperative manner.

Multi-robot coverage has been extensively studied using various models. Depending on whether a map is required for the multirobots, the coverage models may be categorized as off-line and on-line algorithms [13]. Off-line algorithms require a map of workspace for robots (e.g. [14–16]), while on-line algorithms do not need an environmental map (e.g. [17–21]). Previous research on area coverage may be classified into cell-decomposition-based model (e.g. [15, 16, 21–24]), spanning-tree-based approach (e.g. [14, 18, 23–25]), behaviour-based model (e.g. [26–30]), graph-based model (e.g. [20, 31, 32]), depth first search approach (e.g. [18, 19]), Frontier-based model (e.g. [33–37]), and others (e.g. [38–41]).

Many multi-robot coverage algorithms are based on cell decomposition. Cell decomposition methods break continuous space into a finite set of cells. After this decomposition, a connectivity graph is constructed according to the adjacency relationships between the cells. From this connectivity graph, a continuous path can be determined by simply following adjacent

free cells from the initial point to the goal point. Oh et al. [6] developed a triangular cell decomposition method for unknown environments for CAC. This method combines triangular cell decomposition, a template-based approach, and a wall following navigation algorithm for CAC. It can only deal with a single robot CAC. Wagner et al. [21] proposed multi-robot coverage algorithms to explore and cover an unknown environment by approximate cell decomposition approach. The group of multiple robots has limited sensors and no explicit communication. Kurabayashi et al. [15, 16] proposed an exact cell decomposition off-line coverage algorithm for multiple cleaning robots using a Voronoi diagram and boustrophedon approach, where a cost function is defined to obtain a near-optimal solution of the collective coverage task. The approach can avoid overlaps of sweeping areas. The algorithm needs to consider in advance the knowledge of the workspace with known obstacles.

Gabriely and Rimon [42] suggested a spanning tree coverage approach with single robot. It divides up the workspace into discrete cells and generates spanning tree of graph induced by the cells. The robot is able to cover every point precisely once and travel an optimal path in a grid-like representation in the workspace that achieves complete area coverage. Hazon and Kaminka [14] developed a complete and robust multi-robot spanning-tree coverage (MSTC) algorithm based on approximate cell decomposition. Afterwards, the coverage efficiency was improved by a multi-robot forest coverage (MFC) algorithm of approximate cell decomposition proposed by Zheng et al. [23]. Hazon et al. [18] successfully extended the spanning tree work with single robot of Gabriely and Rimon [42] and their off-line work [14] into on-line multi-robot coverage and improved the coverage efficiency.

Behaviour-based strategy for a multi-robot system employs relatively little internal variable states to model the environment and makes fewer assumptions about their environment, thus it is more robust. The cooperative coverage by multirobots with the feature of reactive planning model is implemented by designing individual and team behaviours. Jung et al. [29] combined the advantages of spatial and topological map representations of the environment in a behaviour-based framework for cooperative cleaning of multiple robots. The cooperative multi-robot coverage missions can be accomplished by unifying paths in navigation, cooperation, communication and reactive behaviours [43]. The more detailed explanation of the implementation of the architecture for behaviour-based agents for cooperative multiple cleaning robots are given in [30]. Balch and Arkin [26] proposed a behaviour-based multi-robot approach for coverage, in which the robots are developed with various goal-oriented behaviours for navigation. Recently, Fang et al. [27] used a behaviour-based coverage approach for multirobots to efficiently define the region in which an optimal solution can be found in unknown environments. Most recently, an idea of a leader robot and other follower robots for planning path and controlling robots was proposed using the behaviour-based model [44].

The idea of building up a graph of environment is used for multi-robot coverage. One workspace is decomposed into subregions called cells and therefore a graph may be constructed. The underlying idea of graph-based approach is that multiple robots traverse every edge of the graph to achieve the cooperative coverage. Wagner et al. [20, 31] proposed an approximate cellular decomposition approach for multi-robot coverage to decompose the environment. They employ a dirt grid on the floor for communication among robots. The

robots communicate with each other by leaving traces. A graph is built up for representation of the workspace to be covered. Each edge is assigned to two "smell labels". If an edge is traversed, it is marked by a fresh trace of odour. Recently, to benefit from the graph-based approach, Williams and Burdick [32] constructed a graph for multi-robot navigation. An improved graph representation of the task is applied for boundary coverage problem and a graph algorithm is developed for the boundary coverage problem.

A robot can obtain updated knowledge for its environmental map if it moves to a frontier since frontiers are this type of areas that are on the boundary between open space and uncovered space. With the movement of single robot or multiple robots to successive frontiers, the robots can obtain sufficient information to build up and update the maps for coverage mission. Yamauchi [36, 37] adopted a Frontier-based coverage approach, which leads each robot to the closest unknown region, represented by frontier between the free and unknown workspace to produce a robust autonomous cooperative coverage strategy. The technique builds a global map of the environment, which is analysed to locate the frontiers around the robot and environments. Recently, Burgard et al. [33] and Ko et al. [35] developed algorithms to compute utilities of frontier cells to cover different areas of the unknown environments.

Some previous work combines several approaches to take an advantage of different benefits. For instance, Rekleitis et al. [19] split a terrain by an exact cell decomposition method and a tree was built with each subregion as a node. A centralized depth first search (DFS) algorithm is employed for robots to traverse the unknown region and thus the entire areas are explored and covered. Hazon et al. [18] recently extended the spanning tree work with single robot [42] and their own off-line work [14] into on-line multi-robot coverage. Their spanning trees are constructed by a DFS-like procedure. The effective, robust, and complete multi-robot coverage is implemented. Gossage et al. [34] combined the advantages, in order to obtain robust cooperation, of local Frontier-based approach and global graph-based representation of unknown environments for the cooperation of multirobot. Most recently, Zavlanos and Pappas [45] combined a distributed multi-destination potential field approach and a dynamic assignment algorithm for coverage motion planning of multiple robots.

In some other methods, multirobots collect the incoming sensor information of every single robot in a team to cooperatively perform coverage in unknown environments. In most cases, cell decomposition method is used to split terrain and ensures complete coverage [46]. Butler [38, 39] proposed a distributed cooperative coverage algorithm for multirobot, which performs independently on each robot in a team with a rectilinear environment. The algorithm employs only intrinsic contact sensing knowledge to determine the boundaries of the environment. Recently, Boonpinon and Sudsang [47] developed a multi-robot mapping and area coverage approach using a centroidal Voronoi diagram where a team of robots exchange limited sensory information by explicit communication. Latimer et al. [40] and Rekleitis et al. [41] proposed a coordinated multi-robot approach for a coverage mission while the workspace is typically broken down into distinct regions by Boustrophedon decomposition and different region is covered by robots with back-and-forth motions. Most recently, Schwager et al. [48] suggested a near optimal sensing configuration for coverage by a group of robots by learning the distribution of the sensory information in environments.

Although there have been many studies on multi-robot coverage and most attempt to improve completeness, very few existing coverage algorithms focus on robustness. The MSTC algorithm

proposed by Hazon and Kaminka [14] is robust and complete. The robots with this algorithm would cover cells more than once. Neural network methods have been broadly applied to robot motion planning, control and coverage (e.g. [2], [49, 50]). However, most of them deal with single robot for coverage (e.g. [2, 5, 6]). Some neural network models require learning procedures (e.g. [51, 52], which are computationally expensive and difficult to achieve CAC in real time.

In this chapter, a neural dynamics approach is proposed for multi-robot area coverage applications. Mobile robots have no collisions among themselves and can avoid obstacles and cooperatively work together to improve cleaning productivity effectively. The proposed approach is capable of performing CAC for multirobots, autonomously without any human intervention. Each robot treats the other robots as moving obstacles. The neural activity landscape of each robot is able to guide the robot to follow a reasonable path and to cooperate with other robots. In this paper, the real-time path is generated by employing a neural network algorithm, without either any prior knowledge of the environment or any pre-defined template. No learning procedures are required in the proposed algorithm. The advantage of such CAC strategy using the proposed neural networks is that the robots do not repeat the previous covered locations. The simulation studies demonstrate that the robustness and fault-tolerant can be ensured if one of the robots fails. It is computationally simple and flexible to implement the proposed algorithm on autonomous CAC as no learning procedures and no templates are required. The dynamics of each neuron is characterized by a shunting equation or an additive equation derived from the membrane model for a biological neural system [53]. There are only local lateral connections among neurons. Thus, the computational complexity depends on the neural network size. The varying environment is represented by the dynamic activity landscape of the neural network. Multiple robots share the environmental information, which is collected from the sensors mounted in the workspace, and all the sensors on individual robots.

The effective, complete, and robust cooperate area coverage is achieved by the proposed neural dynamics model. The term "cooperate" is in the sense that multiple robots can work together to achieve a common coverage mission more efficiently and more quickly. "Robust" is in the sense that the multi-robot system does not fatally failed or is not wholly affected by a single robot failure. In this chapter, cleaning robot is used as an example, but the method is applicable for any CAC applications.

The rest of the chapter is organized as follows. In Section 2, the biological inspiration, model algorithm, and stability analysis of this neural dynamics-based approach to real-time collision-free CAC by multirobot are addressed. Several simulation studies aimed to demonstrate the completeness, robustness, and effectiveness of the proposed model for CAC are performed and described in Section 3. Finally, several important properties of the proposed model with CAC are concluded in Section 4.

2. The proposed model

In this section, the originality of the proposed neural network approach to real-time CAC for multiple mobile robots will be briefly introduced. Then, the fundamental concept and model algorithm of the proposed approach will be presented.

2.1. Biological inspiration

In 1952, Hodgkin and Huxley [53] proposed a computational model for a patch of membrane in a biological neural system using electrical circuit elements. In this model, the dynamics of voltage across the membrane, V_m, is described using the state equation technique as

$$C_m \frac{dV_m}{dt} = - \left(E_p + V_m \right) g_p + \left(E_{Na} - V_m \right) g_{Na}$$
$$- (E_K + V_m) g_K$$

(1)

where C_m is the membrane capacitance, E_K, E_{Na}, and E_p are the Nernst potentials (saturation potentials) for potassium ions, sodium ions, and the passive leak current in the membrane, respectively. Parameters g_K, g_{Na} and g_p represent the conductances of potassium, sodium, and passive channels, respectively. This model provides the foundation of the shunting model and leads to a number of model variations and applications [54].

By setting $C_m = 1$ and substituting $x_i = E_p + V_m$, $A = g_p$, $B = E_{Na} + E_p$, $D = E_k - E_p$, $S_i^e = g_{Na}$ and $S_i^i = g_K$ in Eq. (1), a shunting equation is obtained

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i)S_i^e(t) - (D + x_i)S_i^i(t)$$

(2)

where x_i is the neural activity (membrane potential) of the ith neuron. Parameters A, B, and D are nonnegative constants representing the passive decay rate, the upper and lower bounds of the neural activity, respectively. Variables S_i^e and S_i^i are the excitatory and inhibitory inputs to the neuron. This shunting model was first proposed by Grossberg to understand the real-time adaptive behaviour of individuals to complex and dynamic environmental contingencies and has many applications in visual perception, sensory motor control, and many other areas [54]. Research on biologically inspired robots has currently received attention [24, 55, 56].

2.2. Model algorithm

The fundamental concept of the proposed model is to develop a neural network architecture, whose dynamic neural activity landscape represents the dynamically varying environment. By properly defining the external inputs from the varying environment and internal neural connections, the uncovered areas and obstacles are guaranteed to stay at the peak and the valley of the activity landscape of the neural network, respectively. The uncovered areas globally attract the robot in the whole state space through neural activity propagation, while the obstacles have only local effect in a small region to avoid collisions. The real-time collision-free robot area coverage is accomplished based on the dynamic activity landscape of the neural network, the previous robot position and the other robot positions, to guarantee all locations to be covered and the robots to travel along smooth, continuous paths with less turning.

The proposed topologically organized model is expressed in a 2D Cartesian workspace $\mathcal{W}$ of the cleaning robots. The position of the ith neuron in the state space S of the neural network,

denoted by a vector $q_i \in R^2$, uniquely represents a position in $\mathcal{W}$. In the proposed model, the excitatory input results from the uncovered locations and the lateral neural connections, while the inhibitory input results from the obstacles only. Each neuron has local lateral connections to its neighbouring neurons that constitute a subset R_i in $\mathcal{S}$. The subset R_i is called the receptive field of the ith neuron in neurophysiology. The neuron responds only to the stimulus within its receptive field. Thus, the dynamics of the ith neuron in the neural network is characterized by a shunting equation as

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i)\left([I_i]^+ + \sum_{j=1}^{k} w_{ij}[x_j]^+\right)$$
$$- (D + x_i)[I_i]^-$$

(3)

where k is the number of neural connections of the ith neuron to its neighbouring neurons within the receptive field R_i. The external input I_i to the ith neuron at Position (m,n) is defined as:

$$I_i(m, n) = \begin{cases} E & \text{if it is uncovered} \\ 0 & \text{if it is covered} \\ -E & \text{if it is obstacle (another robot)} \end{cases}$$

(4)

where $E \gg B$ is a very large positive constant. The terms $[I_i]^+ + \sum_{j=1}^{n} w_{ij}[x_j]^+$ and $[I_i]^-$ are the excitatory and inhibitory inputs, S_i^e and S_i^i in Eq. (2), respectively. Function $[a]^+$ is a linear-above-threshold function defined as $[a]^+ = \max\{a, 0\}$ and the nonlinear function $[a]^-$ is defined as $[a]^- = \max\{-a, 0\}$. The connection weight w_{ij} from the ith neuron to the jth neuron is given by $w_{ij} = f\left(|q_i - q_j|\right)$, where $|q_i - q_j|$ represents the Euclidean distance between vectors q_i and q_j in the state space, and $f(a)$ is a monotonically decreasing function, such as a function defined as $f(a) = \mu/a$, if $0 \leq a < r_0$; $f(a) = 0$, if $a \geq r_0$, where μ and r_0 are positive constants. Therefore, each neuron has only local lateral connections in a small region $[0, r_0]$. It is obvious that the weight w_{ij} is symmetric, i.e., $w_{ij} = w_{ji}$. A schematic diagram of the neural network in 2D with three-layer (r = 1, 2, and 3) neighbouring neurons with regard to the central neuron C(m,n) is shown in **Figure 1**, where r_0 is chosen as $r_0 = 2$ and r is the number of circles enclosing the central neuron C(m,n). The receptive field of the ith neuron is represented by a circle with a radius of r_0. The ith neuron has only eight lateral connections to its neighbouring neurons that are within its receptive field. The 2D Cartesian workspace in the proposed approach is discretized into squares. The diagonal length of each discrete area is equal to the robot coverage radius that is the size of robot effector or footprint [2]. Each position (grid) uses a number to represent its environmental information. The neurons are placed uniformly on the space to represent covered positions, uncovered positions, and obstacles. In this algorithm, it is necessary to have a flag, denoted by $I_i(m, n)$ in Eq. (4), for a neuron at Position (m, n) to indicate its status that is uncovered, covered, obstacle, or moving obstacle (another robot is regarded as a moving obstacle). This flag may be technically obtained by the sensor information from the current map. A topologically organized discrete map is employed to represent the workspace; each grid uses a number (flag) to represent its environmental information
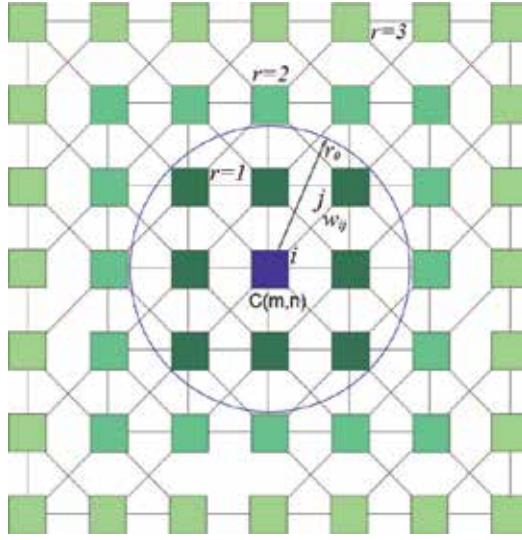
Figure 1. The architecture of a 2D neural network with three-layer (r = 1, 2, and 3) neighbouring neurons with regard to the *i*th neuron C(m,n). The *i*th neuron has only eight lateral connections to its neighbouring neurons that are within its receptive field.

(state). This approach only needs current map, instead of the prior map information. The boundary of the workspace is assumed to be known that can be obtained by wall-following algorithm [57].

The proposed network characterized by Eq. (3) ensures that the positive neural activity can propagate to all the state space, but the negative activity only stays locally. Therefore, the uncovered areas globally attract the robot, while the obstacles only locally prevent the robot from collisions. The positions of the uncovered areas and obstacles may vary over time, e.g., there are moving obstacles (other robots); the covered areas become uncovered again. The activity landscape of the neural network dynamically changes due to the varying external inputs from the uncovered areas and obstacles and the internal activity propagation among neurons. For energy and time efficiency, the robot should travel a shortest path (with least re-visited locations) and make least turning of moving directions. The robot path is generated from the dynamic activity landscape and the previous robot position to avoid least navigation direction changes. For a given current robot position in S (i.e., a position in $\mathcal{W}$), denoted by p_c, the next robot position p_n (also called "command position") is obtained by

$$p_n \quad \Leftarrow \quad x_{p_n} = \max\left\{x_j + cy_j, j = 1, 2, \cdots, k\right\} \tag{5}$$

where c is a positive constant, k is the number of neighbouring neurons of the p_cth neuron, i.e., all the possible next positions of the current position p_c. Variable x_j is the neural activity of the jth neuron, y_j is a monotonically increasing function of the difference between the current to next robot moving directions, which can be defined as a function of the previous position p_p, the current position p_c, and the possible next position p_j, e.g., a function defined as

$$y_j = 1 - \frac{\Delta\theta_j}{\pi} \tag{6}$$

where $\Delta\theta_j \in [0, \pi]$ is the turning angle between the current moving direction and next moving direction, e.g., if the robot moves straight, $\Delta\theta_j = 0$; if goes backward, $\Delta\theta_j = \pi$. Thus, $\Delta\theta_j$ can be given as: $\Delta\theta_j = |\theta_j - \theta_c| = |\text{atan2}\left(y_{p_j} - y_{p_c}, x_{p_j} - x_{p_c}\right) - \text{atan2}\left(y_{p_c} - y_{p_c}, x_{p_c} - x_{p_p}\right)|$. After the current position reaches its next position, the next position becomes a new current position (if the found next position is the same as the current position, the robot stays there without any movement). The current robot position adaptively changes according to the varying environment.

In a multi-robot system, if there exist two robots that work together to sweep in a workspace, it may be viewed as a multiple neural network system [58]. Each robot needs one neural network and all robots share the same workspace information. Each robot treats the other robots as moving obstacles recognized by the sensor information so that they can avoid collisions and cooperatively work together. Each robot has its own neural network which is updated dynamically on the positions of the robot. The environmental knowledge is also dynamically updated, which is sensed by robots via the sensor information. Every position is flagged by a number in Eq. (4). Once one of the multiple robots moves to Position (m, n), the position should be marked as the external input $I_i(m, n) = -E$.

The proposed neural network is a stable system. The neural activity x_i is bounded in the finite interval $[-D, B]$ [54]. The stability and convergence of the present shunting neural network model can also be rigorously proved using a Lyapunov stability theory. By introducing new variables and performing variable substitutions, Eqs. (2) or (3) can be written as

$$\frac{dz_i}{dt} = a_i(z_i)\left(b_i(z_i) - \sum_{j=1}^{N} c_{ij}d_j(z_j)\right) \tag{7}$$

which is Grossberg's general form [54]. It can be proved that Eqs. (2) or (3) satisfies all the three stability conditions required by the Grossberg's general form [54]. The rigorous proof of the stability and convergence of Eq. (7) can be found in [59]. The dynamics of the neural network is guaranteed to converge to an equilibrium state of the system. Eq. (3) combined with the previous robot position ensures to generate complete coverage path. At the beginning, when $t = 1$, the neural activity of all neurons is set to zero. The state of the workspace varies in terms of the dynamics of the neural network described by (3) due to the influence of external inputs. The planned motion ends when the network reaches a steady state.

It is inevitable that multiple cleaning robots have to deal with a deadlock situation in real-world applications. When a cleaning robot arrives in a deadlock situation, i.e., all the neighboring positions are either obstacle or covered locations, all the neural activities of its neighboring locations are not larger than the activity at the current location, because its neighboring locations receive either negative external input (obstacles) or no external input (covered locations), and all the covered neighboring locations passed a longer decay time as they were covered earlier than the current location [see Eq. (3)]. In the proposed model, the neural activity at the deadlock

location will quickly decay to zero due to the passive decay term $-Ax_i$ in Eq. (3). Meanwhile, due to the lateral excitatory connections among neurons, the positive neural activity from the uncovered locations in the workspace will propagate toward the current robot location through neural activity propagation (please see [60]). Therefore, the robot is able to find a smooth path from the current deadlock location directly to an uncovered location. The robot continues its cleaning task until all the locations in the workspace become covered. Thus, the proposed model is capable of achieving complete coverage path planning with deadlock avoidance. The multiple robots will not be trapped in deadlock situations.

The complexity is squarely proportional to the degree of discretization. There are only local connections among neurons. If the workspace is an $N \times N$ square and the number of neurons is $M = N \times N$, N^2 neurons are required and there are $8N^2$ neural connections. If the workspace is a rectangle, the number of neurons required is equal to $M = N_x \times N_y$, where N_x and N_y are the discretized size of the Cartesian workspace. Each neuron has maximal eight local connections. As a result, the total neural connections are $8M$. The computational complexity of the proposed algorithm is $O(N^2)$.

2.3. Implementation issues

There have been many studies on the CAC implementation of mobile robots using various approaches [46, 61–63]. The selection of on-board sensors is equally important as the development of CAC navigation algorithm itself. The performance of multirobots will depend on both CAC navigation algorithm and the placement of on-board sensors. Appropriate amount of sensors is key for the multi-robot system. Use of excessive amount of sensors will cause the increase of cost of robots [62, 64]. Each robot in this multi-robot system has the same configuration. The basic configuration of the robot consists of CPU, memory, sensors, DC motors, wheels, brush or dustpan, and an on-board power supply (e.g., rechargeable battery) [61].

- Sensors equipped on the robots are assumed to be imaging sensors (e.g., camera and position sensor) and rang sensors (e.g., infrared sensors, sonar, ultrasonic sensors or small radar). With on-board sensors, mobile robots can construct current environmental map around the robots [2, 65]. Ultrasonic sensors are employed for range measures due to their simplicity, flexibility, adaptability, low cost, and robustness. The interpretation of the sonar readings is helpful to build an environmental map. The small angular resolution of infrared sensors makes it suitable for CAC with back-and-forth motion. Infrared sensors are capable of reducing angular uncertainty caused by accumulated error of dead reckoning [63]. In addition, several cameras are suspended from the ceiling of workspace that provides global environmental information to robots [66].

- Wheels, DC motors, and discretized environments: each robot is driven by two DC geared motors with two wheels installed to the gear axis. The 2D Cartesian workspace is discretized into squares as most other CAC models. The diagonal length of each discrete grid is equal to a robot sweeping radius, which is the size of the robot effector or footprint (**Figure 2**). The robot in this paper is assumed to be round in shape and a square is embedded in this round as robot's body. A robot sweeping range, which is the size of the
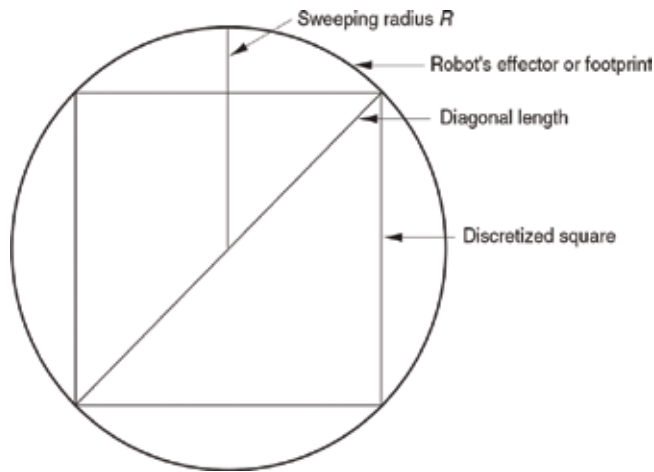
Figure 2. Discretized square enclosed by sweeping area.

robot effector, is proportional to the size of each square. The size of a robot is slightly larger than that of each square. Two wheels driven by DC motors are mounted on shafts with brush or dustpan. The wheels are axle mounted and supported using two sealed ball bearings [see **Figure 3(a)**]. Note that ball bearings do not appear in **Figure 3**. New design is assumed that the wheels have capability to rotate at any directions on the floor make the robots flexibly turn in the workspace. For example, **Figure 3(b)** shows that the robot turns clockwise and counterclockwise 45°. Therefore, sweeping an area can be achieved by traversing the center of that area represented by a rectangular cell. It is assumed that a discrete location represented by that squared cell is regarded to be covered once a robot visits the discrete cell. If a cleaning robot covers every discrete cell, the robot path is considered as a CAC in the workspace.
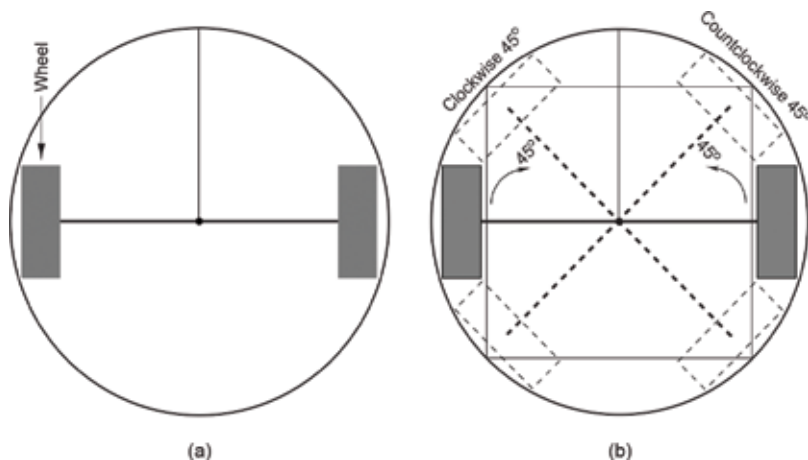
Figure 3. Two wheels driven by DC motors are mounted to the robot. (a) The robot; (b) the robot rotates clockwise and counter-clockwise 45°.

3. Simulation and comparison experiments

The proposed model for CAC is performed by simulation experiments in C++ in this section. The approach is capable of performing CAC for multiple cleaning robots, autonomously without human operation. Two or more mobile robots can cooperatively sweep in indoor environments so as to improve the work productivity. Each robot is required to not only clean its own floor but also cooperate with others to perform the cleaning tasks. In this section, the proposed approach is first applied to multiple robots for cooperative coverage in a corridor-like environment. Then, cooperative coverage in an indoor room environment is studied. Next, it is applied to cooperative coverage in a warehouse environment. Finally, four cleaning robots working together to sweep a sport field is simulated.

3.1. Cooperative coverage in a corridor-like environment

To illustrate the cooperative coverage by a multi-robot system, the proposed model is applied to a cooperative coverage with obstacle avoidance in a corridor-like environment. In the simulation for the multi-robot system, there are two neural network systems for these two robots that share mutual external input signals from the sensor information representing environmental knowledge. Each neural network has 33×28 topologically organized neurons with zero initial neural activities. The model parameters are set as $A = 10$, $B = 1$, and $D = 1$ for the shunting equation; $\mu = 0.7$ and $r_0 = 2$ for the lateral connections; and $E = 100$ for the external inputs (on parameter sensitivity, see [2]). In **Figure 4(a)**, one robot, called Robot 1, represented using solid circle starts to sweep from the lower left corner $S_1(1, 1)$; the other, Robot 2, exhibited by a hollow circle covers from the upper right corner $S_2(31, 26)$. After two robots sweep four columns in their own regions, they encounter walls and then enter to clean in a narrow corridor-like workspace [see **Figure 4(a)**]. Two robots move along smooth zigzag
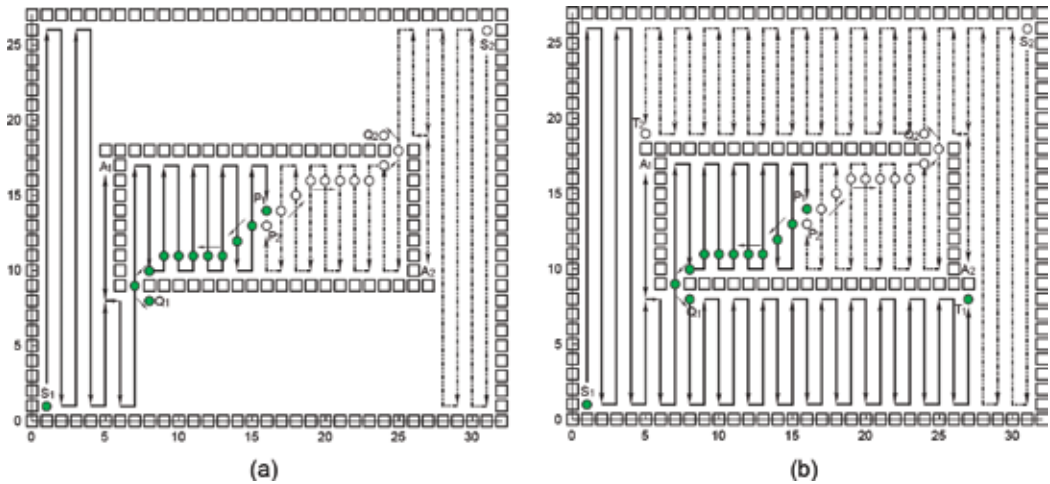
Figure 4. Cooperative coverage in a corridor-like environment. (a) When Robot 1 reaches Position $B_1(8, 8)$ and Robot 2 Position $B_2(24, 19)$; (b) when two robots fulfill the coverage task.

paths by performing back-and-forth motions. Initially, when Robot 1 reaches Position $A_1(5, 17)$ and Robot 2 Position $A_2(27, 10)$, the neural activity landscape of the neural networks is shown in **Figure 5(a)**. In this case, two robots have two neural networks and the neural activities of two neural networks are, respectively, computed through Eq. (3). Consequently, their neural activities may be plotted in a figure as a neural activity landscape of neural network.

These two robots meet at the middle section of the corridor at Positions $P_1(16, 14)$ (flag $I_i(16, 14) = -E$) and $P_2(16, 13)$ (flag $I_i(16, 13) = -E$), respectively. Obviously, Robot 1 and Robot 2 reach the central corridor simultaneously without collision in the central area as shown in **Figure 4(a)**. From the central area, these two robots are able to search point-to-point paths to move to uncovered areas since uncovered areas globally attract them (see [60]). These two robots can follow continuous and smooth paths to achieve $Q_1(8, 8)$ and $Q_2(24, 19)$, respectively. They generate CAC in these two uncovered areas and sweep them as shown in **Figure 4(b)** until they fulfill all the coverage mission at Positions $T_1(27, 8)$ and $T_2(5, 19)$. In fact, the procedure of these two robots cooperatively cover the workspace can be illustrated by the neural activity landscape of the neural networks in **Figure 5**. It demonstrates that after these two robots sweep their own areas, they clean the public aisle areas cooperatively and then move to sweep their own areas again. **Figure 5** shows the neural activity landscape when these two robots move to Positions $A_1(5, 17)$, $A_2(27, 10)$; $B_1(12, 13)$, $B_2(20, 14)$; $C_1(15, 15)$, $C_2(17, 12)$; $D_1(12, 4)$, $D_2(20, 23)$; $E_1(17, 6)$, $E_2(15, 21)$; and $F_1(25, 6)$, $F_2(7, 21)$.

3.2. Cooperative coverage in a warehouse environment

To investigate the flexibility and adaptability of cooperative coverage by multiple robots, the proposed model is applied to a warehouse environment with wall-like obstacles placed in different positions (**Figure 6**). For each robot, the neural network has 32×32 topologically organized neurons with zero initial neural activities and the model chooses the same parameters as the case above.

Robots 1 and 2, respectively, work in the lower half and upper half sections of the workspace. However, they can also assist each other, in other words, one robot can aid to cover the other areas if it has already covered its own column. In this simulation, they start from the same side in the workspace. Robot 1 represented by a solid dot starts from the lower left corner $S_1(1,1)$, whereas Robot 2 by an empty circle sweeps from the upper left corner $S_2(1,30)$.

The wall-like obstacles have influence over the cleaning assignments of two robots. The robot paths that these two robots meet at the middle of the workspace are shown in **Figure 6(a)** when they reach Positions $A_1(1,15)$ and $A_2(1,16)$, where these two robots equally accomplish the coverage tasks. The neural activity landscape of the neural networks when Robots 1 and 2 meet at Positions $A_1(1,15)$ and $A_2(1,16)$ can be found in **Figure 7(a)**.

These two robots are responsible for the different areas in the warehouse and sweep their own regions. Due to the placement of obstacles, these two robots get through different amounts of sweeping assignments. For instance, when Robot 1 arrives at Position $B_1(7,5)$, Robot 2 only reaches Position $B_2(4,23)$. The neural activity landscape of the neural networks is illustrated in **Figure 7(b)**. Similarly, when Robot 1 arrives at Position $C_1(12,15)$, Robot 2 only reaches Position
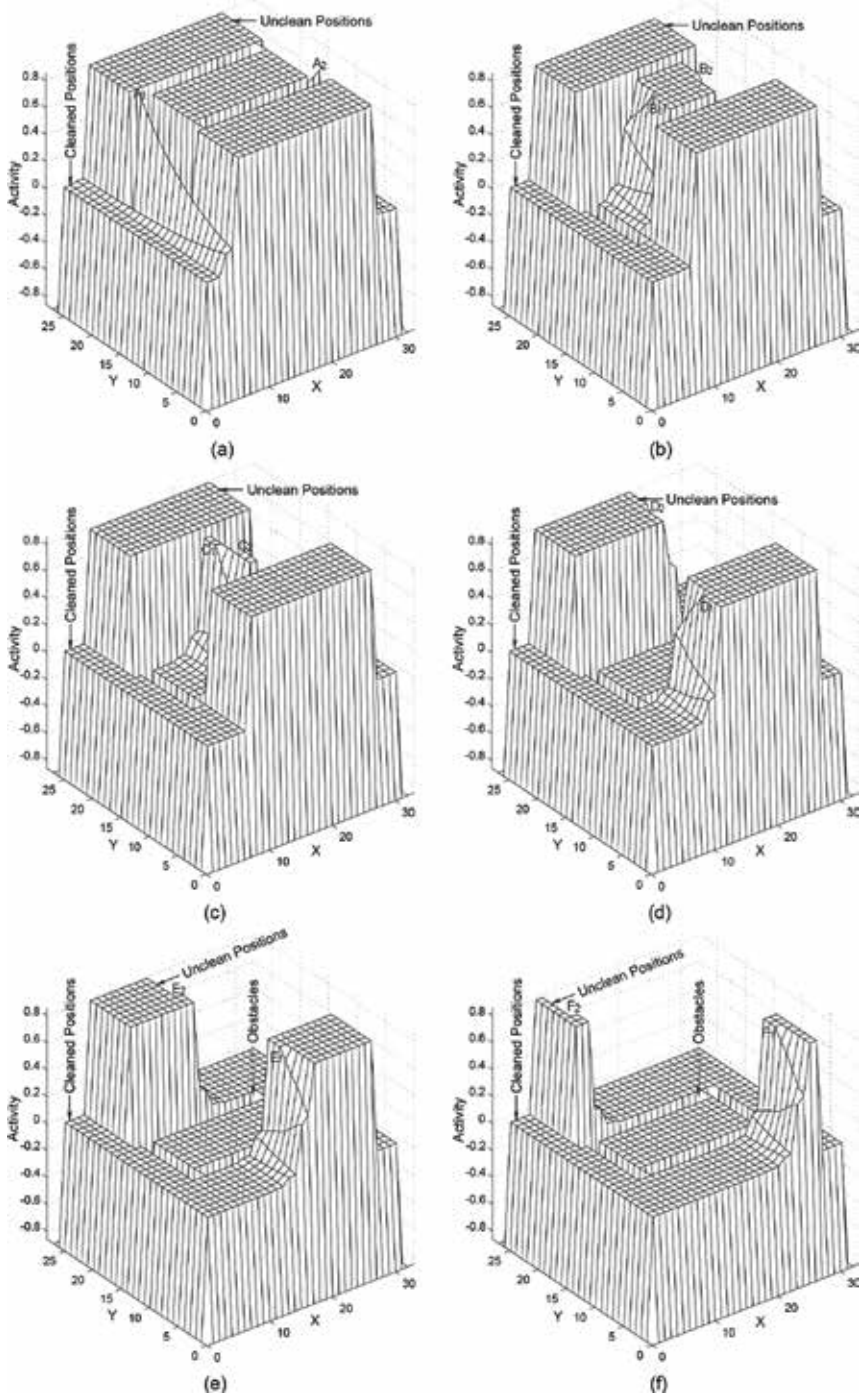
Figure 5. The neural activity landscape of the neural networks for the corridor-like environment case when Robots 1 and 2 reach (a) Positions $A_1(5, 17)$, $A_2(27, 10)$; (b) $B_1(12, 13)$, $B_2(20, 14)$; (c) $C_1(15, 15)$, $C_2(17, 12)$; (d) $D_1(12, 4)$, $D_2(20, 23)$; (e) $E_1(17, 6)$, $E_2(15, 21)$; and (f) $F_1(25, 6)$, $F_2(7, 21)$.
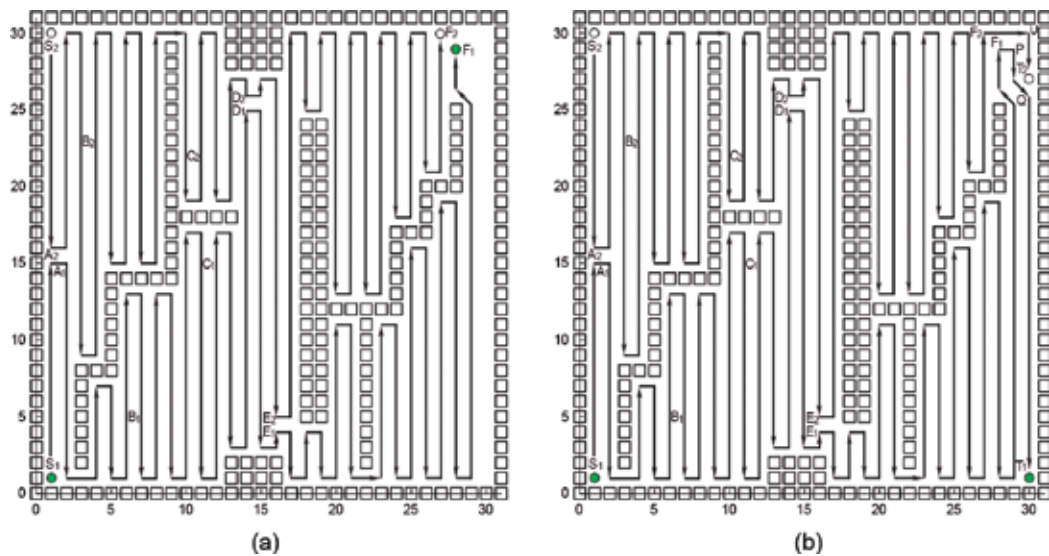
Figure 6. Cooperative coverage in the warehouse environment. (a) Robots 1 and 2 reach Positions $F_1(28,29)$ and $F_2(27,30)$; (b) two robots work cooperatively.

$C_2(10,22)$, in which the corresponding neural activity landscape of the neural networks is exhibited in **Figure 7(c)**. Obviously, when these two robots meet at Positions $D_1(14,25)$ and $D_2(14,26)$, Robot 1 is assisting Robot 2 for coverage work, where the corresponding neural activity landscape of the neural networks is shown in **Figure 7(d)**. Conversely, Robot 2 aids Robot 1 to do coverage work when they arrive at Positions $E_1(16,4)$ and $E_2(16,5)$. They can cooperatively work together without collisions to improve the cleaning productivity efficiently. The neural activity landscape of the neural networks is illustrated in **Figure 7(e)**. Finally, when Robot 1 arrives at $F_1(28,29)$, Robot 2 reaches $F_2(27,30)$, which shows that the Robot 1 has assisted to sweep one column area for Robot 2 as shown in **Figure 6(a)**. These two robots continue to sweep the rest of the workspace. Robot 1 passes Positions $P(29,29)$ and $Q(30,26)$ and reaches the final position $T_1(30,1)$, while Robot 2 attains the final position $T_2(30,27)$ via $U(30,30)$ [see **Figure 6(b)**]. The neural activity landscape of the neural networks is illustrated in **Figure 7(f)**. Ultimately, they reaches Positions $T_1(30,1)$ and $T_2(30,27)$ asynchronously. Therefore, the neural network is able to guide these two robots to complete the coverage task.

3.3. Cooperative coverage by four cleaning robots in a sports field environment

The proposed model is further applied to cooperative coverage in a sports field environment by four cleaning robots, where there exist four neural network systems and the four cleaning robots share mutual external input signals from sensory data representing the environmental information. Each neural network has 20×20 topologically organized neurons with zero initial neural activities and the same model parameters as the above case. In **Figure 8(a)**, the use of various lines is to distinguish the generated paths by the robots. Robot 1 whose paths are represented by solid lines starts to move from the lower left corner $S_1(1,1)$. Robot 2 whose
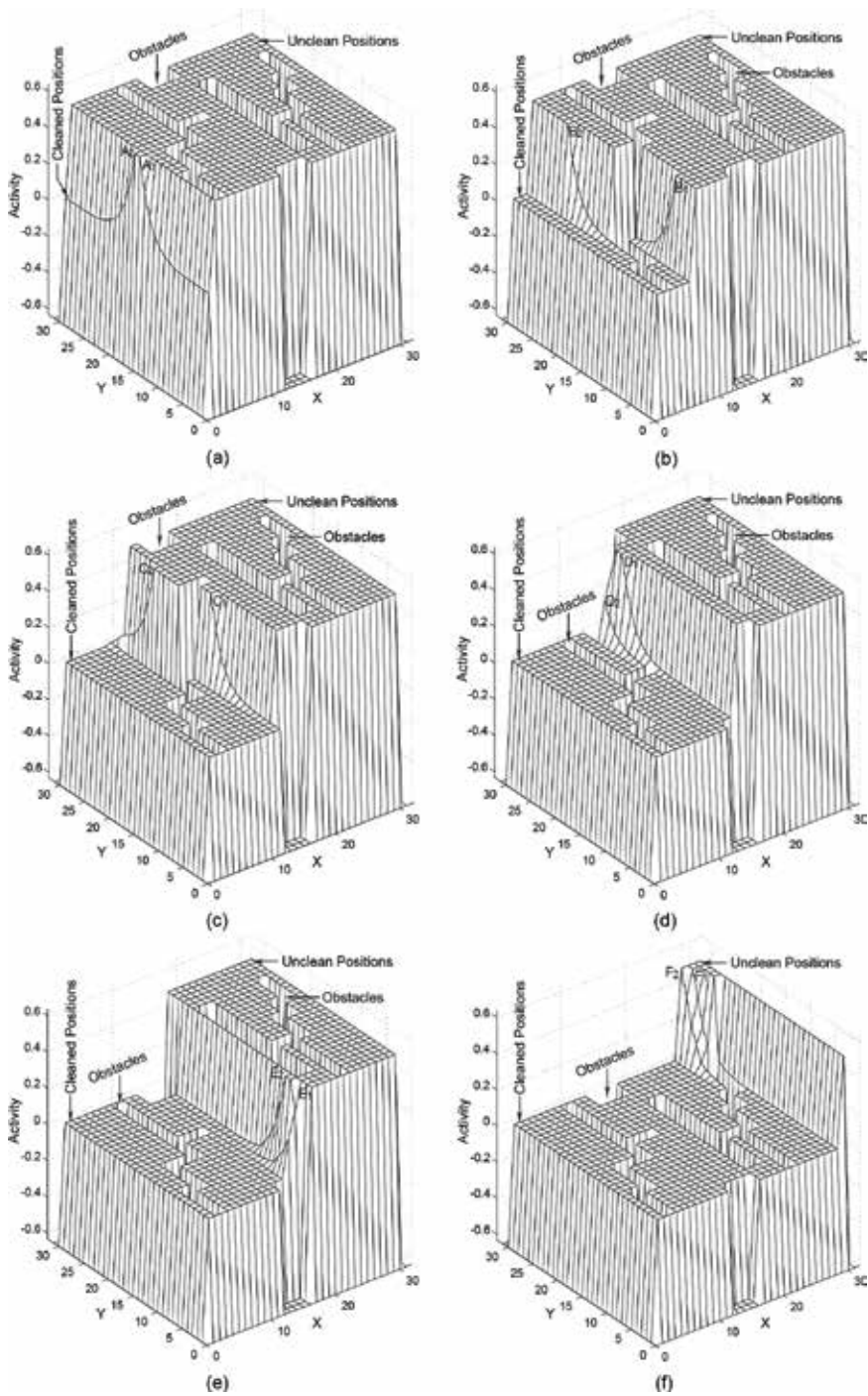
Figure 7. The neural activity landscape of the neural networks for the warehouse environment when Robots 1 and 2 reach (a) Positions $A_1(1, 15)$ and $A_2(1, 16)$; (b) $B_1(7, 5)$ and $B_2(4, 23)$; (c) $C_1(12, 15)$ and $C_2(10, 22)$; (d) $D_1(14, 25)$ and $D_2(14, 26)$; (e) $E_1(16, 4)$ and $E_2(16, 5)$; and (f) $F_1(28, 29)$ and $F_2(27, 30)$.

paths are represented by dashed lines moves from the upper left corner $S_2(1,18)$. Robot 3 whose paths are represented by dash-dotted lines starts from the upper right corner $S_3(18,18)$. Robot 4 whose paths are represented by dash-dot-dot lines sweeps from the lower right corner $S_4(18,1)$. In the simulation, the planned robot paths are shown in **Figure 8(a)**, where the four robots search snake-trail CAC paths and meet at the central area. Because for each neural network, the positive neural activity can propagate to the whole state space of the neural network, each robot can achieve CAC. If one grid is covered by one robot, it will be marked covered by external input signal ($I_i = 0$), another robot will know that it has been covered. When the four robots eventually meet, they have no collision. It shows that these four robots are able to autonomously sweep the whole workspace. They can sweep along zigzag coverage paths and avoid collisions with each other.

After they meet at the central area, where there is a deadlock situation [2], i.e., Robot 1 is at F_1 (9,9); Robot 2 at F_2 (9,10); Robot 3 at F_3 (10,10); and Robot 4 at F_4 (10,9), the robots are able to search point-to-point paths to move to any pre-defined targets. In this simulation as shown in **Figure 8(b)**, Robot 1 goes back to its initial point $G_1(1,1)$. Robot 2, 3, and 4 move back to their initial points $G_2(1,18)$, $G_3(18,18)$, and $G_4(18,1)$, respectively. They may be pre-defined to move to any points based on the demand (see [2]). The targets can globally attract the robots in the whole workspace through neural activity propagation. This case has potential applications in sport fields such as basketball or volleyball match. The four mobile robots can be assigned to clean fields together (e.g., mop sweat on the floor during the volleyball match) and then move back to their starting points during sports contest break, without any human intervention.

3.4. Cooperative coverage in an indoor environment with a robot failure

To verify the robustness and effectiveness of the proposed model, it is applied to a complicated case of cooperative coverage in an indoor environment, where there exist seven sets of
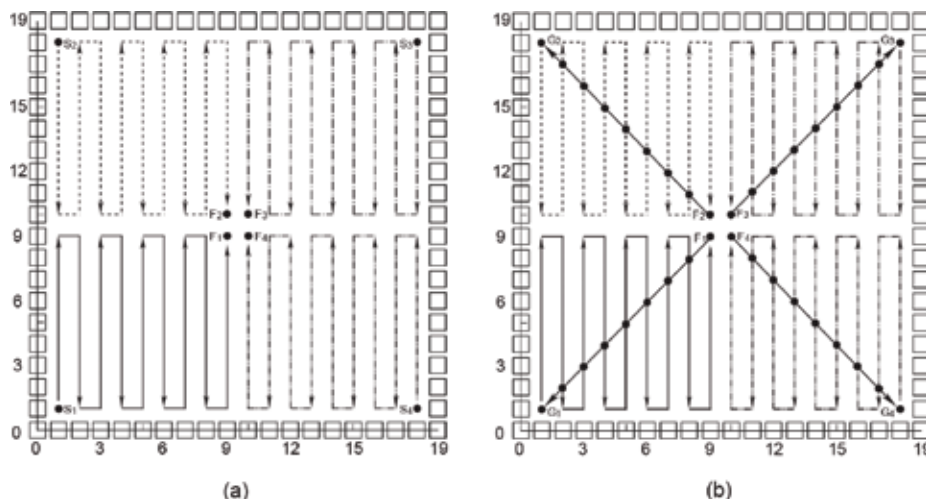
Figure 8. CAC of four mobile robots in a sport field environment. (a) The robot paths when four meet at the centre; (b) the entire robot paths after each robot returns its home position.

obstacles with different sizes and shapes in the workspace as shown in **Figure 9**. Each neural network has 33×28 topologically organized neurons with zero initial neural activities. The model parameters are chosen as $A = 50$, $B = 1$ and $D = 1$ for the shunting equation; $\mu = 0.7$ and $r_0 = 2$ for the lateral connections; and $E = 100$ for the external inputs. In this section, two simulations are performed as shown in **Figure 9**. First, these two robots work cooperatively
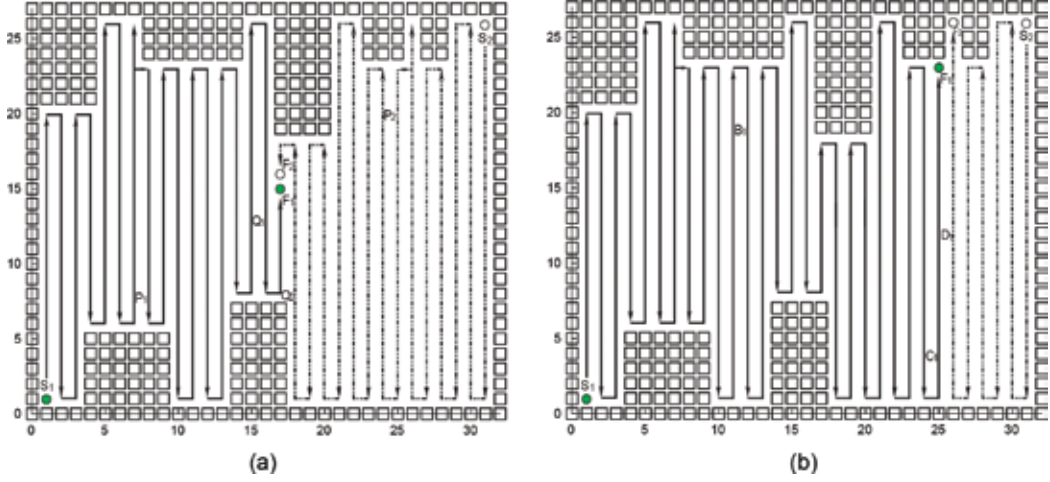
Figure 9. Complete cooperative coverage in an indoor environment with (a) both robots function properly; (b) a robot failure at Position $F_2(26, 26)$.
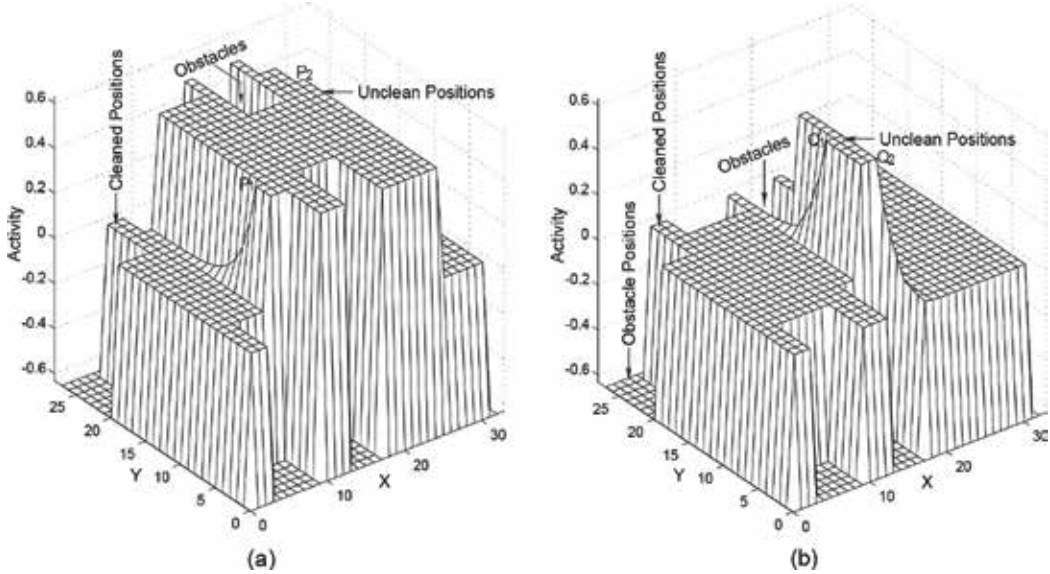
Figure 10. The neural activity landscape of the neural networks for the unstructured environment case when Robot 1 reaches (a) Position $P_1(8, 8)$; (b) Position $Q_1(16, 13)$.

[shown in **Figure 9(a)**]. Second, when Robot 2 fails at one position, Robot 1 is demonstrated to perform the rest of the work.

Robot 1 symbolized by a solid circle starts to move from the lower left corner $S_1(1,1)$ and Robot 2 by an empty circle sweeps from the upper right corner $S_2(31,26)$ as illustrated in **Figure 9**. The real-time collision-free robot paths are shown in **Figure 9(a)**, where the solid lines represent paths of Robot 1 and the dashed lines stand for those of Robot 2. These two robots cover the workspace cooperatively. Robot 1 reaches Position $P_1(8,8)$, while Robot 2 attains Position $P_2(25,20)$ as shown in **Figure 10(a)**. Continuing the work cooperatively, two robots can
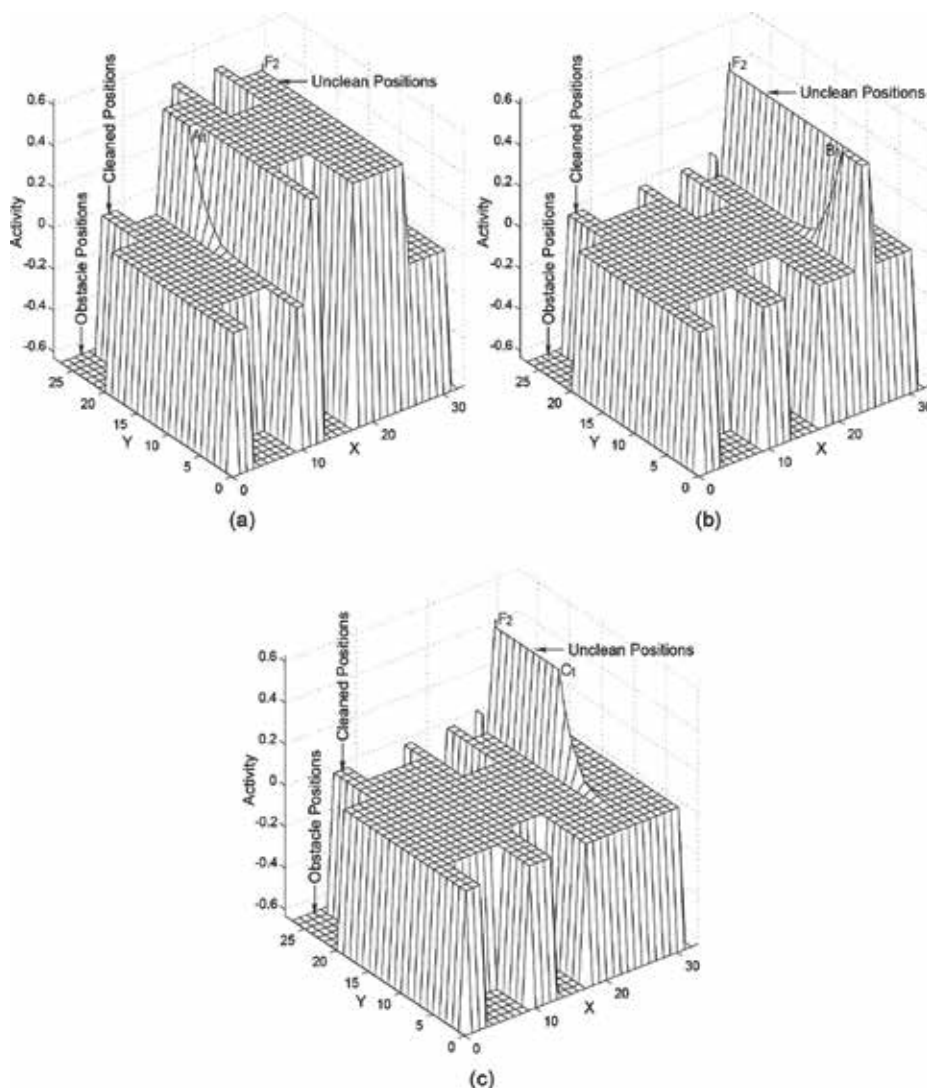
Figure 11. The neural activity landscape of the neural networks for the indoor environment case when Robot 2 fails at $F_2(26,26)$ and Robot 1 reaches (a) Position $A_1(11,19)$; (b) $B_1(24,4)$; (c) $C_1(25,13)$.

approach to achieve complete coverage of the entire workspace as uncovered areas globally attract them to visit. The neural activity landscape of the neural networks illustrates that these robots march on to achieve CAC when Robot 1 arrives at Position $Q_1(16, 13)$ and Robot 2 Position $Q_2(18, 8)$ in **Figure 10(b)**. It shows that these two cleaning robots are capable of autonomously and cooperatively sweeping the whole workspace. Not only can they sweep along a curve path to avoid the irregularly shaped obstacles but also are able to avoid collisions with each other.

Now a simulation is performed to demonstrate the robustness of the proposed model. It is assumed that Robot 2 fails to work at Position $F_2(26, 26)$ as shown in **Figure 9(b)**. Robot 1 can continue to complete the rest of coverage mission even though its partner, Robot 2, was stuck at Position $F_2(26, 26)$ [see **Figure 11(b)**]. The neural activity landscape of the neural networks when Robot 1 reaches Positions $A_1(11, 19)$ and $B_1(24, 4)$ are illustrated in **Figure 11(a)** and **(b)**, respectively. Eventually, when Robot 1 arrives at Position $C_1(25, 13)$ as shown in **Figure 11(c)**, the CAC approaches to an end. In fact, compared **Figure 9(a)** with **Figure 9(b)**, Robot 1 assists Robot 2 for eight columns of sweeping work. The neural activity landscape of the neural networks for the indoor environment case when Robot 2 fails at $F_2(26, 26)$ illustrates the area coverage progress of these two robots (**Figure 11**). Although Robot 2 fails at Position $F_2(26, 26)$, Robot 1 can still be responsible for the rest coverage work. This simulation shows that the proposed approach is robust. The complete coverage can be achieved as long as at least one single robot is able to work.

4. Conclusion

Multiple robots have the capacity for covering the areas more efficiently than a single robot. In this chapter, a biologically motivated neural network approach to cooperative area coverage by a multi-robot system is proposed, which is capable of autonomously accomplishing collision-free cooperative coverage in CAC environments. The effectiveness of the presented paradigm has been discussed and demonstrated through case studies. Multiple robots can work together to achieve a common coverage goal efficiently and robustly.

It is practical to implement the proposed approach in autonomous area coverage as no learning and no templates are required. The robustness and fault-tolerant can be ensured if some robots fail. The model algorithm is computationally simple. The robot path is generated without explicitly searching over the global free workspace or the collision paths, without explicitly optimizing any global cost functions, without any prior knowledge of the dynamic environment, without any templates, and without any learning procedures.

In the future, some research work will be carried out. First, energy-driven multirobot algorithms associated with deep reinforcement learning will be further studied to explore the minimum-energy cleaning robots. Second, the task allocation and impact of number of robots for the cleaning mission will be addressed. Third, the algorithm will be considered to be implemented on an FPGA-based platform. Finally, SLAM and robot vision will be carried out to make the cleaning algorithms more accurate.

Author details

Chaomin Luo[1]*, Gene Eu Jan[2,4], Zhenzhong Chu[3] and Xinde Li[5]

*Address all correspondence to: luoch@udmercy.edu

1 Department of Electrical and Computer Engineering, University of Detroit Mercy, Michigan, USA

2 Graduate Institute of Animation and Film Art, Tainan National University of the Arts, and Department of Electrical Engineering, National Taipei University, Taiwan

3 College of Information Engineering, Maritime University, Shanghai, China

4 Department of Electrical Engineering, National Taipei University, Taiwan

5 School of Automation, Southeast University, Nanjing, China

References

[1] Lawitzky G. A navigation system for cleaning robots. Autonomous Robots. 2000;**9**:255-260

[2] Luo C, Yang SX. A bioinspired neural network for real-time concurrent map building and complete coverage robot navigation in unknown environments. IEEE Transactions on Neural Networks and Learning Systems. 2008;**19**(7):1279-1298

[3] Ortiz F, Pastor JA, Alvarez B, Iborra A, Ortega N, Rodriguez D, Concsa C. Robots for hull ship cleaning. In: Proceedings of IEEE International Symposium on Industrial Electronics; 2007. pp. 2077-2082

[4] Wagner IA, Altshuler Y, Yanovski V, Bruckstein AM. Cooperative cleaners: A study in ant robotics. The International Journal of Robotics Research. 2008;**27**(1):127-151

[5] Oh JS, Park JB, Choi YH. Complete coverage navigation of clean robot based on triangular cell map, In: Proceedings of IEEE International Symposium on Industrial Electronics; Pusan. South Korea; 2001. pp. 2089-2093

[6] Oh JS, Choi YH, Park JB, Zheng YF. Complete coverage navigation of cleaning robots using triangular-cell-based map. IEEE Transactions on Industrial Electronics. 2004;**51**(3): 718-726

[7] Yasutomi F, Takaoka D, Yamada M, Tsukamoto K. Cleaning robot control. In: Proceedings of the IEEE International Conference on Robotics and Automation; Philadelphia. USA; 1988. pp. 1839-1841

[8] Gage DW. Randomized search strategies with imperfect sensors. In: Proceedings of SPIE, Mobile Robots VIII—The International Society for Optical Engineering; Boston. USA; 1994. pp. 270-279

[9]   Najjaran H, Kircanski N. Path planning for a terrain scanner robot. In: Proceedings of the 31st International Symposium on Robotics; Montreal. Canada; 2000. pp. 132-137

[10] Ollis M, Stentz A. Vision-based perception for an automated harvester. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems; Grenoble. France; 1997. pp. 1838-1844

[11] Ollis M, Stentz A. First results in vision-based crop line tracking. In: Proceedings of IEEE International Conference on Robotics and Automation; Minneapolis. USA; 1996. pp. 951-956

[12] Farsi M, Ratcliff K, Johnson PJ, Allen CR, Karam KZ, Pawson R. Robot control system for window cleaning. In: Proceedings of 11th International Symposium on Automation and Robotics in Construction; Brighton. UK; 1994. pp. 617-623

[13] Choset H. Coverage for robotivs - a survey of recent results. Annals of Mathematics and Artificial Intelligence. 2001;**31**:113-126

[14] Hazon N, Kaminka GA. Redundancy, efficiency, and robustness in multi-robot coverage. In: Proceedings of IEEE International Conference on Robotics and Automation; Barcelona. Spain; 2005. pp. 735-741

[15] Kurabayashi D, Ota J, Arai T, Ichikawa S, Koga S, Asama H, Endo I. Cooperative sweeping by multiple mobile robots with relocating portable obstacles. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems; Osaka. Japan; 1996. pp. 1472-1477

[16] Kurabayashi D, Ota J, Arai T, Yoshida E. Cooperative sweeping by multiple mobile robots. In: Proceedings of IEEE International Conference on Robotics and Automation; Minneapolis. USA; 1996. pp. 1744-1749

[17] Ferranti E, Trigoni N, Levene M. Brick & Mortar: An online multi-agent exploration algorithm. In: Proceedings of IEEE International Conference on Robotics and Automation; Roma. Italy, 2007. pp. 761-767

[18] Hazon M, Mieli F, Kaminka GA. Towards robust on-line multi-robot coverage. In: Proceedings of IEEE International Conference on Robotics and Automation; Orlando. USA; 2006. pp. 1710-1715

[19] Rekleitis IM, Dudek D, Milios EE. Multi-robot exploration of an unknown environment, efficiently reducing the odometry error. In: Proceedings of the 15th IEEE International Joint Conference on Artificial Intelligence; Nagoya. Japan; 1997. pp. 1340-1345

[20] Wagner IA, Lindenbaum M, Bruckstein AM. Distributed covering by ant-robots using evaporating traces. IEEE Transactions on Robotics and Automation. 1999;**15**(5):918-933

[21] Wagner IA, Lindenbaum M, Bruckstein AM. MAC versus PC: determinism and randomness as complementary approaches to robotic exploration of continuous unknown domains. The International Journal of Robotics Research. 2000;**19**(1):12-31

[22] Kong CS, New AP, Rekleitis I. Distributed coverage with multi-robot system. In: Proceedings of IEEE International Conference on Robotics and Automation; Orlando, USA. 2006. pp. 2423-2429

[23] Zheng X, Jain S, Koenig S, Kempe D. Multi-robot forest coverage. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems; Edmonton, Canada. 2005. pp. 3852-3857

[24] Zheng X, Koenig S. Robot coverage of terrain with nonuniform traversability. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems; San Diego, USA. 2007. pp. 3757-3764

[25] Agmon N, Hazon N, Kaminka GA. Constructing spanning trees for efficient multi-robot coverage. In: Proceedings of IEEE International Conference on Robotics and Automation; Orlando, USA. 2006. pp. 1698-1703

[26] Balch T, Arkin R. Behavior-based formation control for multirobot teams. IEEE Transactions on Robotics and Automation. 1998;**14**(6):926-939

[27] Fang G, Dissanayake G, Lau H. A behaviour-based optimisation strategy for multi-robot exploration. In: Proceedings of the IEEE Conference on Robotics, Automation and Mechatronics; Singapore. 2004. pp. 875-879

[28] Jung D, Cheng G, Zelinsky A. Experimentals in realising cooperation between autonomous mobile robots. In: Experiments Robotics V—The Fifth International Symposium on Experimental Robotics; Barcelona, Spain. 1997. pp. 609-620

[29] Jung D, Cheng G, Zelinsky A. Robot cleaning: An application of distributed planning and real-time vision. In: Zelinsky A, editor. Filed and Service Robotics. New York, USA: Springer; 1998. pp. 187-194

[30] Jung D, Zelinsky A. An architecture for distributed cooperative-planning in a behaviour-based multi-robot system. Robotics and Autonomous Systems. 1999;**26**:149-174

[31] Wagner IA, Lindenbaum M, Bruckstein AM. Smell as a computational resource—A lesson we can learn from the ant. In: Proceedings of the Fourth Israel Symposium on Theory of Computing and Systems; Jerusalem, Israel. 1996. pp. 219-230

[32] Williams K, Burdick J. Multi-robot boundary coverage with plan revision. In: Proceedings of IEEE International Conference on Robotics and Automation; Orlando, USA. 2006. pp. 1716-1723

[33] Burgard W, Moors M, Stachniss C, Schneider FE. Coordinated multi-robot exploration. IEEE Transactions on Robotics. 2005;**21**(3):376-386

[34] Gossage M, New AP, Cheng CK. Frontier-graph exploration for multi-robot systems in an unknown indoor environment. In: Gini M, Voyles R, editors. Distributed Autonomous Robotic Systems. Japan: Springer; 2006. pp. 51-60

[35] Ko J, Stewart B, Fox D, Konolige K, Limketkai B. A practical, decision-theoretic approach to multi-robot mapping and exploration. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems; Las Vegas, USA: 2003. pp. 3232-3238

[36] B. Yamauchi. Frontier-based exploration using multiple robots. In: Proceedings of the 2nd International Conference on Autonomous Agents; Minneapolis, USA: 1998. pp. 47-53

[37] Yamauchi B. Decentralized coordination for multirobot exploration. Robotics and Autonomous Systems. 1999;**29**(2–3):111-118

[38] Butler Z. Distributed coverage of rectilinear environments [PhD dissertation]. Carnegie Mellon University; 2000

[39] Butler ZJ, Rizzi AA, Hollis RL. Cooperative coverage of rectilinear environments. In: Proceedings of IEEE International Conference on Robotics and Automation; San Francisco, USA. 2000. pp. 2722-2727

[40] Latimer-IV D, Srinivasa S, Lee-Shue V, Sonne S, Choset H, Hurst A. Toward sensor based coverage with robot teams. In: Proceedings of IEEE International Conference on Robotics and Automation; Washington, DC, USA: 2002. pp. 961-967

[41] Rekleitis I, Lee-Shue V, New AP, Choset H. Limited communication, multi-robot team based coverage. In: Proceedings of IEEE International Conference on Robotics and Automation; New Orleans, USA: 2004. pp. 3462-3468

[42] Gabriely Y, Rimon E. Spanning-tree based coverage of continuous areas by a mobile robot. Annals of Mathematics and Artificial Intelligence. 2001;**31**(1–4):77-98

[43] Chen J, Li LR. Path planning protocol for collaborative multi-robot systems. In: Proceedings of IEEE International Symposium on Computational Intelligence in Robotics and Automation; 2005. pp. 721-726

[44] Chueh M, Yeung YLWA, Lei K-PC, Joshi SS. Following controller for autonomous mobile robots using behavioral cues. IEEE Transactions on Industrial Electronics. 2008;**55**(8):3124-3132

[45] Zavlanos MM, Pappas GJ. Dynamic assignment in distributed motion planning with local coordination. IEEE Transactions on Robotics. 2008;**24**(1):232-242

[46] Ribas L, Izquierdo M, Mujal J, Ramon E. A prototype of a low-cost, downsizeable, dynamically reconfigurable unit for robot swarms. In: Proceedings of IEEE International Symposium on Industrial Electronics; 2007. pp. 2167-2172

[47] Boonpinon N, Sudsang A. Constrained coverage for heterogeneous multi-robot team. In: Proceedings of IEEE International Conference on Robotics and Biomimetics; Sanya, China. 2007. pp. 799-804

[48] Schwager M, Slotine JJ, Rus D. Consensus learning for distributed coverage control. In: Proceedings of IEEE International Conference on Robotics and Automation; Pasadena, USA. 2008. pp. 1042-1048

[49] Huang S-J, Lian R-J. A hybrid fuzzy logic and neural network algorithm for robot motion control. IEEE Transactions on Industrial Electronics. 1997;**44**(3):408-417

[50] Su Z, Zeng B, Liu G, Ye F, Xu M. Application of fuzzy neural network in parameter optimization of mobile robot path planning using potential field. In: Proceedings of IEEE International Symposium on Industrial Electronics; 2007. pp. 2125-2128

[51] Arleo A, Smeraldi F, Gerstner W. Cognitive navigation based on nonuniform gabor space sampling, unsupervised growing networks, and reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems. 2004;**15**(3):639-652

[52] Er MJ, Deng C. Obstacle avoidance of a mobile robot using hybrid learning approach. IEEE Transactions on Industrial Electronics. 2005;**52**(3):898-905

[53] Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. Journal of Physiology (London). 1952;**117**:500-544

[54] Grossberg S. Nonlinear neural networks: principles, mechanisms, and architecture. Neural Networks. 1988;**1**:17-61

[55] Svennebring J, Koenig S. Building terrain-covering ant robots. Autonomous Robots. 2004; **16**(3):313-332

[56] Xu WL, Pap J-S, Bronlund J. Design of a biologically inspired parallel robot for foods chewing. IEEE Transactions on Industrial Electronics. 2008;**55**(2):832-841

[57] Domenech JE, Regueiro CV, Gamallo C, Quintia P. Learning wall following behaviour in robotics through reinforcement and image-based states. In: Proceedings of IEEE International Symposium on Industrial Electronics; 2007. pp. 2101-2106

[58] Lee C-Y, Lee J-J. Multiple neuro-adaptive control of robot manipulators using visual cues. IEEE Transactions on Industrial Electronics. 2005;**52**(1):320-326

[59] Grossberg S. Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. IEEE Transactions on Systems, Man, and Cybernetics. 1983;**13**(5):815-926

[60] Yang SX, Luo C. A neural network approach to complete coverage path planning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2004;**34**(1):718-725

[61] Gutierrez-Osuna R, Janet JA, Luo RC. Modeling of ultrasonic range sensors for localization of autonomous mobile robots. IEEE Transactions on Industrial Electronics. 1998;**45**(4): 654-662

[62] Ulusar UD, Akin HL. Design and implementation of a real time planning system for autonomous robots. In: Proceedings of IEEE International Symposium on Industrial Electronics; 2006. pp. 74-79

[63] Fong T, Thorpe C, Baur C. Multi-robot remote driving with collaborative control. IEEE Transactions on Industrial Electronics. 2003;**50**(4):699-704

[64] Luo RC, Su KL. Multilevel multisensor-based intelligent recharging system for mobile robot. IEEE Transactions on Industrial Electronics. 2008;**55**(1):270-279

[65] Won WJ, Ban SW, Lee M. Real time implementation of a selective attention model for the intelligent robot with autonomous mental development. In: Proceedings of IEEE International Symposium on Industrial Electronics; 2005. pp. 1309-1314

[66] Perez DP, Gomez ES, Quintas MM. Simultaneous localization and structure reconstruction of mobile robots with external cameras. In: Proceedings of IEEE International Symposium on Industrial Electronics; 2005. pp. 1321-1326

A Modified Neuro-Fuzzy System Using Metaheuristic Approaches for Data Classification

Mohd Najib Mohd Salleh, Noureen Talpur and
Kashif Hussain Talpur

Additional information is available at the end of the chapter

Abstract

The impact of innovated Neuro-Fuzzy System (NFS) has emerged as a dominant technique for addressing various difficult research problems in business. ANFIS (Adaptive Neuro-Fuzzy Inference system) is an efficient combination of ANN and fuzzy logic for modeling highly non-linear, complex and dynamic systems. It has been proved that, with proper number of rules, an ANFIS system is able to approximate every plant. Even though it has been widely used, ANFIS has a major drawback of computational complexities. The number of rules and its tunable parameters increase exponentially when the numbers of inputs are large. Moreover, the standard learning process of ANFIS involves gradient based learning which has prone to fall in local minima. Many researchers have used meta-heuristic algorithms to tune parameters of ANFIS. This study will modify ANFIS architecture to reduce its complexity and improve the accuracy of classification problems. The experiments are carried out by trying different types and shapes of membership functions and meta-heuristics Artificial Bee Colony (ABC) algorithm with ANFIS and the training error results are measured for each combination. The results showed that modified ANFIS combined with ABC method provides better training error results than common ANFIS model.

Keywords: adaptive neuro-fuzzy inference system, meta-heuristics, classification problem

1. Introduction

The recent advances in artificial intelligence and soft computing techniques have opened new avenues for researchers to explore their applications. These machine learning techniques consist

of several intelligent computing paradigms, including artificial neural networks (ANN), support vector machine (SVM), decision tree, neuro-fuzzy systems (NFS), which have been successfully employed to model various real-world problems [1]. These problems broadly range from engineering to finance, geology and bio-sciences etc.

Among the other soft computing techniques mentioned above, ANFIS is an efficient combination of ANN and fuzzy logic for modeling highly non-linear, complex, and dynamic systems [2]. It has been proved that, with proper number of rules, an ANFIS system is able to approximate every plant. Therefore, ANFIS systems are widely used and play the advantage of good applicability since they can be interpreted as non-linear modeling and conventional linear techniques for state estimation and control [3]. Even though it has been widely used, ANFIS has a major drawback of computational complexities. The number of rules and its tunable parameters increase exponentially when the numbers of inputs are large. Moreover, the standard learning process of ANFIS involves gradient based learning which has prone to fall in local minima. The systems designed in literature generally have few inputs and ANFIS models with large inputs have not been implemented due to curse of dimensionality and many researchers have used meta-heuristic algorithms to tune parameters of ANFIS. This study will modify ANFIS architecture to reduce its complexity. The proposed approach will focus on trying different types of membership functions because ANFIS accuracy highly depends on the type and shape of its membership functions. This research will propose solution to implement ANFIS for the problems with large number of inputs. Therefore, the problem of curse of dimensionality will be addressed in this research.

This study is organized in the following order: Section 2 defines the comprehensive literature review and gap analysis. Section 3 presented the methodology of the research to solve the gaps identified in previous section. Section 4 explains the results and analysis based on the experiments performed in this study. Section 5 summarized the whole research finding and future work.

2. Research background

Recently, Neuro-Fuzzy system has gained more attraction between research communities than other types of fuzzy expert systems since it combines the advantages of learning ability of neural network and reasoning ability of fuzzy logic to solve many non-linear and complex real-world problems with high accuracy. This combination has been broadly used in the areas of education system, medical system, electrical system, traffic control, image processing, predictions and control of linear and nonlinear systems [4]. Since, every technique has few advantages and limitations such as; fuzzy logic is good on describing how they reach to certain decision, but they cannot learn rules themselves while neural networks are good on pattern matching whereas they cannot give a clear picture of how they reach to the certain decision. Adaptability is the main advantage of neural networks; therefore, it has ability to adjust their weights automatically to optimize their behavior as pattern recognizers, decision makers, predictors, etc. In fuzzy expert systems, it is not an easy task to find appropriate membership function parameters and rules. It either needs high level of expertise or a hard practice of trial-and-error [5]. These limitations

are the vital reasons behind the inspiration of building hybrid networks by combining two techniques to overcome the limitations of an individual technique. Among those hybrid networks, adaptive neuro-fuzzy inference system (ANFIS) is one of the best combinations of neural network and fuzzy logic. Next section will be explaining the basic structure and working of ANFIS.

2.1. Adaptive neuro-fuzzy inference system (ANFIS)

The concept of ANFIS was introduced by Jang in 1993, which is a proficient combination of neural network and fuzzy logic. Since ANFIS use fuzzy logic, Fuzzy inference system (FIS) is a useful soft computing technique introducing the concept of fuzzy set theory, fuzzy if-then rules, and fuzzy reasoning; therefore, while designing the ANFIS, the selection of the FIS is the major concern. It is a framework of neuro-fuzzy model that can integrate human expertise as well as adapt itself through learning. As an adaptive neuro-fuzzy model; it has advantage of being flexible, adaptive and effective for non-linear complex problems [6]. Recently, ANFIS has been successfully applied to the applications involving classification, rule-based process controls and pattern recognition.

ANFIS consists of five layers, so nodes of each layer are connected to another layer by directed links. Therefore, to produce the output for a single node, each node performs a particular function on its incoming signals. Hence it is often known as a multilayer feed-forward network. The main objective of the ANFIS is to determine the optimum values of the equivalent fuzzy inference system parameters by applying a learning algorithm.

Figure 1 shows complete ANFIS architecture. The five layers are: (i) fuzzification layer, (ii) product layer, (iii) normalized layer, (iv) defuzzification layer, (v) output layer. The two types of nodes are fixed (circle nodes) and adaptable (square nodes). Two fuzzy if-then rules with two inputs x, y and one output f are considered as:

$$Rule\ 1: If\ x\ is\ A_1\ and\ y\ is\ B_1\ then\ f\ =\ p_1\,x + q_1\,y + r_1$$

$$Rule\ 2: If\ x\ is\ A_2\ and\ y\ is\ B_2\ then\ f\ =\ p_2\,x + q_2\,y + r_2$$

A_1, A_2, B_1, B_2 are non-linear parameters and A_i, B_i are fuzzy membership functions. Whereas $p_1, p_2, q_1, q_2,$ and $r_1, r_2,$ are linear parameters that are the design parameters identified during training process. Parameters in IF part are known as antecedent or premise parameters, whereas, parameters in THEN part are known as consequent parameters. Nodes of layer 1 (Premise part) and layer 4 (consequent part) are trainable or adaptable nodes, while the nodes of layer 2 (product) and layer 3 (normalization) are fixed nodes. To execute the above two rules, the five-layer architecture of ANFIS is explained as following:

2.1.1. Layer 1 (Fuzzification)

This layer is an adaptive layer with the nodes of square shape. Every input of node i in this layer is adaptive membership function to generate the membership degree of linguistic variables. Membership function can be of any shape, i.e. Triangle, Trapezoidal, Gaussian, or generalized Bell function.
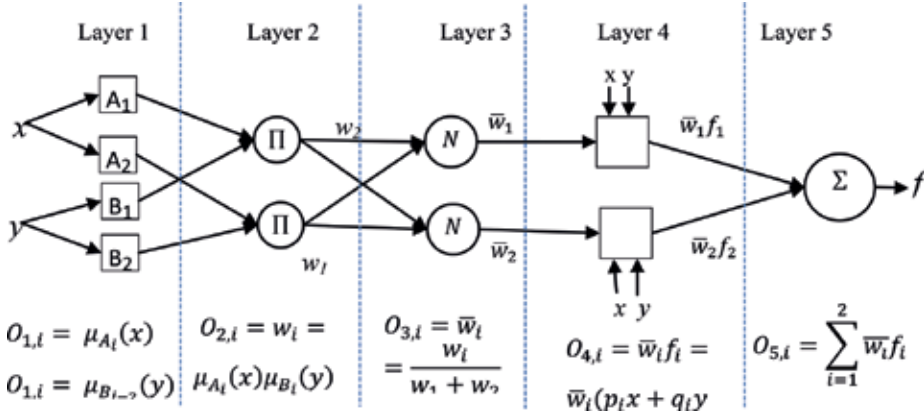
Figure 1. Basic ANFIS's architecture.

$$O_{1,i} = \mu_{A_i}(x), \quad i = 1, 2 \text{ or } O_{1,i} = \mu_{B_{i-2}}(y), i = 3, 4. \tag{1}$$

Here, x and y are the two inputs and if μ_A and μ_B are Gaussian MFs (Eq. 2), they are specified by two parameters center c and width σ, which are referred to as premise parameters.

$$Gaussian\ (x; c, \sigma) = e^{-\frac{1}{2}\left(\frac{x-c}{\sigma}\right)^2} \tag{2}$$

$O_{1,i}$ is the output of layer l and the ith node.

2.1.2. Layer 2 (Product)

These are fixed nodes that represent the product Π to calculate firing strength of a rule. This layer accepts input values from first layer and turns as a membership function to represent fuzzy sets of respective input variables. The output of each node is the product of all the incoming signals that are coming to it.

$$O_{2,i} = w_i = \mu_{A_i}(x)\,\mu_{B_i}(y), \quad i = 1, 2 \tag{3}$$

2.1.3. Layer 3 (Normalization)

Nodes of layer 3 are also fixed nodes. All nodes in this layer are labeled as N. Each node normalizes firing strength of a rule from previous layer by calculating the ratio of the ith rule's firing strength to the sum of all rules' firing strength.

$$O_{3,i} = \overline{w}_i = \frac{w_i}{w_1 + w_2}, \quad i = 1, 2 \tag{4}$$

Where $\overline{w}$ is referred to as normalized firing strength of a rule.

2.1.4. Layer 4 (Defuzzification)

Nodes in this layer are adaptive with node function,

$$O_{4,i} = \overline{w}_i f_i = \overline{w}_i(p_i x + q_i y + r_i), \quad i = 1, 2 \tag{5}$$

where $\overline{w}$ is rule's normalized firing strength and $\{p_i, q_i, r_i\}$ is a first order polynomial. $O_{4,i}$ represents output of layer 4. Parameters in this layer are linear parameters well-known as consequent parameters. These parameters are identified during the training process of ANFIS.

2.1.5. Layer 5 (Overall output)

This single node is called output layer which is labeled as "$\sum$." This layer only does summation of outputs of all rules in previous layer and converts fuzzy result into crisp output.

$$O_{5,i} = \sum_{i=1}^{2} \overline{w}_i f_i = \frac{\sum_{i=1}^{2} \overline{w}_i f_i}{w_1 + w_2} \tag{6}$$

2.2. ANFIS learning algorithm

ANFIS learns and update its all modifiable parameters by using two pass learning algorithm; forward pass and backward pass. ANFIS train its parameters such as c, σ (MF parameters) and p_i, q_i, r_i (consequent parameters) for minimizing error between actual and the desired output; using a hybrid of gradient descent (GD) and least squares estimator (LSE). During forward pass of the learning algorithm, consequent parameters are updated by LSE method and signal is node outputs. During backward pass, the premise parameters are updated by the GD algorithm and the error signals propagate backward from the output layer to until input layer. At this point, neural network learns and train to determine parameter values that can sufficiently fit the training data (**Table 1**).

2.3. Data partitioning

ANFIS can be constructed by partitioning of the input-output data into rule patches. So, this can be accomplished by using three methods such as; genfis1 (grid-partitioning), genfis2 (subtractive clustering) and genfis3 (Fuzzy C-Mean). Grid partitioning employs grid partitioning

	Forward pass	Backward pass
Antecedent parameters	Fixed	GD
Consequent parameters	LSE	Fixed
Signals	Node outputs	Error signals

Table 1. Two pass hybrid learning algorithm for ANFIS.

approach that divides data space into grids based on the number of memberships function per input. Generally, it is appropriate to use grid partitioning only for problems with a less than six number of input variables. The number of rules increases exponentially when the number of inputs increases in underlying system [7]. To avoid this issue, the clustering approach is an effective partitioning method that divides data points into groups (clusters) according to the membership grade or degree. For this reason, clustering based methods seem to be preferred to the grid partitioning techniques. Subtractive clustering method is an effective clustering approach when there is no clear idea about how many clusters will be needed for a given dataset. It reduces the computation time by finding the center of clustering by using the data itself [8]. Fuzzy C-Mean (FCM) algorithm allows one data to belong to two or more clusters. Thus, FCM let membership function value be in the range of 0 until 1 which certainly improve the result [9]. Furthermore, other than partitioning methods, to train ANFIS parameters, many researchers have proposed ANFIS training methods based on meta-heuristic algorithms. Following Section 2.4 discuss more about meta-heuristic algorithms.

2.4. Optimization using meta-heuristics

In meta-heuristic algorithms, *meta* means "higher level," all meta-heuristic algorithms use some trade-off of local search and global exploration. The main components of any meta-heuristic algorithm are: exploitation and exploration. Exploration means generating diverse solutions to explore the search space on the global scale, while exploitation means focusing on the search in a local region by exploiting the information that a current good solution is found in this region. As for ANFIS, structure learning and parameters identification are the two dimensions of ANFIS training. Some have focused on either of the two dimensions, while others have tried to work on both of the issues. The original ANFIS proposed by Jang uses hybrid learning, which uses GD and LSE. However, the drawbacks of GD have opted the researchers to different alternatives such as; ant colony, particle swarm optimization (PSO), firefly algorithm (FFA), cuckoo search algorithm (CSA), artificial bee colony (ABC). Among these training methods in meta-heuristic paradigm, this study will employ ABC (Artificial Bee Colony) optimization algorithm to train the parameters of ANFIS model instead of gradient based learning mechanism.

2.5. Artificial Bee Colony algorithm

Artificial Bee Colony (ABC) is a swarm intelligence-based algorithm which is inspired by the intelligent behavior of honey bees. It was introduced by DervisKaraboga in 2005 to solve optimization problems. Since then, the ABC algorithm has been used infields such as data mining, image processing and numerical problems. As in **Figure 2**, ABC provides a population-based search, the bees searching for food are divided into three groups in ABC: employed bees, onlooker bees and scout bees. Employed bees used to look around the search space to hunt and gather information about the position and quality of food source while the onlooker bees stay in hive and choose the food source based on information given by the employed bees. Scout bees try to replace the abandoned employed bees to search for new food sources arbitrarily [21]. The food source's position is a possible solution to theoptimization problem. The amount of nectar in a food source measured the quality of the problem.
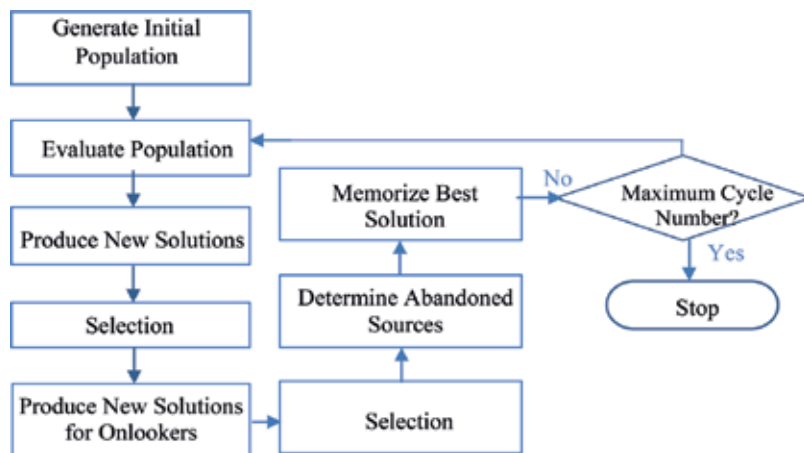
Figure 2. Flowchart of Artificial Bee Colony (ABC) algorithm.

2.6. Comparative study of ANFIS

The successful integration of neural network and fuzzy logic models in the form of ANFIS holds the advantages of solving applications that are highly non-linear and complex. Based on the robustness in results, ANFIS has been implemented in a wide variety of applications including classification tasks, rule-based process control, pattern recognition [4]. Even though ANFIS is one of the best tradeoff between neural network and fuzzy systems, providing smoothness and adaptability in the system because of fuzzy logic interpolation and the neural network back-propagation, on the other hand, the model faces strong computational complexity.

Generally, literature emphasis on three main problems regarding ANFIS to solve computational complexity: reducing rule-base, developing efficient training methods and membership functions selection. A significant increment in rules increase the complexity of ANFIS architecture as rules are generated with all possible combinations of antecedent and quality of result depends on the effectiveness of these rules [10]. Therefore, a careful study of the literature reveals that there have been many techniques to achieve rule-base minimization and accuracy maximization. These techniques include selecting potential and removing non-potential rules from the entire ANFIS knowledge-base, such as; Karnaugh map (K-Map) [11]. Moreover, other than rule-base minimization issue, training the parameters of ANFIS model is one of the main issues encountered when the model is applied to the real-world problems. The original ANFIS architecture that was introduced by Jang has drawbacks, since it uses hybrid learning algorithm which is the combination of GD and LSE. Because of using GD, it has problem to be likely trapped in local [6]. To cope with this many researchers have proposed meta-heuristic algorithms; such as, genetic algorithm (GA), particle swarm optimization (PSO), artificial bee colony (ABC) etc. Similarly, a new hybrid method was introduced in literature by Orouskhani and Mansouri [12]. This study modified Cat Swarm Optimization for training the parameters of ANFIS by updating antecedent parameters. On the other hand,

Karaboga and Kaya [13] proposed an Adaptive and Hybrid Artificial Bee Colony (*a*ABC) algorithm to train all parameters of ANFIS. Najafi [14] employed PSO with ANFIS to optimize and train parameters for prediction of viscosity of mixed oils. Along with issue such as; rule base and training methods; ANFIS also suffer from uncertainty. Therefore, techniques like fuzzy logic have been applied because fuzzy logic with the help of membership function is capable to describe uncertainty [15]. As ANFIS also practice fuzzy logic, therefore correct choice of membership functions is most important factor in building the ANFIS model. Although, various studies in literature can be found for the choice of membership function in fuzzy inference system such as; Suntae [16] did the comparison of membership functions in security robot system for decision making. Similarly, the key focus of Saha [17] study was to investigate the best membership function in applications of sign language. Adil [18] compared the effects of different types of membership functions to determine the performance of fuzzy logic controller. But for the case of neuro-fuzzy systems like ANFIS, such studies have been barely found.

As ANFIS accuracy is highly dependent to the type and shape of membership function, therefore, this study will try different types and shapes of membership function to get best suitable membership function for ANFIS model. Additionally, this study will focus on modifying the standard ANFIS architecture by reducing the fourth layer as consequent part of rules contains more number of parameters, plus; one of the best meta-heuristic approach will be used to tune the parameters of ANFIS model. The final model will be ideal for dataset having large number of inputs. The proposed approach to achieve the goal of this study is further explained in next section which defines research methodology.

3. Research methodology

The research methodology starts by collecting classification datasets. This research will solve six real world classification problems Iris, Breast Cancer, Car evaluation, Teacher Assistant, Glass Identification and Seeds based on small to large number of inputs (4–10 inputs) taken from the University of California Irvine Machine Learning Repository (UCIMLR). For splitting the data into training and testing purpose, according to literature [19, 20] most researchers practiced the 70:30 ratio (70% training, 30% testing) because the more data applied for the training, the more optimal and accurate results a system generates. Therefore, in this study the 70% of the dataset instances were selected for training set and the remaining 30% of the dataset instances were chosen for testing set.

Right after collecting, preprocessing and partitioning the selected data, in modification phase, different types and shapes of membership functions will be examined in standard ANFIS architecture while solving six classification problems taken from UCI repository. ANFIS architecture will be modified to lessen its computational complexities because ANFIS's basic architecture contains five layers and use GD and updating parameters consumes a lot of computational time when the number of parameters are large. The dataset collected will be then applied in

modified ANFIS architecture to compare and validate results between standard ANFIS architecture and modified ANFIS architecture. The outcome of this research will be modified ANFIS architecture as stated in proposed approach (**Figure 3**: Research Framework).

3.1. Proposed approach

Among various clustering techniques available, grid partitioning (genfis1), subtractive clustering (genfis2) and fuzzy C-Mean clustering (genfis3) most widely used during clustering into fuzzy inference system (FIS). It acts as a model that will reflect the relationship between the different input parameters. As it is observed from the comparative analysis and literature review that those clustering approaches are very useful to generate better accuracy for ANFIS model, since it generates maximum number of rules by considering all possibilities, but it also increases computational cost as consequent part of rules contains more number of parameters. Therefore, fourth layer which holds linear coefficients shares most of the computational cost of training algorithm because consequent part holds more number of trainable parameters than the parameters in premise part. Following example explains number of trainable parameters holds by premise and consequent part:

$$F(n, m, p) = n \times m \times p + m^n \times (n + 1)$$

Where n represents number of inputs and each input n is partitioned into m number of membership functions and parameters need to be trained or modified in membership function are labeled as p. According to this, total number of premise trainable parameters are $n \times m \times p$. Similarly, consequent trainable parameters per rule are $n + 1$ and m^n expresses total number of rules in system. Therefore, the total number of consequent trainable parameters are $m^n \times (n + 1)$.
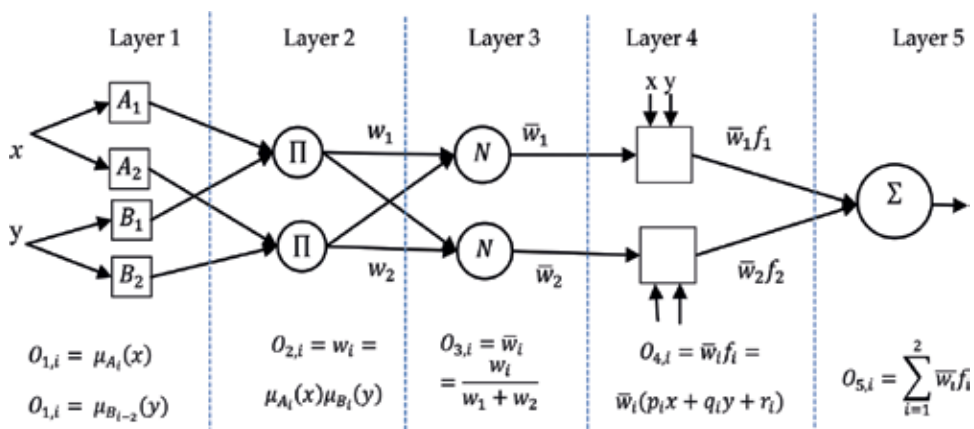
Figure 3. Standard architecture of ANFIS with function associated to each layer.

For instance; If dataset applied in ANFIS model having $n = 4$ inputs, $m = 2$ (membership functions) for each input and membership type of triangular function with parameters $p = 3$, the number of premise and number of consequent trainable parameters will be as follows:

$$F(n, m, p) = n \times m \times p + m^n \times (n + 1)$$

$$F(4, 2, 3) = 4 \times 2 \times 3 + 2^4 \times (4 + 1)$$

$$F(4, 2, 3) = 24 + 80 = 104$$
$$= \text{Premise} + \text{Consequent} + \text{Total trainable parameters}$$

Thus, out of the total trainable parameters 104, the number of premise trainable parameters is 24 and consequent trainable parameters are 80 which is far more than premise parameters. Therefore, removal of fourth layer may contribute in reduction of computation hence ANFIS architecture can be reduced to four layers. Following **Figure 3** shows standard architecture of ANFIS with 5 layers and their associated functions in each layer. This research will modify ANFIS architecture by removing fourth layer which contain adaptable nodes and holds more parameters to update. As observed in **Figure 3**, the calculation performed by fourth layer practice this following equation with the node function:

$$f_i = p_i x + q_i y + r_i$$
$$\text{where } O_{4,i} = \overline{w}_i f_i = \overline{w}_i (p_i x + q_i y + r_i), \qquad i = 1, 2$$

Where $\overline{w}$ is rule's normalized firing strength from third layer and $\{p_i, q_i, r_i\}$ is a first order polynomial from f_i. Parameters in fourth layer are referred to as consequent parameters and are identified during the training process of ANFIS. This research will try to remove two extra parameter p_i and q_i from the function f_i in fourth layer as following $f_i = p_i x + q_i y + r_i$.

To make it $f_i = r_i$, this approach will reduce two trainable parameters and likewise make third layer as an adaptable to merge the fourth layer function into third layer as:

$$O_{3,i} = \overline{w}i \times \text{fi}(r_i) \tag{7}$$

So, the resultant ANFIS will be a modified architecture that is more simple architecture comprising total four layers and less trainable parameters. For example, following **Figure 4** shows proposed approach to modify ANFIS architecture to make it simple and less complex:

According to the modified architecture of ANFIS, the total parameters trained by system will be reduced. By taking the same example as standard ANFIS model. If dataset applied in ANFIS model having $n = 4$ inputs, $m = 2$ (membership functions) for each input and membership type of gaussian function with parameters $p = 2$, the number of premise and number of consequent trainable parameters will be as follows:
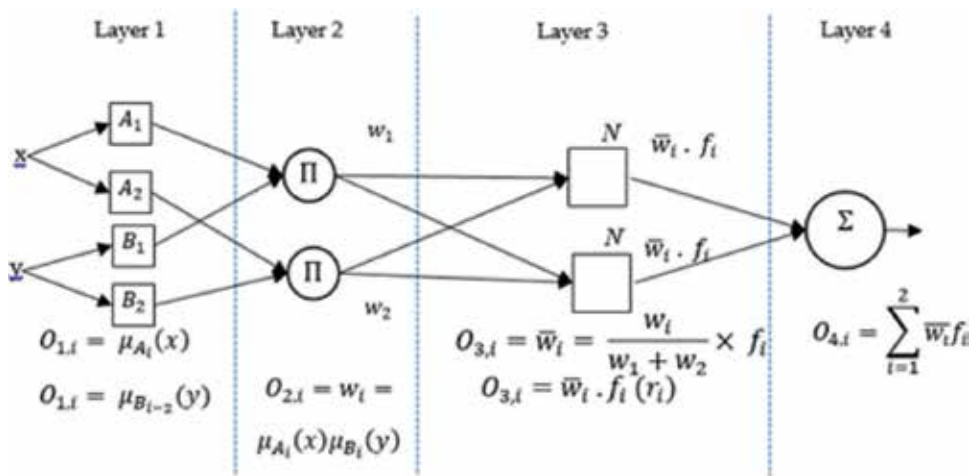
Figure 4. Proposed approach to modify ANFIS architecture.

$$F(n, m, p) = n \times m \times p + m^n \times 1$$

$$F(4, 2, 3) = 4 \times 2 \times 2 + 2^4 \times 1$$

$$F(4, 2, 3) = 16 + 16 = 32$$

$$= \text{Premise} + \text{Consequent} = \text{Total trainable parameters}$$

Hence, it can be observed that out of 32 total trainable parameters, the premise parameters are 16 and consequent parameters are 16. The consequent parameters in modified ANFIS model are far less than the consequent parameters in standard ANFIS model.

Therefore, after modifying the architecture, different datasets with small to large amount of inputs will be applied in modified ANFIS to compare results in terms of computation time and efficiency with performance validation matrix MSE percentage of accuracy, number of trainable parameters and number of epochs.

Furthermore, training algorithm also plays an important role to train ANFIS model. It is observed form the literature review section, that the meta-heuristic algorithms, such as ant colony, particle swarm optimization (PSO), firefly algorithm (FFA), cuckoo search algorithm (CSA), artificial bee colony (ABC) and many more algorithms have been developed and used to train ANFIS model by researchers recently. Among these training methods in meta-heuristic paradigm, this study will employ ABC (Artificial Bee Colony) optimization algorithm to train the parameters of ANFIS model instead of gradient based learning mechanism. Therefore, this study will not only reduce computational complexity of ANFIS architecture,

but this approach will also apply efficient training mechanism to train ANFIS parameters by using meta-heuristic swarm intelligence-based approach.

4. Experimental results and discussions

To achieve objectives of the research, this section discusses and evaluates the experimental results of the study in detail by following the proposed research framework carefully, explained in previous section. Firstly, the experiments have been performed with ANFIS models genfis1, genfis2 and genfis3 to investigate the best suitable membership function for input parameters as ANFIS model. Secondly, standard ANFIS model consists of five layers was modified by employing ABC (Artificial Bee Colony) optimization algorithm to train the parameters of ANFIS model instead of gradient based learning mechanism. Therefore, this study will reduce computational complexity of ANFIS architecture. The performance of proposed modified ANFIS model was compared with another three clustering genfis1, genfis2 and genfis3 based ANFIS models with the measurement criteria of MSE, percentage of accuracy and number of trainable parameters with the number of epochs. These simulations were performed to solve the classification problems; thus, six benchmark classification datasets naming Iris, Teacher Assistant Evaluation, Car Evaluation, Seeds, Breast Cancer and Glass Identification were taken from University of California Irvine Machine Learning Repository (UCIMLR) ranging from 4 to 10 input attributes.

4.1. Membership function

To evaluate the results of membership functions on performance of ANFIS, Fuzzy Logic Toolbox™ was used in MATLAB to employ ANFIS models with clustering methods genfis1, genfis2 and genfis3 into FIS as input model. Furthermore, most of the settings for ANFIS models were used as default as mentioned in the toolbox, however the distinguishing changes are presented in **Table 2**.

As ANFIS with grid-partitioning method (genfis1) offers to try and test different types of membership functions, it is noteworthy to mention that ANFIS with grid-partitioning (genfis1) was used to test 4 basic types of membership functions, however subtractive clustering (genfis2) and fuzzy c-mean clustering (genfis3) use Gaussian types of membership function by default in MATLAB toolbox. The performance of ANFIS with different partitioning methods and membership functions is evaluated using Root Mean Square Error (RMSE). Similarly, performance of ANFIS with different membership functions have been

ANFIS	Membership function type	Number of membership functions	Epochs	Error tolerance
genfis1	trapmf, gbellmf, trimf, gaussmf	2	200	0
genfis2	gaussmf	10	200	0
genfis3	gaussmf	10	200	0

Table 2. ANFIS settings for membership functions.

ranked from smallest to largest (rank 1–6) according to the sum of training and testing RMSE for each dataset. Furthermore, the average of these ranks is being computed to generate overall rank in the group of membership functions. **Table 3** presents the final result of comparison in terms of average of train and test RMSE, average of ranks and overall ranks.

The overall results of the experiments from **Table 3** shows that among all three models of ANFIS (i.e. genfis1, genfis2, and genfis3) and membership functions, ANFIS with subtractive clustering (genfis2) and Gaussian membership function performed best in all classification datasets. Similarly, the result of ANFIS with grid partitioning method which is genfis1 shows that gaussian membership function have performed best and achieved best RMSE as compared to three other shapes i.e., trapezoidal, bell, and triangular because Gaussian membership function draws smooth curve which represents the data points

Dataset	RMSE and rank	trapmf (genfis1)	gbellmf (genfis1)	trimf (genfis1)	gaussmf (genfis1)	gaussmf (genfis2)	gaussmf (genfis3)
D1	Train. RMSE	0.0361	0.0244	0.0102	0.0629	0.0551	0.1552
	Test RMSE	0.6219	0.4111	0.1983	0.2195	0.1234	0.3461
	Rank	6	4	2	3	1	5
D2	Train. RMSE	0.5186	0.4909	0.5090	0.4317	0.1021	0.4162
	Test RMSE	2.2147	1.5992	4.7986	1.1171	1.8323	1.8443
	Rank	5	3	6	1	2	4
D3	Train. RMSE	0.1714	0.1714	0.1769	0.1758	0.1283	0.1872
	Test RMSE	0.2826	0.2432	0.2225	0.2216	0.5201	0.2185
	Rank	5	4	2	1	6	3
D4	Train. RMSE	0.0274	0.0158	0.0555	0.0158	7.7889e-05	0.0297
	Test RMSE	2.1954	0.9718	1.3917	6.6173	0.7233	0.5780
	Rank	5	3	4	6	2	1
D5	Train. RMSE	0.1154	0.1065	0.1065	0.1065	0.1065	0.1442
	Test RMSE	2.1406	1.4741	0.9554	1.4568	0.5469	2.9828
	Rank	5	4	2	3	1	6
D6	Train. RMSE	0.0613	0.0443	2.4258	0.1074	0.4597	0.3135
	Test RMSE	31.4304	206.4092	2.4367	29.1106	2.0372	2.6016
	Rank	5	6	3	4	1	2
	Avg. Train. RMSE	0.1555	0.1422	0.5473	0.1500	0.1420	0.2077
	Avg. Test RMSE	6.4809	35.1848	1.6672	6.4571	0.9639	1.4285
	Avg. Rank	5.1667	4	3.1667	3	02.1667	3.5
	Overall Rank	**6**	**5**	**3**	**2**	**1**	**4**

Table 3. Simulation results of membership function.

effectively with minute differences. Therefore, as a conclusion, this experiment shows that the Gaussian membership function is best option to employ with ANFIS model to solve classification problems.

Table 4 shows the simulation results for the performance of modified ANFIS architecture with other ANFIS models. To evaluate the performance, the proposed modified ANFIS model

Classification problem	ANFIS model	MSE	Accuracy (%)	Trainable parameters	Epochs
Iris	genfis1	0.00395	99.8021	96	200
	genfis2	0.00300	99.8481	96	200
	genfis3	0.02400	98.7956	96	200
	Modified ANFIS model	0.00310	99.8412	32	200
Teaching assistant evaluation	genfis1	0.1863	90.6817	212	200
	genfis2	0.0104	99.4787	212	200
	genfis3	0.1732	91.3388	212	200
	Modified ANFIS model	0.0400	97.9973	52	200
Car evaluation	genfis1	0.0309	98.4547	472	200
	genfis2	0.0164	99.1769	472	200
	genfis3	0.0350	98.2478	472	200
	Modified ANFIS model	0.0337	98.3136	88	200
Breast cancer	genfis1	0.0113	99.4328	5156	200
	genfis2	0.0113	99.4328	5156	200
	genfis3	0.0207	98.9603	5156	200
	Modified ANFIS model	0.0102	99.4901	546	200
Glass identification	genfis1	0.0115	99.4232	11,304	200
	genfis2	0.2113	89.4337	11,304	200
	genfis3	0.0983	95.0858	11,304	200
	Modified ANFIS model	0.0177	99.1177	1064	200
Seeds	genfis1	2.4964×10^{-4}	99.9875	1052	200
	genfis2	261.5473	−12977.36	1052	200
	genfis3	8.8209×10^{-4}	99.9558	1052	200
	Modified ANFIS model	1.170724×10^{-4}	99.9941	156	200

Table 4. Performance analysis of modified ANFIS model and standard ANFIS model.

was compared with other three ANFIS models, genfis1(grid partitioning), genfis2(subtractive clustering) and genfis3(Fuzzy C Mean), respectively. In this experiment, 6 benchmark classification datasets were employed on the ANFIS models. The performance measurement criteria used for evaluation were MSE, percentage of accuracy, number of trainable parameters and number of epochs. The summary of the simulation based on performance of modified ANFIS model and genfis1(grid partitioning), genfis2(subtractive clustering) and genfis3(Fuzzy C Mean), respectively, is shown in **Table 4**.

According to the **Table 4**, it is clearly demonstrated that the proposed modified ANFIS model with ABC optimization algorithm outperformed than standard ANFIS models genfis1 with grid partitioning method while solving classification problems. From the results, it can be concluded that the modified ANFIS model shows better MSE and reasonable percentage of accuracy by training the less number of trainable parameters with equal number of epochs as compare to standard ANFIS model. Hence, removing the fourth layer and reducing the parameters for training not only solved the issue of computational complexity of standard ANFIS model but it is also can be considered as a perfect solution to save the training cost of the model.

5. Conclusion and future work

Based on the results of experiments, this section summarized that as ANFIS accuracy highly depends on the shape and type of membership function. To achieve this, ANFIS with three partitioning method (i) grid-partitioning (genfis1), (ii) subtractive clustering (genfis2), and (iii) fuzzy c-mean clustering (genfis3) have been used to evaluate the membership functions. The simulation results clarify that gaussian membership function is best fit to employ with ANFIS model to solve classification problems. Moreover, as the standard ANFIS architecture holds five layers and uses gradient based learning that increases the complexity of ANFIS architecture when the number of inputs are large, thus, this research proposed a modified ANFIS architecture by reducing the fourth layer as well as reducing the parameters in consequent part, reducing the burden of training the parameters. Apart from that, one of the meta-heuristics optimization algorithm; artificial bee colony (ABC) algorithm have been used to train the parameters of ANFIS model instead of using gradient decent. The performance of proposed modified ANFIS model was compared with another three ANFIS models genfis1(grid partitioning), genfis2(subtractive clustering) and genfis3(Fuzzy C Mean). The result in Table 4.17 proves that that the designed modified ANFIS model can be implemented by researchers while solving classification problems with confidence.

Furthermore, for future work, more experiments can be performed to find out the appropriate membership function for standard ANFIS model and modified ANFIS model to deal with problems other than classification problems such as time series and clustering problems. This research focused on finding out the good membership function for ANFIS model for classification problems. Additionally, this study modified ANFIS model to make it less complex and instead of using typical two pass learning algorithms to train the parameters, this study applied one of the meta-heuristic approach which is ABC optimization algorithm. As meta-heuristic approach provide huge options for optimization algorithms such as ant colony (AC), particle swarm optimization (PSO), firefly algorithm (FFA), cuckoo search algorithm

(CSA), and genetic algorithm (GA) to train the ANFIS parameters. Hence, in future, other meta-heuristic algorithms can be applied in modified ANFIS model for comparison purpose to find out best meta-heuristic algorithm for the modified ANFIS model.

Author details

Mohd Najib Mohd Salleh*, Noureen Talpur and Kashif Hussain Talpur

*Address all correspondence to: najib@uthm.edu.my

Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia

References

[1] Buragohain M, Mahanta C. A novel approach for ANFIS modelling based on full facto-rial design. Applied Soft Computing. Amsterdam, The Netherlands: Elsevier Science Publishers B. V. 2008;**8**(1):609-625

[2] Aghbashlo M et al. An exergetically-sustainable operational condition of a photo-biohydrogen production system optimized using conventional and innovative fuzzy techniques. Renewable Energy. 2016;**94**:605-618

[3] Liu P, Leng W, Fang W. Training ANFIS model with an improved quantum-behaved particle swarm optimization algorithm. Mathematical Problems in Engineering. 2013. Article ID: 595639, 10 pages

[4] Kar S, Das S, Ghosh PK. Applications of neuro fuzzy systems: A brief review and future outline. Applied Soft Computing. Amsterdam, The Netherlands: Elsevier Science Publishers B. V. 2014;**15**:243-259

[5] Das S, Ghosh PK, Kar S. Hypertension diagnosis: A comparative study using fuzzy expert system and neuro fuzzy system. In: IEEE International Conference on Fuzzy Systems (FUZZ), 2013; IEEE; 2013. pg 1-7, 978-1-4799-0020-6, DOI: 10.1109/FUZZ-IEEE.2013.6622434

[6] Salleh MNM, Hussain K. A review of training methods of ANFIS for applications in busi-ness and economics. International Journal of u-and e-Service, Science and Technology, International Conference on Recent Trends in Computer Science and Electronics Engineering (RTCSE'16), Kuala Lumpur, Malaysia. 2016;**9**(7):165-172

[7] Abdulshahed AM, Longstaff AP, Fletcher S. A novel approach for ANFIS modelling based on Grey system theory for thermal error compensation. In: 14th UK Workshop on Computational Intelligence (UKCI), 2014; IEEE; 2014. DOI: 10.1109/UKCI.2014.6930155

[8] Le T, Altman T. A new initialization method for the fuzzy c-means algorithm using fuzzy subtractive clustering. In: Proceedings of International Conference on Information and Knowledge Engineering, 2011 (IKE'11); Las Vegas, USA; 2011

[9] Samat NA, Salleh MNM. A study of data imputation using fuzzy c-means with particle swarm optimization. In: Herawan T, et al., editors. Recent Advances on Soft Computing

and Data Mining: The Second International Conference on Soft Computing and Data Mining (SCDM-2016) Proceedings; August 18-20, 2016; Bandung, Indonesia. Cham: Springer International Publishing; 2017. pp. 91-100

[10] Akrami SA, El-Shafie A, Jaafar O. Improving rainfall forecasting efficiency using modified adaptive neuro-fuzzy inference system (MANFIS). Journal of Water Resources Management. 2013;**27**(9):3507-3523. DOI: 10.1007/s11269-013-0361-9

[11] Soh AC, Kean KY. Reduction of ANFIS-rules based system through K-map minimization for traffic signal controller. In: 12th International Conference on Control, Automation and Systems (ICCAS), 2012; Jeju Island, Korea: IEEE

[12] Orouskhani M, Mansouri. A hybrid method of modified cat swarm optimization and gradient descent algorithm for training ANFIS. International Journal of Computational Intelligence and Applications. 2013;**12**(02):1350007

[13] Karaboga D, Kaya E. An adaptive and hybrid artificial bee colony algorithm (aABC) for ANFIS training. Applied Soft Computing. Amsterdam, The Netherlands: Elsevier Science Publishers B. V. 2016;**49**:423-436

[14] Najafi A. Implementing a PSO-ANFIS model for prediction of viscosity of mixed oils. Journal of Petroleum Science and Technology. 2017;**35**(2):155-162

[15] Ibarra L, Ponce P, Molina A. A new approach to uncertainty description through accomplishment membership functions. Expert Systems with Applications. 2015;**42**(21):7895-7904

[16] Suntae K. A study of fuzzy membership functions for dependence decision-making in security robot system. Neural Computing and Applications. 2017;**28**(1):155-164

[17] Saha S, Bhattacharya S, Konar A. Comparison between type-1 fuzzy membership functions for sign language applications. In: Proceedings of 2016 International Conference on Microelectronics, Computing and Communications (MicroCom 2016), Durgapur, India: IEEE; 2016

[18] Adil O. Comparison between the effects of different types of membership functions on fuzzy logic controller performance. International Journal of Emerging Engineering Research and Technology. 2015;**3**(3):76-83

[19] Gholami V et al. Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers. Journal of Hydrology. 2015;**529**(Part 3):1060-1069

[20] Shrivastava H, Sridharan S. Conception of data preprocessing and partitioning procedure for machine learning algorithm. International Journal of Recent Advances in Engineering & Technology (IJRAET). 2013;**1**(3). ISSN: 2347-2812

[21] Uddin J. Optimization of ANFIS using artificial bee colony algorithm for classification of Malaysian SMEs. In: Recent Advances on Soft Computing and Data Mining: The Second International Conference on Soft Computing and Data Mining (SCDM-2016) Proceedings; August 18-20, 2016; Bandung, Indonesia: Springer International Publishing; 2017. ISSN 2194-5357

Engineering and Technology Applications

Differential Evolution Algorithm in the Construction of Interpretable Classification Models

Rafael Rivera-Lopez and Juana Canul-Reich

Abstract

In this chapter, the application of a differential evolution-based approach to induce oblique decision trees (DTs) is described. This type of decision trees uses a linear combination of attributes to build oblique hyperplanes dividing the instance space. Oblique decision trees are more compact and accurate than the traditional univariate decision trees. On the other hand, as differential evolution (DE) is an efficient evolutionary algorithm (EA) designed to solve optimization problems with real-valued parameters, and since finding an optimal hyperplane is a hard computing task, this metaheuristic (MH) is chosen to conduct an intelligent search of a near-optimal solution. Two methods are described in this chapter: one implementing a recursive partitioning strategy to find the most suitable oblique hyperplane of each internal node of a decision tree, and the other conducting a global search of a near-optimal oblique decision tree. A statistical analysis of the experimental results suggests that these methods show better performance as decision tree induction procedures in comparison with other supervised learning approaches.

Keywords: machine learning, classification, evolutionary algorithms

1. Introduction

Knowledge discovery refers to the process of nontrivial extraction of potentially useful and previously unknown information from a dataset [1]. Within the stages of this process, data mining stands out since it allows analyzing the data and producing models for their representation. In particular, machine learning provides data mining with useful procedures to build these models, since many of the techniques aimed at information discovery are based on inductive learning. Decision trees (DTs), artificial neural networks (ANN), and support vector

machines (SVMs), as well as clustering methods, have been widely used to build predictive models. The ability to track and evaluate every step in the information extraction process is one of the most crucial factors for relying on the models gained from data mining methods [2]. In particular, DTs are classification models characterized by their high levels of comprehensibility and robustness. Knowledge learned via a DT is understandable due to its graphical representation [3], and also DTs can handle noise or data with missing values and make correct predictions [4].

On the other hand, soft-computing-based approaches have been widely used to solve complex problems in almost all areas of science and technology. These approaches try to imitate the process of human reasoning when solving a problem with the objective of obtaining acceptable results in a reasonable time. For the case of data mining, soft computing techniques such as ANN, metaheuristics (MHs), fuzzy logic methods, and other approaches have been used as tools to solve the data mining challenges. In particular, an MH is a general algorithmic template based on intelligent processes and behaviors observed in both nature and other disciplines [5]. Evolutionary algorithms (EAs) are one type of MH that have been successfully applied for providing near-optimal solutions for many computationally complex problems in almost all areas of science and technology. The effectiveness of EAs is due to two factors: (1) they combine a clever exploration of the search space to identify promising areas and (2) they perform an efficient exploitation of these areas aiming to improve the known solution or solutions. EAs are inspired by evolutionary theories that synthesize the Darwinian evolution through natural selection with the Mendelian genetic inheritance. In particular, differential evolution (DE) algorithm is an EA designed for solving optimization problems with variables in continuous domains that, instead of implementing traditional crossover and mutation operators, it applies a linear combination of several randomly selected candidate solutions to produce a new solution [6].

MHs have been previously applied to build DTs, and there exist several surveys that describe their implementation [7–11]. Some approaches apply a recursive partitioning strategy in which an MH finds a near-optimal test condition for each internal node of a DT; however, the approach most commonly used is to perform a global search in the solution space with the aim of finding near-optimal DTs. Since DE is one of the most powerful EA to solve real-valued optimization problems, and the task of finding a near-optimal oblique hyperplane with real-valued coefficients is an optimization problem in a continuous space, in this chapter, two DE-based methods to induce oblique DTs are described: one implementing a recursive partitioning strategy to find the most suitable oblique hyperplane of each internal node of a decision tree, and the other conducting a global search of a near-optimal oblique decision tree. A statistical analysis of the experimental results suggests that these methods show better performance as decision tree induction procedures in comparison with other supervised learning approaches.

The rest of this chapter is organized as follows: Section 2 provides a set of basic definitions about DTs and the DE algorithm. The induction of oblique DTs by means of MH-based approaches is described in Section 3. The constituent elements of the DE-based methods described in this chapter is discussed in Section 4, and the experimental results are discussed in Section 5. Finally, Section 6 describes the conclusion and the future work.

2. Background

Machine learning methods are an essential tool in emerging disciplines such as data science [12] and business intelligence [13] since they provide efficient predictive models constructed from the data previously collected. DTs, ANN, and SVMs, as well as clustering methods, have been widely used to build these models. A DT is a hierarchical model using an ordered sequence of decisions to predict the class membership of new unclassified instances. An ANN consists of many nonlinear elements connected by links associated with weighted variables operating in parallel [14], in which learning is performed iteratively as the network processes the training instances, trying to simulate the way a human being learns from previous experiences. Finally, one SVM finds the hyperplane that best separates the training instances into two different classes using a set of functions called kernels. The optimal hyperplane is described with a combination of entry points known as support vectors [15].

A DT is an acyclic connected graph with a single root node used as one classification model induced through a set of training instances. A DT contains zero or more internal nodes and one or more leaf nodes [16]. Each internal node evaluates a test condition consisting of a combination of one or more attributes of the dataset, and each leaf node has a class label. The arcs joining an internal node with their successor nodes are labeled with the possible outcomes of its test condition. Each DT branch represents a sequence of decisions made by the model to determine the class membership of a new unclassified instance. The DT induction (DTI) process commonly implements a recursive partition strategy. In each stage of this process, the most appropriate test condition to split the training set is selected according to some partition criterion. As a result of evaluating the training instances with this test condition, two or more instances subsets are created which are assigned to the successor nodes of the current internal node. This process is recursively applied until a stop criterion is reached. If the number of attributes used in the test conditions of the tree internal nodes is regarded, two types of DT can be constructed: axis-parallel or multivariate DTs. An axis-parallel DT is a univariate DT that evaluates a single attribute in each test condition to split the training set. On the other hand, oblique DTs and nonlinear DTs are multivariate DTs in which a linear combination and a nonlinear composition of attributes are utilized in the test conditions of a DT, respectively. Multivariate DTs commonly show better performance, and they are smaller than univariate DTs, but they require more computational effort to induce them. In particular, an oblique hyperplane divides the instance space into two halfspaces, and it is defined as follows:

$$\sum_{j=1}^{d} h_j x_j + b > 0 \qquad (1)$$

where d is the number of attributes in the dataset, x_j is the value of the j-th attribute, h_j is a real-valued coefficient in the hyperplane, and b represents the independent term of the hyperplane. **Figure 1** shows an axis-parallel DT induced from the iris dataset [17] using the J48 method [18], and **Figure 2** shows a near-optimal oblique DT constructed from the same dataset by the DE-based method implementing a global search strategy. Iris dataset has four attributes, three
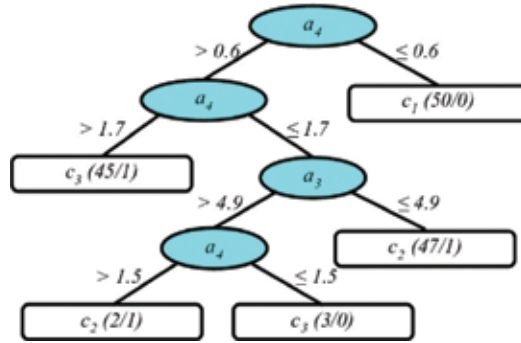
Figure 1. An axis-parallel DT induced from the iris dataset.
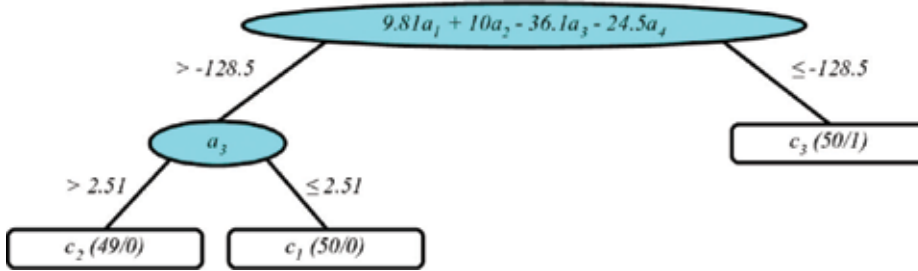
Figure 2. An oblique DT induced from the iris dataset.

class labels, and 150 instances. It is clear that the oblique DT is more compact and more accurate than its axis-parallel version, but it has been proved that to find the oblique hyperplane that minimizes the number of misclassified instances both above and below is an NP-hard problem [19].

On the other hand, MHs are general algorithmic templates that can be easily adapted to solve almost all optimization problems [20]. MHs are nature-inspired procedures using stochastic components to find a near-optimal solution and have several parameters that need to be fitted to the specific problem [21]. In accordance with the number of candidate solutions used in its search procedure, MHs have been grouped in single-solution-based MHs and population-based MHs [22]. Single-solution-based MHs implement intelligent search procedures that iteratively replace a candidate solution with a neighboring solution with the aim of reaching a near-optimal solution. Simulated annealing (SA) and Tabu search (TS) are two well-known single-solution-based MHs. Population-based MHs use a group of candidate solutions in each step of their iterative process. The most commonly used population-based MHs are related to EAs and Swarm intelligence (SI) methods. Genetic algorithms (GA), genetic programming (GP), evolutionary strategies (ES) and DE are the most prominent EAs, and ant colony optimization (ACO) and particle swarm optimization (PSO) are examples of SI methods.

In particular, DE is an effective EA designed to solve optimization problems with real-valued parameters [6]. DE evolves a *population* $X = \{x^1,, x^2, ...,, x^{NP}\}$ of NP *chromosomes* by applying

mutation, crossover, and selection operators with the aim to reach a near-optimal solution. Several DE variants differing in the implementation of the mutation and crossover operators have been described in existing literature. In this chapter, the standard DE algorithm, named `DE/rand/1/bin` in agreement with the nomenclature adopted to refer DE variants, is used as a procedure to find a near-optimal solution. At each iteration of this evolutionary process, known as a *generation*, a new population of chromosomes is generated from the previous one. For each $i \in \{1, \dots, NP\}$ in the g-th generation, x^i is taken from the X_{g-1} population, and it is used to build a new vector u^i by applying the mutation and crossover operators. Vectors x^i and u^i are known as the *target vector* and the *trial vector*, respectively. To build a new chromosome, instead of implementing a traditional mutation operator, DE first applies a linear combination of several chromosomes randomly chosen from the current population ($x^{r_1}, x^{r_2},$ and x^{r_3}) to construct a mutated vector $v^i = x^{r_1} + F(x^{r_2} - x^{r_3})$, where F is a user-specified value representing a scale factor applied to control the differential variation. Next, the crossover operator determines each parameter in u^i from either x^i or v^i, based on a stochastic decision. If a random value is less than a crossover factor (CF), the j-th parameter value of u^i is taken from v^i, otherwise its value is $u^i_j = x^i_j$. Finally, a one-to-one tournament is applied to determine which vector, between x^i and u^i, is selected as a member of the new population X_g. **Figure 3** shows a scheme of the application of the DE operators to build a new chromosome for the next population.

DE has been used in conjunction with several machine learning techniques to implement classification methods [23–27]. It has been mainly applied to optimize the parameters of classification methods or to conduct preprocessing tasks in a data mining process. DE has several advantages in comparison with other MHs, and since mutation operator is based on a linear combination of several randomly chosen individuals, DE exhibits a good trade-off between its exploitation and exploration skills [28]. On the other hand, although DE requires the definition of a smaller number of parameters compared to other MHs, its performance is sensitive to the values selected for CR, F, and NP.
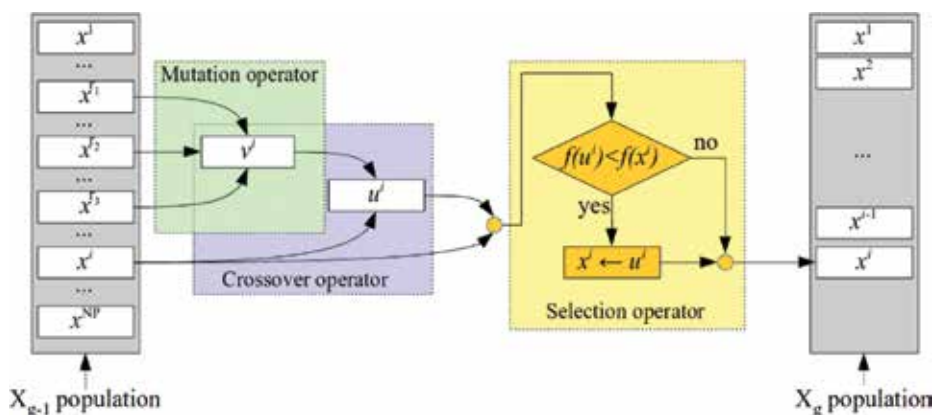
Figure 3. DE operators applied to build a new chromosome for the next population.

3. Induction of oblique decision trees using metaheuristics

Several MHs have been used to induce DTs with methods implementing a recursive partitioning strategy. One timeline of these methods is shown in **Figure 4**. Single-solution-based MHs such as SA and TS have been used to induce DTs through this strategy. SA is applied in the simulated annealing of decision trees (SADT) method [29] that iteratively perturbs one randomly selected coefficient to build a new hyperplane, and in one variant of the oblique classifier 1 (OC1) system [30], named OC1-SA [31], that disturbs simultaneously several coefficients of the best axis-parallel hyperplane found by the OC1 algorithm. TS is used in the linear discriminant and TS (LDTS) method [32] and in the linear discrete support vector DT with TS (LDSDT$_{TS}$) method [33]. Furthermore, EAs such as ES, GA, and DE also have been applied to build an oblique DT through this strategy. The OC1-ES algorithm [31] and the multimembered ES oblique DT (MESODT) method [34] obtain a near-optimal hyperplane using the $(1+1)$-ES and the (μ, λ)-ES, respectively. Furthermore, GA evolves a population of hyperplanes encoded: (1) with a binary chromosome in the binary tree-GA (BTGA) algorithm [35] and in the HereBoy for DT (HBDT) method [36] and (2) with a real-valued chromosome in the OC1-GA algorithm [31] and in the procedures described by Krętowski [37], and by Pangilinan and Janssens [38]. Finally, DE is applied in an OC1 variant named OC1-DE algorithm [39].

On the other hand, several MH-based approaches implementing a global search strategy have been described in the existing literature. GA evolves a population of variable-length chromosomes in the generalized decision tree inducer (GDTI) method [40] and in the evolutionary full tree induction (EFTI) method [41]. Other GA-based approaches such as the Global EA for oblique DTI (GEA-ODT) procedure [42, 43] and the tree analysis with randomly generated
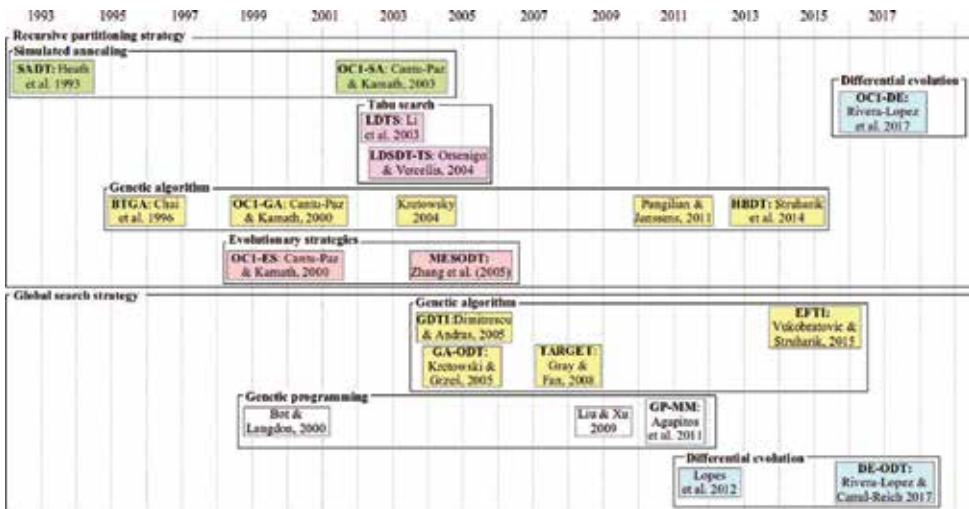
Figure 4. Timeline of the MH-based approaches to induce oblique DTs.

and evolved trees (TARGET) algorithm [44] use trees as chromosomes. Furthermore, the standard GP is applied by Liu and Xu [45], the strongly typed GP is used by Bot and Langdon [46, 47], and the grammar-based GP is utilized in the GP with margin maximization (GP-MM) method [48]. Finally, DE is implemented in the perceptron DT (PDT) method [49, 50] and in the DE for inducing oblique DTs (DE-ODT) method [51].

4. DE-based methods to induce oblique decision trees

In this chapter, two methods to induce an oblique DT using the DE/rand/1/bin algorithm are described. The first method, named OC1-DE, is similar to the OC1 system and its variants, but it applies DE to find a near-optimal hyperplane at each internal node of an oblique DT [39]. The second one, named DE-ODT method, implements a global search strategy to induce oblique DTs [51].

4.1. OC1-DE method to search near-optimal oblique hyperplanes

The OC1-DE method is based on the OC1 system [30] and the OC1-GA procedure [31]. The OC1 system applies a two-step process to find a better hyperplane. First, it finds the best axis-parallel hyperplane splitting the instance set. Next, it applies two procedures to disturb the hyperplane coefficients:

- *Sequential perturbation*: This is a deterministic rule that adjusts the hyperplane coefficients, taking one at a time and looking for its optimal value.

- *Random vector perturbation*: When the sequential perturbation reaches a local optimum, a random vector is added to the current hyperplane with the aim of looking elsewhere in the solutions space.

Finally, this procedure returns as the best hyperplane to the one selected between the best-perturbed hyperplane and the best axis-parallel hyperplane. OC1 uses several criteria to evaluate the quality of the candidate hyperplanes such as information gain [52] and three criteria introduced by Heath [19]: max minority, minority sum, and sum of impurities. Induced DT is pruned by removing sub-trees whose impurity value is less than a predefined threshold value. An improved OC1 version [53] uses the elements defined in the CART method [54]: the Gini index and the twoing rule as splitting criteria and the cost-complexity pruning method.

On the other hand, the OC1-GA method is an OC1 variant encoding the hyperplane coefficients in a real-valued chromosome. First, the axis-parallel hyperplane that best splits the training instances is obtained. This hyperplane is copied to 10% of the initial population and the remaining hyperplanes are randomly created. Then, OC1-GA evolves this population to find a near-optimal hyperplane evaluating its quality through the twoing rule. Oblique DT is then induced in a recursive partitioning strategy, and it is pruned using the cost-complexity pruning method.

The DE implementation to find a near-optimal hyperplane at each internal node of an oblique DT is shown in the **Algorithm 1**. First, the axis-parallel hyperplane that best splits a set of training instances is obtained (line 1). It is copied to 10% of the initial population, as is proposed in [31], and the remaining hyperplanes are randomly created (line 2). Each random hyperplane is constructed considering that almost two instances with different class labels are separated by the hyperplane. Next, this population is evolved through several generations using the DE operators (lines 3–19), and the best hyperplane in the population is selected (line 20). Finally, the OC1-DE algorithm returns the hyperplane selected between the best axis-parallel hyperplane and the best oblique hyperplane produced by DE (line 21).

The hyperplane returned by the OC1-DE is used as the test condition of a new internal node that is added in an oblique DT. This hyperplane is used to split the training instances into two subsets. The OC1-DE method is recursively applied using each subset until a leaf node is created as all instances in the subset have the same class label or a threshold value of unclassified instances is reached. The quality of the hyperplane is obtained using the twoing rule. Finally, a pruning procedure is applied in order to reduce the overfitting of DT produced and to improve its predictive power.

Algorithm 1 Hyperplane selection using the OC1-DE algorithm.

1: $h^{AP} \leftarrow$ Best axis-parallel hyperplane for the training instances
2: $X_0 \leftarrow$ Initial population of random hyperplanes with 10% of copies of h^{AP}
3: **for** g in $[1,\dots,\text{it}]$ **do**
4:     $X_g \leftarrow \emptyset$
5:     **for each** $i \in \{1,\dots,NP\}$ **do**
6:         $h^i \leftarrow$ The trial hyperplane from X_{g-1}
7:         $\{h^{r1}, h^{r2}, h^{r3}\} \leftarrow$ Randomly selected hyperplanes from X_{g-1}
8:         $v^i \leftarrow h^{r1} + F(h^{r2} - h^{r3})$ ▷ DE mutation operator
9:         **for each** $j \in \{1,\dots,|h^i|\}$ **do**
10:            $r \leftarrow$ A uniformly distributed random number from $[0,1]$
11:            $l \leftarrow$ A randomly chosen index from $\{1,\dots,|x^i|\}$
12:            $u^i_j \leftarrow \begin{cases} v^i_j & \text{if } r \leq CR \text{ or } j = l, \\ h^i_j & \text{otherwise.} \end{cases}$ ▷ DE crossover operator
13:        **end for**
14:        **if** $fitness(u^i) < fitness(h^i)$ **then**   ▷ DE selection operator
15:            $h^i \leftarrow u^i$
16:        **end if**
17:        $X_g \leftarrow X_g \cup \{h^i\}$
18:    **end for**
19: **end for**
20: $h^{best} \leftarrow$ Best hyperplane in X_g
21: **return** $\begin{cases} h^{AP} & \text{if } fitness(h^{AP}) < fitness(h^{best}), \\ h^{best} & \text{otherwise.} \end{cases}$

4.2. DE-ODT method to induce oblique decision trees

The DE-ODT method implements a global search strategy in which the DE algorithm is applied to find a near-optimal oblique DT, where each real-valued chromosome encodes only a feasible oblique DT.

4.2.1. Linear representation of oblique decision trees

In the DE-ODT method, each chromosome represents the internal nodes of a complete binary oblique DT stored in a fixed-length real-valued vector (**Figure 5**). This vector encodes the set of hyperplanes used as test conditions of the oblique DT. Vector size is determined using both the number of attributes and the number of class labels of the training set whose model is induced. Since each internal node of an oblique DT has a hyperplane as its test condition, the size of the real-valued vector x^i used to encode each i-th candidate solution in the population is fixed as $n_e(d + 1)$, where n_e is the estimated number of internal nodes of a complete binary oblique DT. Considering that: (1) an oblique DT is more compact than its univariate version and since (2) the DT size is related to the structure of the training set, the DE-ODT method determines the value of n_e based on both the number of attributes and the number of class labels (s) in it.

If the number of internal nodes of a complete binary DT with height H is $2^{H-1} - 1$ and the number of leaf nodes of the same DT is 2^{H-1}, two heights can be obtained: $H_i = \lceil \log_2(d + 1) + 1 \rceil$, and $H_l = \lceil \log_2(s) + 1 \rceil$. Using these equations, n_e is determined as follows:

$$n_e = 2^{\max(H_i, H_l) - 1} - 1, \tag{2}$$

and, the size of the real-valued parameter vector representing a sequence of n_e hyperplanes for a training set with d attributes is computed as follows:

$$n = n_e(d + 1). \tag{3}$$

As an example, if a hypothetical dataset with three numerical attributes and three class labels is used to induce an oblique DT, then $d = 3$ and $s = 3$. In this case, $H_i = \lceil \log_2(4) + 1 \rceil = 3$ and $H_l = \lceil \log_2(3) + 1 \rceil = 3$. Finally, $n_e = 2^{\max\{3,3\}} - 1 = 7$. This implies that the oblique DT could have seven internal nodes. Finally, one chromosome representing a candidate solution in the evolutionary process of this problem has 28 real-valued parameters.
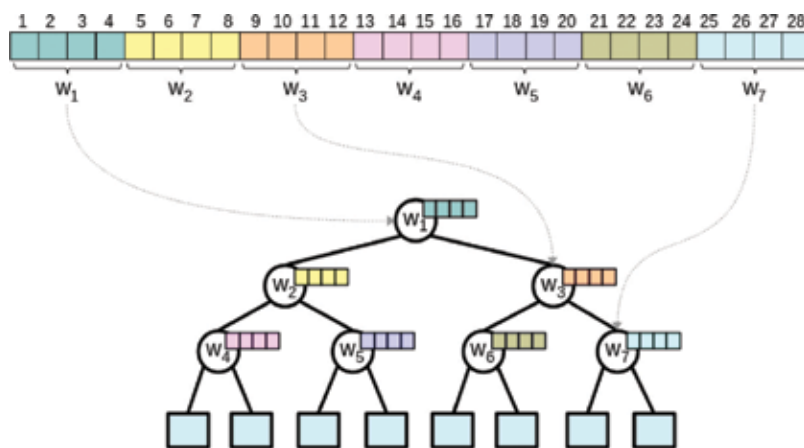
Figure 5. Linear encoding scheme for the internal nodes of a complete binary oblique tree.

4.2.2. Induction of feasible oblique decision trees

The DE-ODT method applies the following steps to map an oblique DT from a chromosome x^i of the population:

1. **Hyperplanes construction:** x^i is used to build the vector w^i representing the sequence of candidate hyperplanes utilized in the internal nodes of a partial DT. Since the values of x^i represent the hyperplane coefficients contained in these nodes, the following criterion applies: Values $\{x^i_1, ..., x^i_{d+1}\}$ are assigned to the hyperplane h^1, the values $\{x^i_{d+2}, ..., x^i_{2d+2}\}$ are assigned to the hyperplane h^2, and so on. For each $j \in \{1, ..., n_e\}$, and for each $k \in \{1, ..., d+1\}$, the k-th coefficient of h^j is designed as follows

$$h^j_k = x^i_{(j-1)(d+1)+k}. \tag{4}$$

These hyperplanes are assigned to the elements of w^i: h^1 is assigned to w^i_1, h^2 is assigned to w^i_2, and so on. **Figure 6** shows an example of the construction of a set of hyperplanes from one chromosome for the hypothetical dataset previously described. Once w^i is completed, it is used to create a partial DT with only internal nodes.

2. **Partial oblique decision tree construction:** w^i is used to create the partial tree (pT^i). First, the element in the initial location of w^i is used as the root node of pT^i. Next, the remaining elements of w^i are inserted in pT^i as successor nodes of those previously added so that each new level of the tree is completed before placing new nodes at the next level, in a similar way to the breadth-first search strategy. **Figure 7** shows an example of the construction of a partial oblique DT from w^i. In this figure, it can be observed that w^i_1 is selected as the tree root node, w^i_2 and w^i_3 are placed as the successor nodes of w^i_1, w^i_4 and w^i_5 are designed as the successor nodes of w^i_2, and so on.

3. **Oblique decision tree completion:** The final stage of the mapping scheme adds leaf nodes in pT^i using the training set. In this stage, one instance set is assigned to a node (the
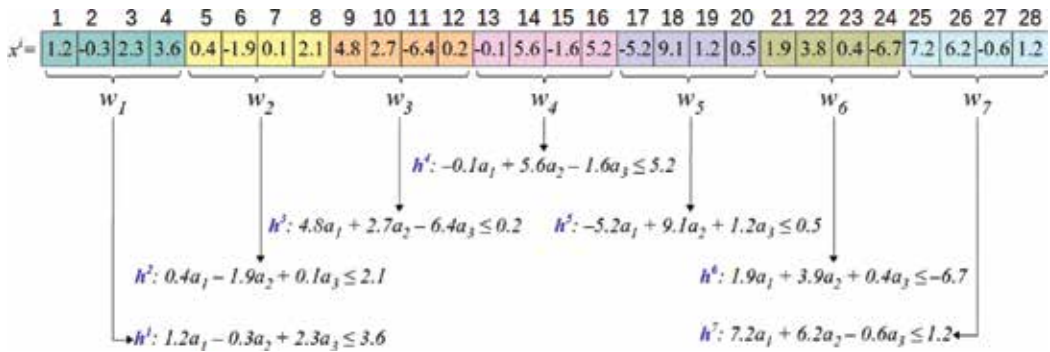
Figure 6. Construction of a set of hyperplanes from x^i.
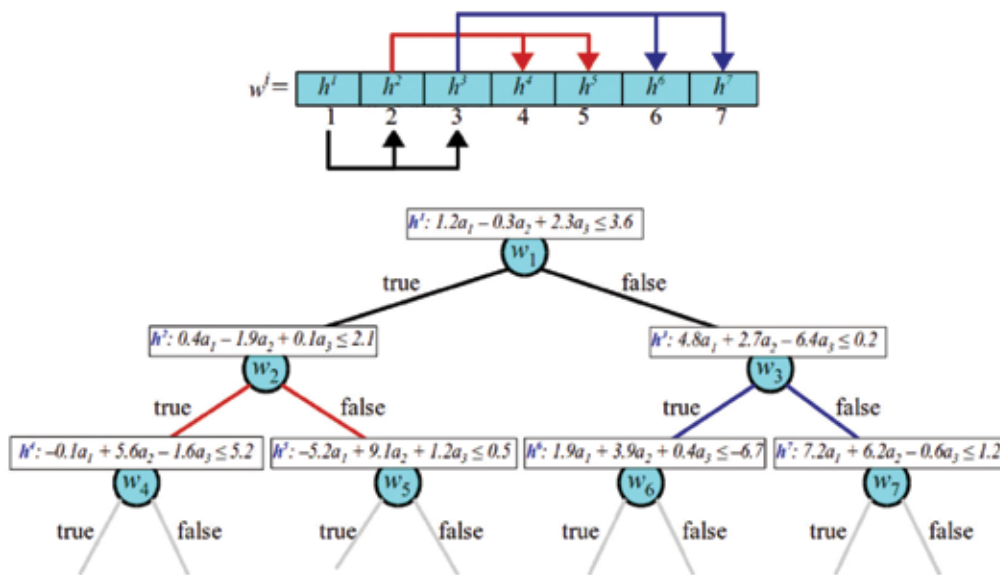
Figure 7. Construction of a partial oblique DT with only internal nodes.

complete training set for the root node of the tree), and it is labeled as an internal node. To evaluate each instance in this set using the hyperplane associated to the internal node, two instances subsets are created, and they are assigned to the successor nodes of this node. This assignment is repeated for each node of the partial DT. If the internal node is located at the end of a branch of the DT, then two leaf nodes are created, and they are designated as successor nodes of this node. The instances subsets created are assigned to these leaf nodes. On the other hand, if all instances in the set assigned to the internal node have the same class label, it is labeled as a leaf node and its successor nodes are removed, if they exist. **Figure 8** shows an example of this tree-completion procedure. **Figure 8** shows that all the instances assigned to w_3 and w_5 have the same class label, so they are designated as leaf nodes, and the successor nodes of w_3 are removed from the tree. On the other hand, since w_4 is the ending node of a branch, its instance set is split using its hyperplane, the instances subsets produced are assigned to two new leaf nodes, and their majoritarian classes are assigned as their class labels. It can be observed that this tree has three internal nodes and four leaf nodes.

4.2.3. General structure of the DE-ODT method

The **Algorithm 2** shows the structure of the DE-ODT method described in this chapter. This procedure requires to identify the training set used to induce an oblique DT, as well as the three control parameters applied by the DE algorithm (CR, F, and NP) and the threshold value (τ) used to determine if a node is labeled as a leaf node. First, the DE-ODT method gets the attributes vector (a), the vector of class labels (c), and the instance set (ι) from the dataset whose
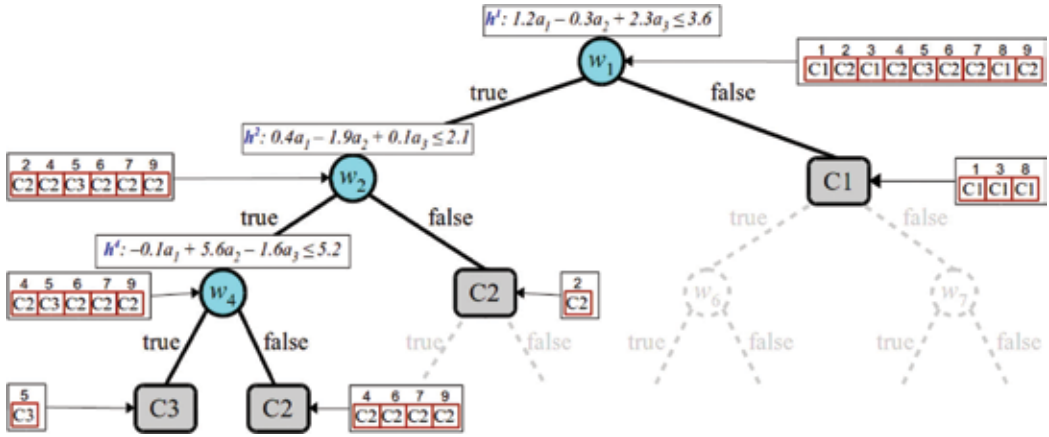
Figure 8. Completion of an oblique DT using pT^i and the training instances.

model must be built (line 1). Next, the value of d and n are computed (lines 2–4). Then, the DE algorithm evolves a population of real-valued chromosomes encoding oblique DTs. DE selects the best candidate solution x^{best} in the last population as the result of its evolutionary process (line 5). After that, a near-optimal oblique DT is constructed applying the procedures described in the previous paragraphs (lines 6–8). Since the DE-ODT method uses an a priori definition of the size of the chromosome, it is possible that some leaf nodes in the DT do not meet the following conditions: that the size of its instances subset is less than τ or that all the instances in the subset belong to the same class. The DE-ODT method refines the DT by replacing nonoptimal leaf nodes with sub-trees (line 9). Finally, the oblique DT is pruned to reduce the possible overfitting generated by applying this refinement (line 10).

This procedure allows inducing feasible oblique DTs with a different number of nodes, although they are represented with a fixed-length parameter vector.

Algorithm 2 General structure of DE-ODT method.

1: $(a,c,\iota) \leftarrow$ READTRAININGSET($trainingSet$)
2: $d \leftarrow |a|$ ▷ Number of attributes
3: $n_e \leftarrow$ Number of estimated internal nodes computed using (1.2)
4: $n \leftarrow n_e(d+1)$ ▷ Size of the real-valued chromosome
5: $x^{best} \leftarrow$ DIFFERENTIALEVOLUTION(CR, F, NP, n)
6: $w \leftarrow$ HYPERPLANESCONSTRUCTION(x^{best})   ▷ Nodes vector
7: $pT \leftarrow$ PARTIALDTCONSTRUCTION(w)   ▷ DT with only internal nodes
8: $T \leftarrow$ DTCOMPLETION(pT, ι, τ)   ▷ Complete DT
9: $T \leftarrow$ DTREFINEMENT(T, τ)   ▷ Refined DT
10: $T \leftarrow$ DTPRUNING(T, τ)   ▷ Pruned DT

---

5. Experimental study

In this chapter, the experimental study carried out to analyze the performance of the DE-based methods for DTI is detailed. First, a description of the datasets used in this study, as well as the

definition of the parameters of each method, are given. Then, both the model validation technique used in the experiments and the statistical tests applied to evaluate the results obtained are outlined. Finally, a discussion about the performance of the DE-based methods is provided.

5.1. Experimental setup

A benchmark of 20 datasets chosen from the UCI machine learning repository [55] is used to carry out the experimental study. These datasets have been selected as their attributes are numerical, their instances are classified into two or more classes, and most of them are imbalanced datasets. **Table 1** shows the description of these datasets. To ensure that the comparison of the results achieved by the DE variants with those produced by other approaches is not affected by the treatment of the data, all datasets used in this study do not have missing values. Also, the data are not preprocessed, filtered, or normalized, that is, they are used as they are obtained from the UCI repository.

The DE-based methods are implemented in the Java language using the JMetal library [56]. The mutation scale factor is linearly decreased from 0.5 to 0.1 as the evolutionary process progresses, and the crossover rate is fixed at 0.9. The decrement in the F value allows more exploration of search space at the beginning of the evolutionary process, and with the passage of the generations, it tries to make a better exploitation of promising areas of this space [57]. The population size is adjusted to $5n$, with 250 and 500 chromosomes as lower and upper bound, respectively. These bounds are used to ensure that the population is not so small as not to allow a reasonable exploration of the search space and is not so large as to impact the runtime of the algorithm. Furthermore, the fitness function used in the DE-ODT method computes the training accuracy of each DT in population, and the twoing rule is used as fitness value in the OC1-DE method. The best oblique DT induced by these methods is pruned using the error-based pruning (EBP) approach [58]. Finally, the threshold value used to determine

Dataset	Instances	Attributes	Classes	Class distribution	Dataset	Instances	Attributes	Classes	Class distribution
Glass	214	9	7	70|76|17|0|13|9|29	Diabetes	768	8	2	500|268
Balance-scale	625	4	3	288|49|288	Heart-statlog	270	13	2	150|120
Iris	150	4	3	50 instances per class	Australian	690	14	2	307|383
Ionosphere	351	34	2	126|225	Wine	178	13	3	59|71|48
Sonar	208	60	2	97|111	Vehicle	846	18	4	212|217|218|199
Liver-disorder	345	6	2	145|200	Page-blocks	5473	10	5	4913|329|28|88|115
Blood-t	748	4	2	570|178	Breast-tissue-6	106	9	6	22|21|14|15|16|18
Movement-libras	360	90	15	24 instances per class	Parkinsons	195	22	2	48|147
Seeds	210	6	3	70 instances per class	Segment	2310	19	7	330 instances per class
Ecoli	336	7	8	143|77|52|35|20|5|2|2	Spambase	4601	57	2	1813|2788

Table 1. Description of datasets used in the experiments.

whether a node should be labeled as one leaf node is set to be two instances, and the DT size is defined as the number of leaf nodes of the oblique DT.

In this study, a repeated stratified 10-fold cross-validation (CV) procedure is used to estimate the predictive performance of the DE-based methods, and the Friedman test [59] is applied to carry out a statistical analysis of the results produced by these methods as compared to them with those obtained by other classification methods. This nonparametric statistical test evaluates the statistical significance of the experimental results through computing the p-value without making any assumptions about the distribution of the analyzed data. This p-value is used to accept or to reject the null hypothesis H_0 of the experiment which holds that the performance of the compared algorithms does not present significant differences. If the p-value does not exceed a predefined significance level, H_0 is rejected and the Bergmann-Hommel (BH) post hoc test [60] is conducted to detect the differences between all existing pairs of algorithms. These statistical tests are applied using the *scmamp* R library [61].

Dataset	J48		sCART		OC1-DE		DE-ODT	
Glass	67.62	(4)	71.26	(2)	**71.31**	(1)	68.97	(3)
Diabetes	74.49	(3)	74.56	(2)	73.37	(4)	**75.79**	(1)
Balance-scale	77.82	(4)	78.74	(3)	**93.92**	(1)	91.97	(2)
Heart-statlog	78.15	(2)	78.07	(3)	74.11	(4)	**81.11**	(1)
Iris	94.73	(3)	94.20	(4)	96.73	(2)	**97.17**	(1)
Australian	84.35	(4)	85.19	(2.5)	85.19	(2.5)	**85.61**	(1)
Ionosphere	89.74	(3)	88.86	(4)	91.11	(2)	**92.28**	(1)
Wine	**93.20**	(1)	89.49	(4)	92.58	(2)	91.88	(3)
Sonar	73.61	(3)	70.67	(4)	77.65	(2)	**79.34**	(1)
Vehicle	72.28	(2)	69.91	(4)	**72.32**	(1)	71.33	(3)
Liver-disorders	65.83	(4)	66.64	(3)	67.63	(2)	**71.16**	(1)
Page-blocks	96.99	(2)	96.76	(4)	96.88	(3)	**97.07**	(1)
Blood-t	78.20	(2)	77.86	(3)	76.35	(4)	**78.70**	(1)
Breast-tissue-6	34.81	(3)	32.45	(4)	34.91	(2)	**38.85**	(1)
Movement-libras	69.31	(2)	65.64	(3)	**75.11**	(1)	55.63	(4)
Parkinsons	84.72	(4)	86.31	(3)	**87.95**	(1)	86.43	(2)
Seeds	90.90	(3.5)	90.90	(3.5)	**93.76**	(1)	91.79	(2)
Segment	**96.79**	(1)	95.83	(3)	95.93	(2)	94.78	(4)
Ecoli	82.83	(4)	83.15	(3)	83.51	(2)	**84.72**	(1)
Spambase	92.68	(2)	92.35	(3)	92.19	(4)	**93.94**	(1)
Average ranking		2.825		3.250		2.175		1.750

Table 2. Average accuracies obtained by the DTI algorithms and the DE-based methods.

The results obtained with the DE-based methods are compared with those achieved by several supervised learning methods available on the WEKA data mining software [62]. First, the accuracy and the size of the DTs got by these algorithms are compared with those obtained by the J48 method [63] and by the SimpleCART (sCART) [54] procedure. Next, the accuracy of the DTs constructed with the DE-based procedures is compared with those achieved using the following classification methods: Naïve Bayes (NB) [64], multilayer perceptron (MLP) [65], radial basis function neural network (RBF-NN) [66], and random forest (RF) [67].

5.2. Comparison with DTI methods

Table 2 and **Figure 9** show the average accuracies of the DTs induced by the DTI algorithms as well as those achieved by the OC1-DE method. In **Table 2**, the best result for each dataset is highlighted with bold numbers, and the numbers in parentheses refer to the ranking reached by each method for each dataset. The last row in this table indicates the average ranking of each method. It is observed that the DE-based methods produce better results than those generated by the other DTI algorithms.

A statistical test of the experimental results is conducted to evaluate the performance of the DE-based methods. First, the Friedman test is run and its resulting statistic value is 16.197 for four methods and 20 datasets, which has a p-value of 1.033×10^{-3}. When evaluating this p-value with a significance level of 5%, H_0 is rejected. Next, the BH post hoc test is applied to
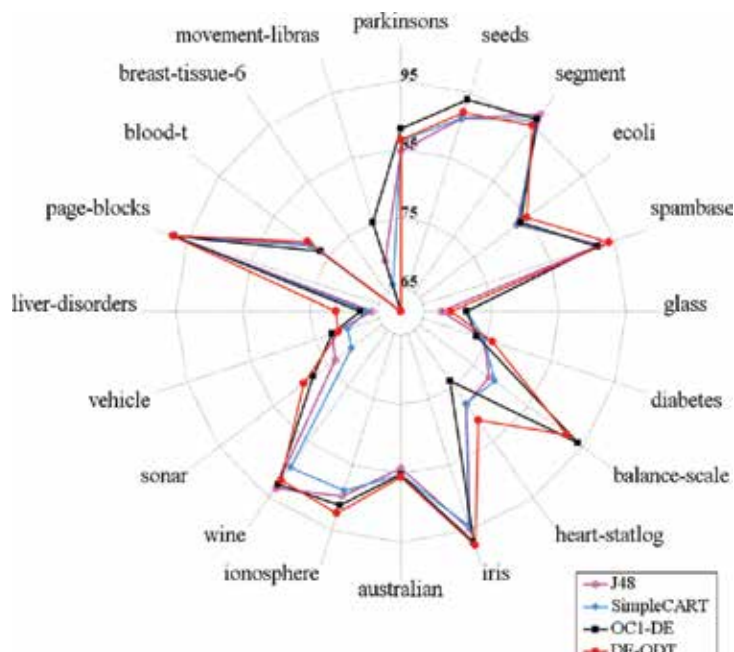
Figure 9. Graphical comparison of the average accuracies obtained by the DTI algorithms and the DE-based methods.

Method	AR	OC1-DE		DE-ODT	
		Unadjusted	BH	Unadjusted	BH
J48	2.825	1.1134e-01	1.1134e-01	8.4584e-03	**2.5375e-02**
sCART	3.250	8.4584e-03	**2.5375e-02**	2.3856e-04	**1.4131e-03**
OC1-DE	2.175	—	—	2.9786e-01	5.9572e-01
DE-ODT	1.750	2.9786e-01	5.9572e-01	—	—

Table 3. The *p*-values for multiple comparisons among DTI algorithms and the DE-based methods.
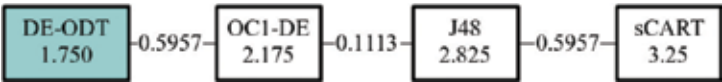
Figure 10. The *p*-value graph of the DTI algorithms and the DE-based methods.

Dataset	J48		sCART		OC1-DE		DE-ODT	
Glass	23.58	(4)	**8.00**	(1)	21.61	(3)	11.08	(2)
Diabetes	22.20	(3)	**3.00**	(1)	41.55	(4)	14.77	(2)
Balance-scale	41.60	(4)	13.00	(2)	15.24	(3)	**5.01**	(1)
Heart-statlog	17.82	(4)	16.00	(2)	17.43	(3)	**7.23**	(1)
Iris	4.64	(3)	5.00	(4)	**3.00**	(1)	3.37	(2)
Australian	25.75	(4)	**5.00**	(1)	21.90	(3)	15.64	(2)
Ionosphere	13.87	(4)	**3.00**	(1)	7.20	(2)	7.73	(3)
Wine	5.30	(3)	5.00	(2)	5.48	(4)	**4.71**	(1)
Sonar	14.45	(4)	10.00	(2)	10.24	(3)	**6.13**	(1)
Vehicle	69.50	(3)	80.00	(4)	56.74	(2)	**44.25**	(1)
Liver-disorders	25.51	(4)	**3.00**	(1)	22.65	(3)	6.60	(2)
Page-blocks	42.91	(4)	**22.00**	(1)	38.70	(3)	24.56	(2)
Blood-t	**6.50**	(1)	10.00	(3)	22.39	(4)	8.46	(2)
Breast-tissue-6	22.45	(4)	**8.00**	(1)	14.09	(3)	8.97	(2)
Movement-libras	47.52	(4)	30.00	(3)	**27.46**	(1)	29.07	(2)
Parkinsons	10.24	(4)	7.00	(2)	7.11	(3)	**4.85**	(1)
Seeds	7.42	(4)	6.00	(3)	4.78	(2)	**3.17**	(1)
Segment	41.21	(4)	41.00	(3)	30.53	(2)	**27.91**	(1)
Ecoli	18.59	(4)	15.00	(3)	12.57	(2)	**7.06**	(1)
Spambase	103.37	(4)	75.00	(3)	74.42	(2)	**31.70**	(1)
Average ranking		3.65		2.15		2.65		1.55

Table 4. Average DT sizes obtained by the DTI methods.

find all the possible hypotheses which cannot be rejected. **Table 3** shows both the average rank (AR) of the results yielded by each method and the *p*-values computed by comparing the average accuracies achieved by the DE-based procedures versus those obtained by the other DTI methods. The *p*-values highlighted with bold numbers indicate that H_0 is rejected for this pair of methods since they show different performance. Unadjusted *p*-values are calculated with the average ranks of the two methods being compared, as is described by Demšar in [68]. These values are used by the BH post hoc test to compute the corresponding adjusted *p*-values. **Table 3** shows that the DE-ODT method has a better performance than the other DTI methods since it has the lowest average rank (1.750), and its results are statistically different than these methods. **Figure 10** shows a graph where the nodes represent the compared methods and the edges joining two nodes indicate that the performance of these methods does not present significant differences. The values shown in the edges are the *p*-values computed by the BH post hoc test. This figure is based on that obtained using the *scmamp* library, and in it is observed that the DE-based methods are not statistically different between them, and that the DE-ODT method is statistically different with the DTI methods. This statistical results indicate that the DE-ODT method is the better DTI method to build oblique DT.

On the other hand, the average sizes of the DTs constructed by the DE-based algorithms and also of those induced by the J48 and the sCART methods are shown in **Table 4** and **Figure 11**. Similar to **Table 2**, the best result for each dataset in **Table 4** is highlighted with bold numbers, and the numbers in parentheses refer to the ranking reached by each method for each dataset. These results indicate that the DE-ODT method produces the most compact DTs. Also, it is observed that the size of the DTs built for the OC1-DE method has less complexity than those yielded by the J48 method.
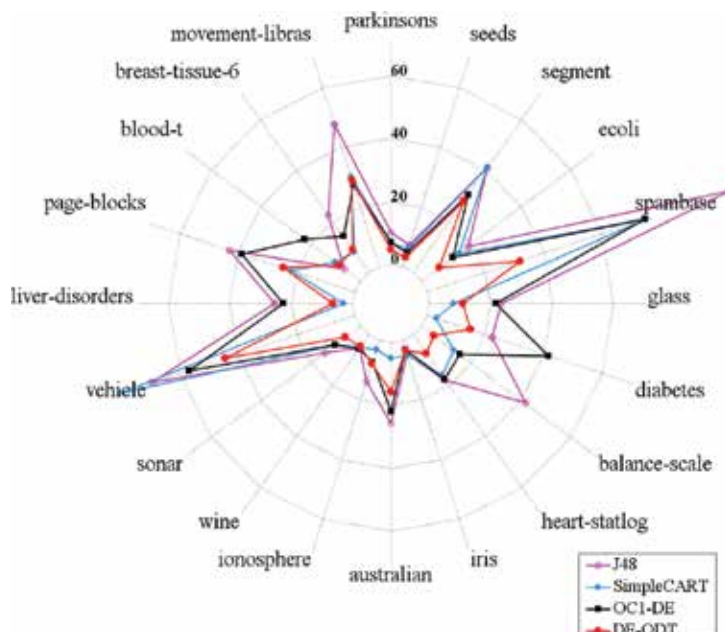
Figure 11. Average DT sizes of several DTI methods.

Dataset	NB		MLP		RBF-NN		RF		OC1-DE		DE-ODT	
Glass	49.44	(6)	67.29	(4)	65.09	(5)	**79.95**	(1)	71.31	(2)	68.97	(3)
Diabetes	75.76	(3)	74.75	(4)	74.04	(5)	**76.18**	(1)	73.37	(6)	75.79	(2)
Balance-scale	90.53	(4)	90.69	(3)	86.34	(5)	81.71	(6)	**93.92**	(1)	91.97	(2)
Heart-statlog	**83.59**	(1)	79.41	(5)	83.11	(2)	82.41	(3)	74.11	(6)	81.11	(4)
Iris	95.53	(5)	96.93	(2)	96.00	(4)	94.73	(6)	96.73	(3)	**97.17**	(1)
Australian	77.19	(6)	83.42	(4)	82.55	(5)	**86.77**	(1)	85.19	(3)	85.61	(2)
Ionosphere	82.17	(6)	91.05	(5)	91.71	(3)	**93.39**	(1)	91.11	(4)	92.28	(2)
Wine	97.47	(4)	**98.03**	(1.5)	97.70	(3)	**98.03**	(1.5)	92.58	(5)	91.88	(6)
Sonar	67.69	(6)	81.59	(2)	72.60	(5)	**84.47**	(1)	77.65	(4)	79.34	(3)
Vehicle	44.68	(6)	**81.11**	(1)	65.35	(5)	75.14	(2)	72.32	(3)	71.33	(4)
Liver-disorders	54.87	(6)	68.72	(3)	65.04	(5)	**72.99**	(1)	67.63	(4)	71.16	(2)
Page-blocks	90.01	(6)	96.28	(4)	94.91	(5)	**97.54**	(1)	96.88	(3)	97.07	(2)
Blood-t	75.28	(5)	78.46	(2)	78.22	(3)	73.62	(6)	76.35	(4)	**78.70**	(1)
Breast-tissue-6	**46.42**	(1)	35.47	(5)	41.13	(3)	45.19	(2)	34.91	(6)	38.85	(4)
Movement-libras	64.14	(5)	80.50	(2)	75.50	(3)	**82.89**	(1)	75.11	(4)	55.63	(6)
Parkinsons	70.10	(6)	**91.44**	(1)	81.49	(5)	91.38	(2)	87.95	(3)	86.43	(4)
Seeds	90.52	(6)	**95.24**	(1)	91.67	(5)	93.57	(3)	93.76	(2)	91.79	(4)
Segment	80.17	(6)	96.21	(2)	87.31	(5)	**98.07**	(1)	95.93	(3)	94.78	(4)
Ecoli	85.51	(2)	84.85	(3)	83.30	(6)	**86.25**	(1)	83.51	(5)	84.72	(4)
Spambase	79.56	(6)	91.19	(4)	81.31	(5)	**95.65**	(1)	92.19	(3)	93.94	(2)
Average ranking		4.800		2.925		4.350		2.125		3.700		3.100

Table 5. Average accuracies obtained by several classification methods.
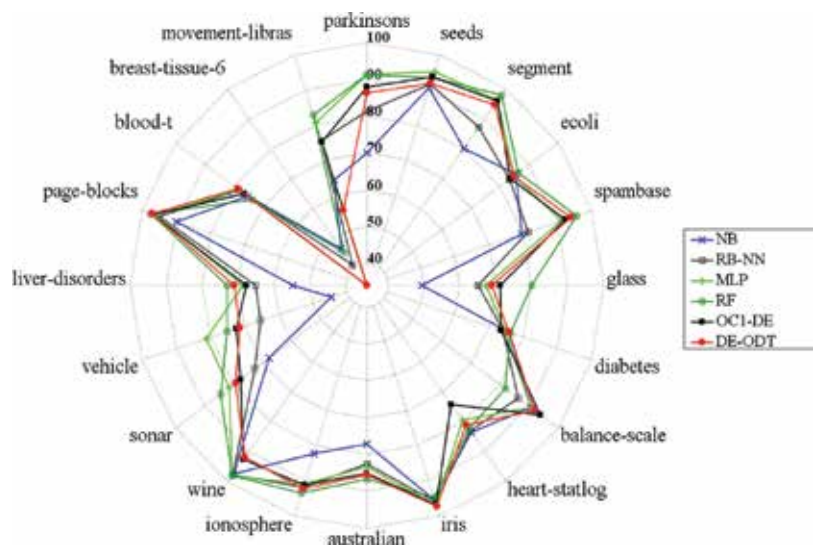
Figure 12. Graphical comparison of the average accuracies obtained by several classification methods.

Method	AR	OC1-DE		DE-ODT	
		Unadjusted	BH	Unadjusted	BH
NB	4.800	6.2979e-02	3.7787e-01	4.0591e-03	**2.8414e-02**
MLP	2.925	1.9019e-01	7.6079e-01	7.6737e-01	8.9374e-01
RBF-NN	4.350	2.7118e-01	7.7679e-01	3.4610e-03	1.3844e-01
RF	2.125	7.7623e-03	5.4336e-03	0.9342e-02	3.9736e-01
OC1-DE	3.700	—	—	3.1049e-01	7.6079e-01
DE-ODT	3.100	3.1049e-01	7.6079e-01	—	—

Table 6. The *p*-values for multiple comparisons among several classification methods.

5.3. Comparison with other classification methods

Table 5 and **Figure 12** show the average accuracies got by several classification methods as well as those obtained by the DE-based methods. In this table, we can observe that the RF algorithm and the MLP method construct more accurate classifiers than the others, and also that the DE-based procedures induce DTs with better accuracy than the models built by both the RBF-NN algorithm and the NB method.

The Friedman statistics computed by analyzing the results got by these six methods with 20 datasets is 27.661, and the corresponding *p*-value is 4.24×10^{-5} so that H_0 is rejected. The BH post hoc test is then applied to find all possible hypotheses that cannot be refused. **Table 6** shows the results of these tests, and **Figure 13** shows the graph corresponding to these *p*-values. The value highlighted with bold in **Table 6** indicates that the DE-ODT method is statistically different with the NB method, only.
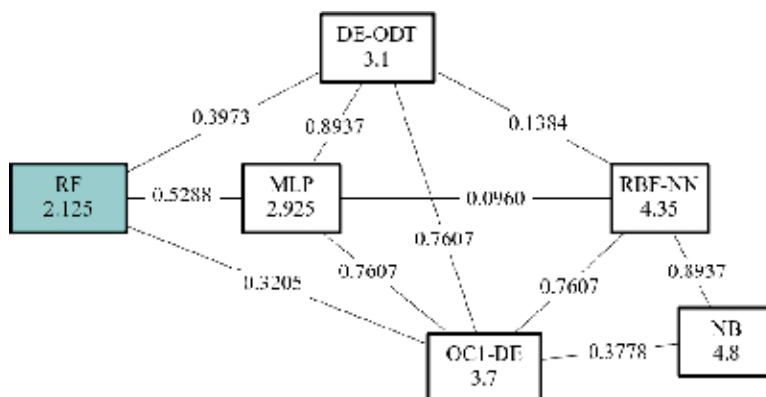
Figure 13. The *p*-value graph of the classification methods.

The p-values obtained by the BH post hoc test point out that the RF method is statistically different only with the RBF-NN algorithm and the NB method, and that both the MLP method and the DE-ODT procedure are statistically different with the NB method. The comparison between the remaining pairs of algorithms indicates that they have a similar performance. The RF method is the best ranked in this comparison, and the AR of the DE-ODT procedure places it as the third best classification method.

6. Conclusions

In this chapter, two DE-based methods to induce oblique DTs are described. The OC1-DE method implements a recursive partitioning strategy to find a near-optimal hyperplane which is used as test condition of an oblique DT. On the other hand, in the DE-ODT method, a global search in the space of oblique DTs is conducted with the aim of finding a near-optimal tree. The DE-ODT method estimates the size of the chromosome encoding a complete tree based on both the number of attributes and the number of classes of the dataset whose model is constructed. This method also defines a scheme to map feasible oblique DTs from this chromosome.

The experimental results obtained indicate that these DE-based methods are better DTI methods, since they build more accurate and compact oblique DTs than those induced by the J48 and the sCART procedures. The DE-ODT method is better than the OC1-DE since its search procedure uses intelligent search procedures combining their exploration and exploitation skills, thus providing a better way to discover the relationships between the attributes used in the training set, and although the search process is only guided by the accuracy of the DT, the models constructed are more compact than those produced by the methods that implement a recursive partitioning strategy. Among the other compared methods, the results got by the OC1-DE method are better than those obtained by the other methods, since it uses a linear combination of attributes in each test condition of the tree, and it produce better hyperplanes than the axis-parallel hyperplanes.

Even though the results yielded by the DE-based variants are not better than those produced by the RF algorithm and the MLP-based classifier, they are statistically equivalent. An advantage of these methods is that it constructs models whose decisions and operations are easily understood, and although the RF method also builds DTs, its voting scheme makes it very difficult to trace the way in which the model takes its decisions.

In this chapter, an analysis of the run time of the algorithms is not performed, since it is known that MHs consume more computational time than other approaches because they work with a group of candidate solutions, unlike the traditional methods where only one DT is induced from the training set. It is important to mention that for many practical applications, the construction of the model is conducted in one offline procedure, so the time of its construction is not a parameter that usually impacts the efficiency of the built model.

Author details

Rafael Rivera-Lopez[1] and Juana Canul-Reich[2]*

*Address all correspondence to: juana.canul@ujat.mx

1 Departamento de Sistemas y Computación, Instituto Tecnológico de Veracruz, Veracruz, VER, México

2 División Académica de Informática y Sistemas, Universidad Juárez Autónoma de Tabasco, Cunduacán, TAB, México

References

[1] Frawley WJ, Piatetsky-Shapiro G, Matheus CJ. Knowledge discovery in databases: An overview. AI Magazine. 1992;**13**(3):57

[2] Stiglic G, Kocbek S, Pernek I, Kokol P. Comprehensive decision tree models in bioinformatics. PLoS One. 2012;**7**(3):1-13

[3] Huysmans J, Dejaeger K, Mues C, Vanthienen J, Baesens B. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decision Support Systems. 2011;**51**(1):141-154

[4] Nettleton DF, Orriols-Puig A, Fornells A. A study of the effect of different types of noise on the precision of supervised learning techniques. Artificial Intelligence Review. 2010;**33**(4):275-306

[5] Du KL, Swamy MNS. Search and Optimization by Metaheuristics. Switzerland: Springer; 2016

[6] Storn R, Price K. Differential evolution – A simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization. 1997;**11**(4):341-359

[7] Galea M, Shen Q, Levine J. Evolutionary approaches to fuzzy modelling for classification. The Knowledge Engineering Review. 2004;**19**(1):27-59

[8] Espejo PG, Ventura S, Herrera F. A survey on the application of genetic programming to classification. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2010;**40**(2):121-144

[9] Kokol P, Pohorec S, Štiglic G, Podgorelec V. Evolutionary design of decision trees for medical application. Data Mining and Knowledge Discovery. 2012;**2**(3):237-254

[10] Barros RC, Basgalupp MP, Carvalho ACPLF, Freitas AA. A survey of evolutionary algorithms for decision-tree induction. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2012;**42**(3):291-312

[11] Kolçe E, Frasheri N. The use of heuristics in decision tree learning optimization. International Journal of Computer Engineering in Research Trends. 2014;**1**(3):127-130

[12] Alurkar AA, Ranade SB, Joshi SV, Ranade SS, Sonewar PA, Mahalle PN, Deshpande AV. A proposed data science approach for email spam classification using machine learning techniques. In: 2017 Internet of Things Business Models, Users, and Networks; November 2017; pp. 1-5

[13] Mishan MT, Kushan AL, Fadzil AFA, Amir ALB, Anuar NB. An analysis on business intelligence predicting business profitability model using naive Bayes neural network algorithm. In: 2017 7th IEEE International Conference on System Engineering and Technology (ICSET). Shah Alam, Malasya: IEEE; 2017; pp. 59-64

[14] Lippmann RP. An introduction to computing with neural nets. ASSP Magazine. 1987;**4**(2): 4-22

[15] Abe S. Support Vector Machines for Pattern Classification. London, UK: Springer; 2005

[16] Murthy SK. On Growing Better Decision Trees from Data. PhD thesis, The Johns Hopkins University; 1997

[17] Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936;**7**(2):179-188

[18] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. San Francisco, CA, USA: Morgan Kaufmann; 2005

[19] Heath DG. A Geometric Framework for Machine Learning [PhD thesis]. Johns Hopkins University; 1993

[20] Birattari M. Tuning Metaheuristics: A Machine Learning Perspective, Volume 197 of Studies in Computational Intelligence. Berlin Heildelberg: Springer; 2009

[21] Boussaïd I, Lepagnot J, Siarry P. A survey on optimization metaheuristics. Information Sciences. 2013;**237**:82-117

[22] Talbi EG. Metaheuristics: From Design to Implementation. Hoboken, NY, USA: Wiley; 2006

[23] Li J, Ding L, Li B. Differential evolution-based parameters optimisation and feature selection for support vector machine. International Journal of Computational Science and Engineering. 2016;**13**(4):355-363

[24] Leema N, Nehemiah HK, Kannan A. Neural network classifier optimization using differential evolution with global information and back propagation algorithm for clinical datasets. Applied Soft Computing. 2016;**49**:834-844

[25] Geetha K, Baboo SS. An empirical model for thyroid disease classification using evolutionary multivariate Bayesian prediction method. Global Journal of Computer Science and Technology. 2016;**16**(1):1-9

[26] García S, Derrac J, Triguero I, Carmona CJ, Herrera F. Evolutionary-based selection of generalized instances for imbalanced classification. Knowledge-Based Systems. 2012; **25**(1):3-12

[27] Tušar T. Optimizing accuracy and size of decision trees. In: Proceedings of the 16th International Electrotechnical and Computer Science Conference (ERK-2007), Portorož; Slovenia, 2007; pp. 81-84

[28] Neri F, Tirronen V. Recent advances in differential evolution: A survey and experimental analysis. Artificial Intelligence Review. 2010;**33**(1–2):61-106

[29] Heath DG, Kasif S, Salzberg S. Induction of oblique decision trees. In: Bajcsy R, editor. Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93); Chambéry; France, 1993. pp. 1002-1007

[30] Murthy SK, Kasif S, Salzberg S, Beigel R. OC1: A randomized algorithm for building oblique decision trees. In: AAAI'93. Vol 93. AAAI press; 1993. pp. 322-327

[31] Cantú-Paz E, Kamath C. Inducing oblique decision trees with evolutionary algorithms. IEEE Transactions on Evolutionary Computation. 2003;**7**(1):54-68

[32] Li XB, Sweigart JR, Teng JTC, Donohue JM, Thombs L, Wang SM. Multivariate decision trees using linear discriminants and tabu search. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans. 2003;**33**(2):194-205

[33] Orsenigo C, Vercellis C. Discrete support vector decision trees via tabu search. Computational Statistics & Data Analysis. 2004;**47**(2):311-322

[34] Zhang K, Xu Z, Buckles BP. Oblique decision tree induction using multimembered evolution strategies. In: Dasarathy BV, editor. Proceeding of Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security, SPIE 2005. Vol. 5812. Orlando:, Florida SPIE; 2005. pp. 263-270

[35] Chai BB, Zhuang X, Zhao Y, Sklansky J. Binary linear decision tree with genetic algorithm. In: Proceedings of the 13th International Conference on Pattern Recognition (ICPR'96). Track D: Parallel and Connectionist Systems. Vol. IV. Vienna: IEEE; 1996. pp. 530-534

[36] Struharik R, Vranjkovic V, Dautovic S, Novak L. Inducing oblique decision trees. In: Proceedings of the 12th International Symposium on Intelligent Systems and Informatics (SISY–2014), Subotica, Serbia: IEEE; 2014. pp. 257-262

[37] Krętowski M. An evolutionary algorithm for oblique decision tree induction. In: Rutkowski L, Siekmann J, Tadeusiewicz R, Zadeh LA, editors. Proceedings of the 7th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2004). LNAI. Vol. 3070. Zakopane, Poland: Springer; 2004. pp. 432-437

[38] Pangilinan JM, Janssens GK. Pareto-optimality of oblique decision trees from evolutionary algorithms. Journal of Global Optimization. 2011;**51**(2):301-311

[39] Rivera-Lopez R, Canul-Reich J, Gámez JA, Puerta JM. OC1-DE: A differential evolution based approach for inducing oblique decision trees. In: Rutkowski L, Korytkowski M, Scherer R, Tadeusiewicz R, Zadeh LA, Zurada JM, editors. Proceedings of the 16th International Conference in Artificial Intelligence and Soft Computing (ICAISC 2017). LNCS. Vol 10245. Zakopane, Poland: Springer; 2017. pp. 427-438

[40] Dumitrescu D, András J. Generalized decision trees built with evolutionary techniques. Studies in Informatics and Control. 2005;**14**(1):15-22

[41] Vukobratovic B, Struharik R. Evolving full oblique decision trees. In: Proceedings of the 16th IEEE International Symposium on Computational Intelligence and Informatics (CINTI 2015). Budapest, Hungary: IEEE; 2015. pp 95-100

[42] Krętowski M, Grześ M. Global induction of oblique decision trees: An evolutionary approach. In: Kłopotek MA et al editors. IIPWM'05, Volume 31 of ASC. Berlin Heildelberg: Springer; 2005. pp. 309-318

[43] Krętowski M, Grześ M. Evolutionary learning of linear trees with embedded feature selection. In: Rutkowski L et al editors. ICAISC 2006. LNAI. Volume 4029 of LNAI-Springer; 2006. pp. 400-409

[44] Gray JB, Fan G. Classification tree analysis using TARGET. Computational Statistics & Data Analysis. 2008;**52**(3):1362-1372

[45] Liu KH, Xu CG. A genetic programming-based approach to the classification of multiclass microarray datasets. Bioinformatics. 2009;**25**(3):331-337

[46] Bot MCJ, Langdon WB. Application of genetic programming to induction of linear classification trees. In: Poli R et al., editors. EuroGP 2000. LNCS. Vol. 1802. Berlin Heildelberg: Springer; 2000. pp. 247-258

[47] Bot MCJ, Langdon WB. Improving induction of linear classification trees with genetic programming. In: Whitley LD et al., editors. GECCO-2000. San Francisco, CA, USA: Morgan Kaufmann; 2000. pp. 403-410

[48] Agapitos A, O'Neill M, Brabazon A, Theodoridis T. Maximum margin decision surfaces for increased generalisation in evolutionary decision tree learning. In: Silva S et al editors. EuroGP 2011. LNCS. Volume 6621. Berlin Heildelberg: Springer; 2011. pp. 61-72

[49] Lopes RA, Freitas ARR, Silva RCP, Guimarães FG. Differential evolution and perceptron decision trees for classification tasks. In: Yin H, Costa JAF, Barreto G, editors. Proceedings of the 13th International Conference Intelligent Data Engineering and Automated Learning (IDEAL 2012). LNCS. Vol. 7435. Natal, Brazil: Springer; 2012. pp. 550-557

[50] Freitas ARR, Silva RCP, Guimarães FG. Differential evolution and perceptron decision trees for fault detection in power transformers. In: Snášel V et al, editors. SOCO Models in Industrial & Environmental Applications. AISC. Volume 188. Berlin Heildelberg: Springer; 2013. pp. 143-152

[51] Rivera-Lopez R, Canul-Reich J. A global search approach for inducing oblique decision trees using differential evolution. In: Mouhoub M, Langlais P, editors. Proceedings of the 30th Canadian Conference on Artificial Intelligence (AI 2017), volume 10233 of LNCS. Edmonton, Canada: Springer; 2017. pp. 27-38

[52] Quinlan JR. Induction of decision trees. Machine Learning. 1986;**1**(1):81-106

[53] Murthy SK, Kasif S, Salzberg S. A system for induction of oblique decision trees. Journal of Artificial Intelligence Research. 1994;**2**(1):1-32

[54] Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. Boca Raton, FL, USA: Chapman and Hall; 1984

[55] Lichman M. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences; 2013

[56] Durillo JJ, Nebro AJ. jMetal: A Java framework for multi-objective optimization. Advances in Engineering Software. 2011;**42**(10):760-771

[57] Das S, Konar A, Chakraborty UK. Two improved differential evolution schemes for faster global search. In: Beyer HG, editor. Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation (GECCO'05), Washington, DC, USA: ACM; 2005. pp. 991-998

[58] Quinlan JR. C4.5: Programs for Machine Learning. San Mateo, CA, USA: Morgan Kaufmann; 1993

[59] Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association. 1937;**32**(200):675-701

[60] Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika. 1988;**75**(2):383-386

[61] Calvo B, Guzmán-Santafé R. scmamp: Statistical comparison of multiple algorithms in multiple problems. The R Journal. 2016;**8**(1):248-256

[62] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An update. SIGKDD Explorations Newsletter. 2009;**11**(1):10-18

[63] Witten IH, Frank E, Trigg LE, Hall MA, Holmes G, Weka SJC. Practical machine learning tools and techniques with Java implementations. Technical Report 11, Department of Computer Science. New Zeland: Waikato; 1999

[64] John GH Langley P. Estimating Continuous Distributions in Bayesian Classifiers. In: Besnard P, Hanks S, editors. Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI'95), San Francisco, CA, USA: Morgan Kaufmann; 1995. pp. 338-345

[65] Murtagh F. Multilayer perceptrons for classification and regression. Neurocomputing. 1991;**2**(5):183-197

[66] Frank E. Fully supervised training of Gaussian radial basis function networks in WEKA. Technical Report 04, Department of Computer Science. New Zeland: Waikato; 2014

[67] Breiman L. Random forests. Machine Learning. 2001;**45**(1):5-32

[68] Demšar J. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research. 2006;**7**:1-30

Advanced Content and Interface Personalization through Conversational Behavior and Affective Embodied Conversational Agents

Matej Rojc, Zdravko Kačič and Izidor Mlakar

Additional information is available at the end of the chapter

Abstract

Conversation is becoming one of the key interaction modes in HMI. As a result, the conversational agents (CAs) have become an important tool in various everyday scenarios. From Apple and Microsoft to Amazon, Google, and Facebook, all have adapted their own variations of CAs. The CAs range from chatbots and 2D, carton-like implementations of talking heads to fully articulated embodied conversational agents performing interaction in various concepts. Recent studies in the field of face-to-face conversation show that the most natural way to implement interaction is through synchronized verbal and co-verbal signals (gestures and expressions). Namely, co-verbal behavior represents a major source of discourse cohesion. It regulates communicative relationships and may support or even replace verbal counterparts. It effectively retains semantics of the information and gives a certain degree of clarity in the discourse. In this chapter, we will represent a model of generation and realization of more natural machine-generated output.

Keywords: co-verbal behavior generation, affective embodied conversational avatars, humanoid robot behavior, multimodal interaction, unity, EVA framework

1. Introduction

One of the key challenges in the modern human-machine interaction (HMI) is the generation of more natural, more personalized, and more human-like human-machine interaction [1]. As a result, the conversational agents (CAs) are gaining interest and traction, especially due to the fact that most of the user devices are already capable to support multimedia and the concept of conversational agents (CAs). Apple, Microsoft, Amazon, Google, Facebook,

etc., already have adapted their own variations of CAs. Moreover, the newest technologies, such as the Amazon Echo and Google Home, which are positioned as supporting multiuser and highly personalized interaction in collocated environments (e.g., homes and ambient-assisted living environments), integrate virtual agents supporting both visual and auditory interactions [2–4]. Thus, exploring the conversational models and challenges around CA supported interaction represents a timely topic. The production of conversational behavior, e.g., socially shared information and attitude, incorporates much more than just speech verbal exchange. Namely, it is multimodal and multilayered, since it entails multiple verbal and nonverbal signals that are correlated in a dynamic and highly unpredictable settings. One might even say that the social interaction involves synchronized signal verbal and nonverbal channels. The verbal channels carry symbolic/semantic interpretation of message through the linguistic and paralinguistic features of interaction, while the co-verbal channels serves as an orchestrator of communication [5–7]. Thus, such an interaction facilitates full embodiment of the collocutors. It also exploits physical environment in which the interaction is positioned in [8–10]. Further, the co-verbal behavior is actually equally relevant as speech. Namely, it actively contributes to the information presentation and understanding, as well as the discourse itself. It establishes semantic coherence and regulates communicative relationships. It may support or even replace the verbal communication in order to clarify or reinforce the information provided by the verbal counterparts [11–13]. The co-verbal behavior goes well beyond an add-on or a style of information representation. For instance, spatial orientation of the face and eye gaze are key nonverbal cues that shape the footing of the conversational participants [14]. Through the co-verbal responses, the listeners may signal their interest, attention, and understanding [15]. As a result, the role of co-verbal (and nonverbal) behavior in human communication and in human-machine interaction has been increasingly scrutinized over the last few decades, within a wide range of contexts [16–21]. Embodied conversational agents (ECAs) are nowadays the most natural selection for the generation of affective and personalized agents. ECAs are those CAs that can facilitate full virtual body and the available embodiment in order to incorporate humanlike responses. The ECA technology ranges from chatbots and 2D/3D realizations in a form of talking heads [22–24] to fully articulated embodied conversational agents engaged in various concepts of HMI, including sign language [25], storytelling [26], companions [27], and virtual hosts within user interfaces, and even used as moderators of various concepts in ambient-assisted living environments [28–32].

In this chapter we present novel expressive conversational model for facilitating human-like conversations and a solution for affective and personalized human-machine interaction. The model facilitates (i) a platform for the generation of "conversational" knowledge and resources, (ii) a framework for planning and generation of (non-)co-verbal behavior, and (iii) a framework for delivery of affective and reactive co-verbal behavior through attitude, emotion, and gestures synchronized with the speech. Namely, the EVA expressive conversational model is outlined in Section 2. The main idea is to formulate various forms of co-verbal behavior (gestures) with respect to arbitrary and unannotated text and broader social and conversational context. The "conversational" knowledge and resources required are generated via annotation of spontaneous dialog and through the corpus analysis as presented in Section 3.

In Chapters 4 and 5, how these resources are integrated into the two-folded approach of the automatic co-verbal behavior generation is then described. The presented approach involves (a) the problem of behavior formulation (intent and behavior planning) and (b) the problem of behavior realization (animation via ECA). Finally, we conclude with synthesis of affective co-verbal behavior within interfaces and final remarks.

2. EVA conversational model for expressive human-machine interaction

In order to cope also with the complexity in multiparty conversations, and in order to apply the knowledge to various concepts in human-machine interaction in a form of conversational behavior, we have envisaged and deployed an advanced EVA conversational model, which is used (a) to study the nature of natural behavior of human-collocutors; (b) to create conversational knowledge in form of linguistic, paralinguistic verbal, and nonverbal features; (c) and to test theories and to apply knowledge in various conversational settings as part of situation understanding or as a part of output generation processes. The presented EVA conversational model is outlined in **Figure 1**. As can be seen, it consists of the following three cooperative frameworks/platforms: *conversational analysis platform*, *EVA framework*, and *EVA realization framework*.
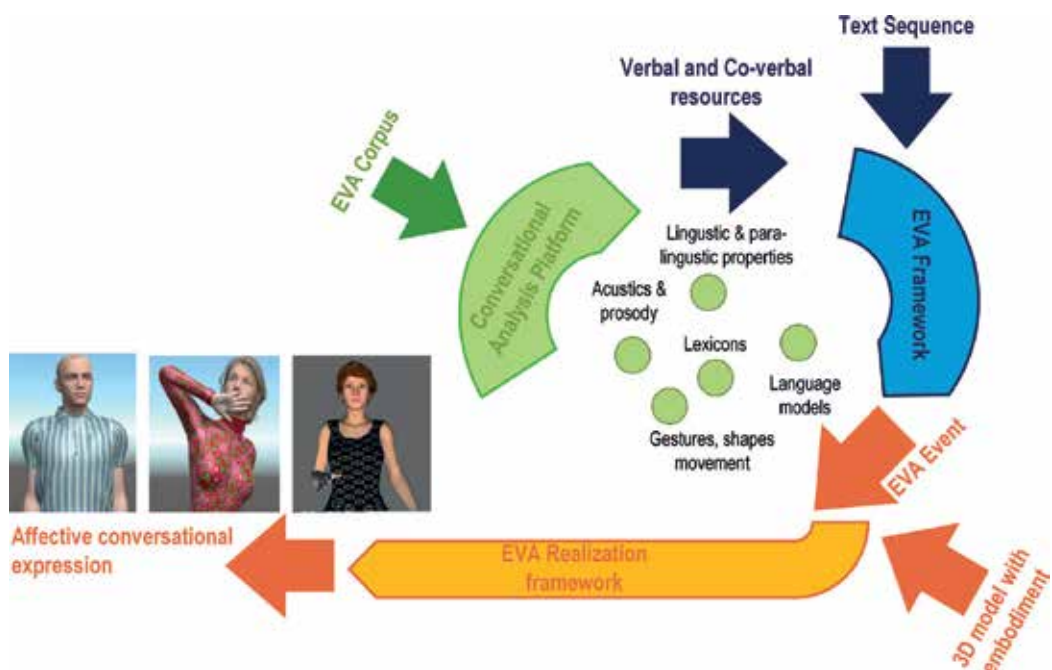
Figure 1. EVA conversational model for generation and exploitation of expressive human-like machine responses.

The model builds on the notion that verbal to co-verbal alignment and synchronization are driving forces behind affective and social interaction. Thus, the *conversational analysis platform* is used for analyzing how linguistic and paralinguistic features interplay with embodiments during complex, spontaneous, and multiparty interactions. Its processing is based on multi-modal corpus, named EVA corpus [33], the EVA annotation scheme developed to describe the complex relations of co-verbal behavior (proposed in [34]), and *EVA framework* capable (a) to capture various contexts in the "data" and (b) to provide the basis to analytically investigate into various multidimensional correlations among co-verbal behavior features.

Communication is in its fundaments a multimodal process. In order to describe the nature of face-to-face interaction, we have chosen to build upon the concept of "multimodality in interaction" over linguistic basis [2]. We extended this concept with a cognitive dimension consisted of various concepts, such as emotions, sentiment, and communicative intent. We also blend it with the meta-information represented as a set of paralinguistic features, such as facial expressions and gestures, prosody, pitch, dialog function and role, etc., [17, 35]. As outlined in **Figure 1**, the annotated EVA corpus is used to build resources required for the planning and generation of conversational behavior, both verbal part (e.g., text-to-speech (TTS) synthesis) and (non-)co-verbal part (conversational behavior synthesis). These are lexicons, language models, semiotic grammar of communicative intent, lexicon of conversational shapes, gestures and movements, acoustic and prosodic properties, and other linguistic and paralinguistic features (e.g., word/syllable segmentation, sentence type, sentiment, etc.) that are used within behavior generation rules and machine-learned behavior generation models. The resources generated within the *conversation analysis platform* are fed to the *EVA framework*. The main idea of the proprietary *EVA framework* proposed in [36] is to evoke a social response in human-machine interaction through affective synthetic response generated on arbitrary and unannotated texts. Thus, the EVA behavior generation model within the *EVA framework* is data-driven and also driven by the text-to-speech (TTS) synthesis engine. The model is modular and merged with the TTS engine's architecture, into the first omni-comprehensive TTS system's engine as proposed in [37]. The output of the EVA framework represents the complete co-verbal behavior described by using the proprietary procedural description language, named EVA-Script, and the synthesized speech, both perfectly synchronized at the phoneme level. The co-verbal behavior is described within an EVA event, where it represents a contextual link between the language, context-independent motor skills (shapes, movements, and poses that conversational agent can display), and the context-dependent intent for which the behavior has been generated (e.g., situation, relation, communicative function, etc.). The behavior is already adapted to the nature and capabilities of the virtual entity representing it. The EVA event then represents input for the *EVA realization framework*.

The *EVA realization framework* is built on the premises that natural multimodal interaction is much more than speech accompanied with the repetitive movements of the limbs and face. Namely, natural interaction entails multiple behavior variations that are correlated in dynamic and highly unpredictable settings [6]. It also incorporates various social and inter-personal signals in order to "color" the final outcome and can dynamically adapt to various intra- and interpersonal contexts as well as various situational contexts [4, 38]. The role of this framework is to transform the co-verbal descriptions contained in EVA events into

articulated movement generated by the expressive virtual entity, e.g., to apply the EVA-Script language onto the articulated 3D model EVA in the form of animated movement. Further, the framework contains the *animation-parameters builder* and the *animation-realization engine*. Both are maintained within the *EVA behavior realizer* and implemented as a series of standalone modules. The *animation-parameters builder* is used to understand the specified behavior and to adapt it to the restriction of the targeted agent. Thus, it transforms the EVA-Script sequence into animation parameters that are then mapped to different control units of the ECA's 3D articulated model, while the *animation-realization engine* is responsible for scheduling and execution of the translated animation parameters in a form of sequences of parallel/sequential transformation of ECA-related resources (e.g., meshes textures, bones, morphed shapes, etc.). These resources actually constitute virtual agent's embodiment in a form of hand/arm gestures, posture and body configuration, head movement and gaze, facial expressions, and lip sync. Finally, the *EVA realization framework* also incorporates procedures required for efficient integration of the embodied conversational agent into various user interfaces. In the next section, we will present particular modules of the *EVA conversational model* in more detail. Firstly, we will describe the *conversational analysis platform*, which represents the basis not only for the generation of communicative behavior but also for understanding the nature of complex conversational behavior and face-to-face interaction as a whole.

3. Conversation analysis and annotation scheme

Conversation analysis represents a powerful tool for analyzing language and human co-verbal behavior in various aspects of social communication [39]. Namely, interaction through dialog is an act of conveying information, in which humans can convey information through a variety of methods, such as speaking, body language (gestures and posture) and facial expression, and even social signals [40]. Interpersonal communication can involve the transfer of information between two or more collocutors that use verbal and nonverbal methods and channels. Symbolic/semantic interpretation of message is presented through linguistic and paralinguistic features, while the co-verbal part in general serves as an orchestrator of communication [5]. The concept of the co-verbal behavior has become one of the central research paradigms and one of the important features of human-human interaction. It has been investigated from various perspectives, e.g., from anthropology and linguistics and psychosociological fields to companions, communication and multimodal interfaces, smart homes, ambient assisted living, etc. The multimodal corpora represent the results of various research efforts. They are the tools through which researchers analyze the inner workings of interaction. The knowledge generated by using such multimodal corpora and annotation schemes, therefore, represents a key resource for better understanding the complexity of the relations between verbal and co-verbal parts of human-human communication. It provides insights into understanding of several (social) signals and their interplay in the natural exchange of information. In the following section, we will represent EVA corpus, a corpus of spontaneous and multiparty face-to-face dialog. We will also outline the EVA annotation scheme designed as a tool for corpus analytics and generation of verbal and nonverbal resources [33].

3.1. The multimodal EVA corpus

Among video corpora, television (TV) interviews and theatrical plays have shown themselves to be very usable resource of spontaneous conversational behavior for the analytical observation, and annotation of co-verbal behavior and emotions, used during conversation. In general, TV discussions represent a mixture of institutional discourse, semi-institutional discourse and casual conversation. Material used in existing corpora is often subject to certain restrictions in order to reduce the conversational noise, such as time restriction, strict agenda, strict scenario and instructions to implement targeted concepts, and technical features (camera direction and focus, editing) that further influence especially communicative function of co-verbal behavior and its expressive dimensions (speech, gestures, facial displays). However, the conversational noise, if properly analyzed and incorporated, may unravel a lot of features and contexts that model the natural multimodal conversational expressions. Namely, by exploiting the casual nature and noise in the material, as we do with the EVA corpus, we can take into consideration the complete interplay of various conversation phenomena, such as dialog, emotional attitude, prosody, communicative intents, structuring of information, and the form of its representation. All these can give us a true insight into how informal communication works, what stimuli triggers conversational phenomena, and how do these impulses interact and reflect on each other. Such relations can then provide synthetic agents with the basis for the multimodal literacy, namely, the capacity to construct meaning through understanding of situation and responding to some not predefined situation. The conversational setting in the EVA corpus is totally relaxed and free and is built around a talk show that follows some script/scenario; however, the topics discussed are highly unpredictable, changeable, informal, and full of humor and emotions. Further, although sequencing exists, it is performed highly unorderly as are also the communicative functions. This guarantees a highly causal and unordered human discourse, with lots of overlapping statements and roles, vivid emotional responses, and facial expressions. The goals of the corpus and the annotation scheme are built around (semiotic) communicative intent as the driving force for the generation of co-verbal and communicative behavior. The communicative intent is a concept through which we are able to correlate the intent of the spoken information (defined, e.g., through part-of-speech (POS) tags, prosodic features, and classification of interpretation through meaning) with co-verbal behavior (gestures). Human face-to-face interactions are multimodal and go well beyond pure language and semantics. Within the EVA corpus and corpus analysis outlined in this section, we decided for the extension of semantics by applying the concept of communicative intent and other linguistic and paralinguistic features, such as dialog role and functions, attitude, sentiment and emotions, prosodic phrases, pitch, accentuation, etc., to the observed exchange of information.

3.2. The EVA annotation scheme

In order to capture and analyze conversational phenomena in EVA corpus, the video material is annotated by following the EVA annotation scheme that incorporates linguistic and paralinguistic features and interlinks them with nonverbal movement [37, 41]. The annotation process is performed separately for each speaker. The formal model of the annotation

scheme is outlined in **Figure 2**. In addition to symbolic conversational correlations, the presented scheme also targets the analysis of the form of movement in high resolution. This allows us to test and evaluate also the low-level correlation between movement and prosody, communicative intent, and other linguistic and paralinguistic features. As a result, we can analyze the face-to-face interactions in greater detail. Further, through the extracted knowledge, we are able to pair features into complex stimuli used for triggering the generation of the conversational artifacts and to improve the understanding of the situation through multimodality. As can be seen in **Figure 2**, the annotation session per speaker is separated into symbolical annotation (e.g., annotation of function) and into annotation of the form (e.g., annotation of visualization). Each of the annotated features (linguistic, paralinguistic, and movement/shape related) is captured on a separate track and interlinked with spoken content and movement via a shared timeline. In this way we are able to analyze and search for various multidimensional relationships between conversational artifacts and identify and establish temporal and symbolic links between verbal and co-verbal features of complex multiparty interaction.

As outlined in **Figure 2**, the EVA annotation scheme has the capacity not only to establish links between features on the symbolic level but also to interlink the form of co-verbal movement and its manifestation (e.g., the visualization) with symbolic artifacts, such as dialog role, emotions, lemma, POS tags, sentence type, phrase breaks, prominence, sentiment, and semiotic intent. This is quite important for investigating into the multidimensional interlinks between various features. For instance, the co-verbal behavior may originate as a reflection of attitude/emotion or even be a supportive artifact in the implementation of the communicative function (e.g., feedback, turn taking, turn accepting, sequencing, etc.), while the verbal behavior primarily used for representation of information may also reflect attitude/emotion or be adjusted to serve as a part of the implementation of a communicative function. Through
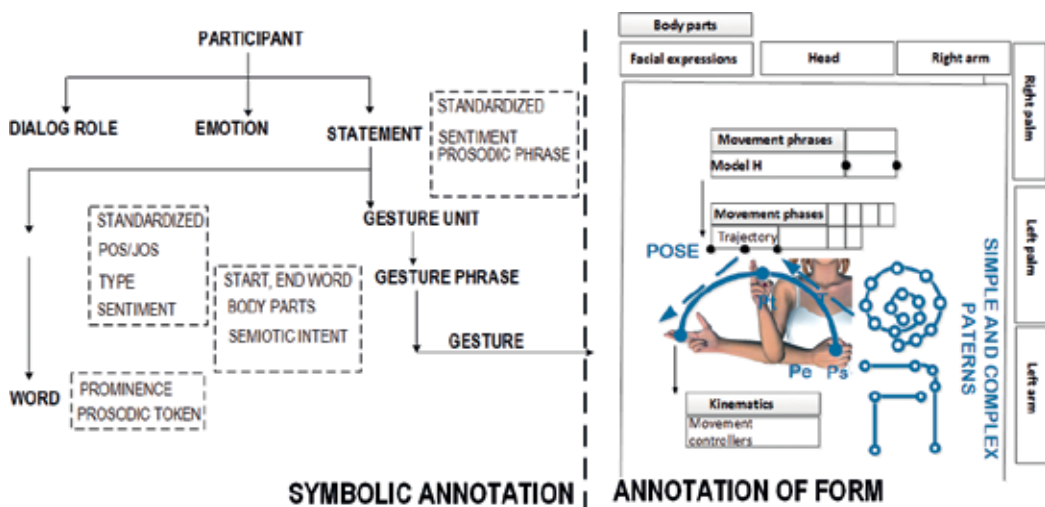

Figure 2. The formal model of the EVA annotation scheme and the topology of the annotation of the form.

the annotation scheme, all artifacts are interconnected through temporal domain and can be related among each other in numerous ways and combinations.

Overall, the symbolic annotation allows us to identify and describe in detail the nature of communicative acts performed during information exchange. The annotation of the form, on the other hand, then describes the shapes and movements generated during these symbolically defined communication acts. Thus, the concept of EVA annotation is based on the idea that symbolic relations and concepts are established on the functional/symbolic level and realized via body expressions, e.g., hand gestures (left and right arm and hands), facial expression, head movement, and gaze. During each symbolic act, the movement of each body part (hands, arms, head, and face) can be described with movement phrase, movement phases, transitions, and the articulators propagating the observed movement. The movement phrase describes the full span of movement phases, where each movement phase contains a mandatory stroke and optional preparation, hold, and retraction phases. Further, each movement phase identifies a pose at the beginning Ps and a pose at the end Pe, where poses are "interconnected" with a trajectory that identifies the path over which the observed body parts propagate from the start pose to the end pose. The trajectory T is a parametric description of propagation, which includes the partitioning of the trajectory T into movement primes (simple patterns), such as linear and arc, each defined through the intermediate poses.

4. Advanced co-verbal behavior generation by using EVA framework

The EVA behavior generation model proposed in [36] is used to convert general unannotated texts into co-verbal behavior description automatically. The model integrates several nonverbal elements that are associated (correlated) with the verbal behavior (speech). Therefore, general texts can be presented as multimodal output, consisting of spoken communication channel as well as synchronized visual communication channel. The EVA behavior generation model performs synchronization of the verbal and nonverbal elements that is necessary in order to achieve desired naturalness, in the domain of meaning (intent) and in the temporal domain. Further, the EVA model generates the co-verbal behavior descriptions and the verbal behavior simultaneously. The EVA model distinguishes between the behavior generation and behavior realization step. **Figure 3** outlines the expressive conversational behavior generation module, which consists of the following three concurrent processes: intent classification, behavior planning, and speech synthesis. The *speech synthesis process* converts general text into speech signal and also represents a source of linguistic and prosodic features that are used for planning and synchronizing the nonverbal behavior. Further, the *intent classification process* identifies the nature of the spoken content through pattern matching incorporating linguistic and prosodic features, where the intent of the input text is defined in the form of classification of linguistic expressions into semiotic classes. The result is a set of possible interpretations of the input text. Further, the *behavior planning process* involves filtering of several interpretations, the pose/gesture selection process based on target cost calculation mechanism, and the temporal synchronization step based on prosodic and acoustic features obtained during synthesizing the speech signal. As outlined in **Figure 3**, the EVA behavior generation model
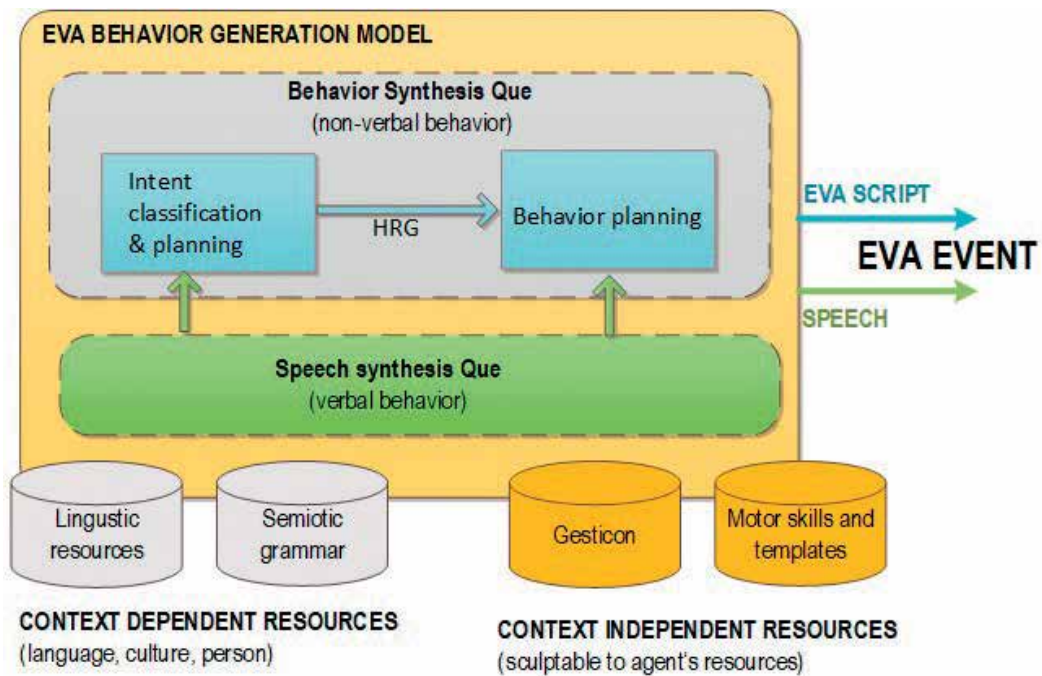
Figure 3. The architecture of the EVA behavior generation model for the generation of expressive co-verbal behavior.

facilitates language–/context-dependent and language–/context-independent resources. The language–/context-dependent resources are linguistic resources and the nonlinguistic semiotic grammar. The linguistic resources include lexicons, language models, rules, and corpuses, while the nonlinguistic grammar includes sets of semiotic rules for pairing language with communicative intent. The language–/context-independent resources are *Gesticon* and the repository of *motor skills*. *Gesticon* couples the context-independent *motor skills* with the semiotic nature of the intent. Namely, we associate a given semiotic pattern (communicative intent) with a unique set of features describing the manifestation of the shape/pose. The semiotic pattern incorporates a semiotic class/subclass, the movement phase within which the pose manifestation is observed, and the POS tag of the word represented as the nucleus of meaning. The unique set of features describing the manifestation of shape/pose incorporates body-pose identification, representing a pair of initial and final pose, a general trajectory of hand movement, semantic word relation, and minimal and maximal temporal values within which the gesture was observed to be carried out, and the number of occurrences of the given gesture that was observed in the EVA corpus. The *semiotic grammar* and *Gesticon* are created and populated with patterns and associated with unique sets based on the analysis and annotation discussed in Section 3.

The conceptual EVA behavior generation model has been actualized in the form of the EVA engine outlined in **Figure 4**. The EVA engine converts a general text into the speech signal accompanied by humanlike synchronized gesticulation and lip movement. The EVA engine
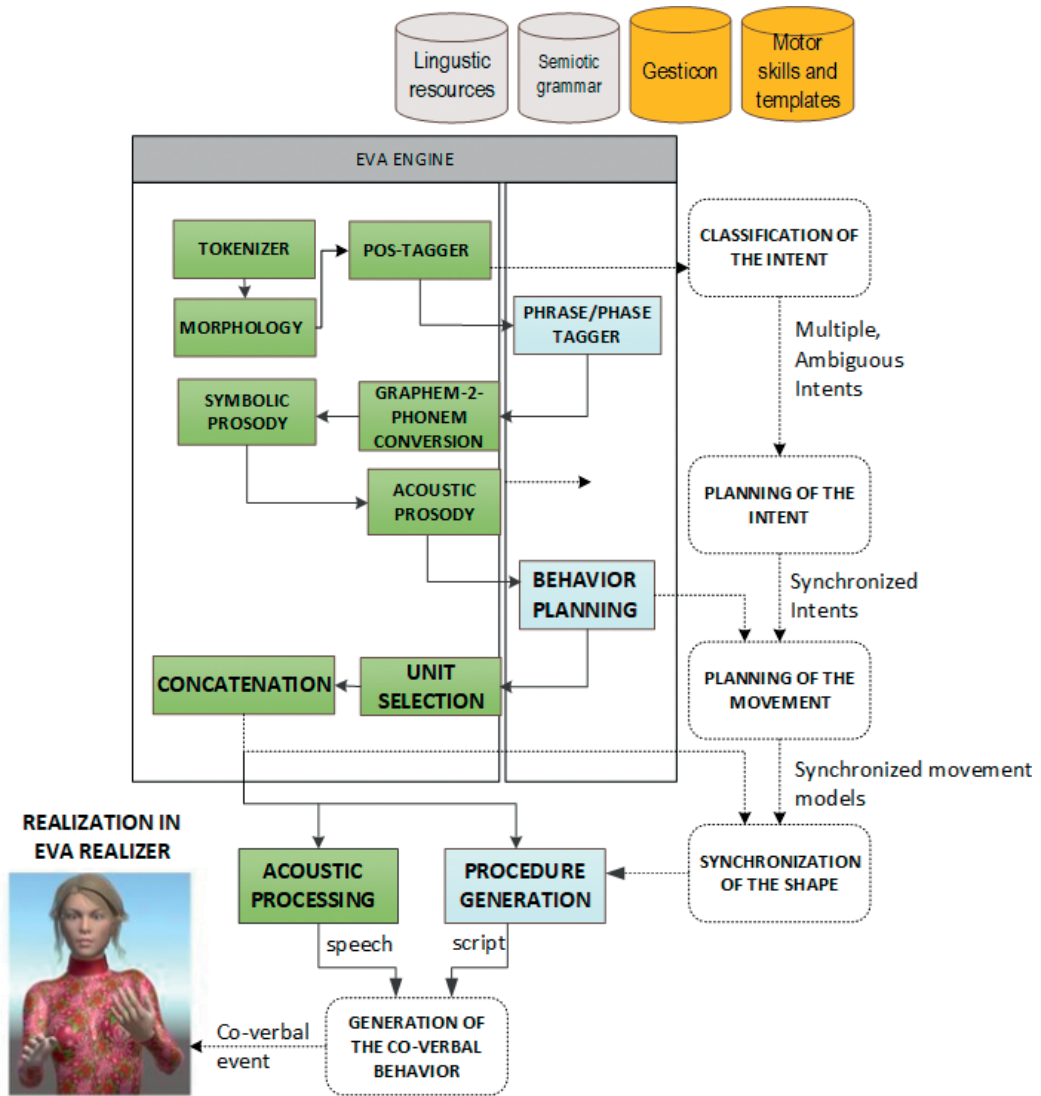
Figure 4. EVA engine: implementation of the EVA behavior generation model.

is composed of (a) processing steps for text-to-speech synthesis as proposed in the TTS system PLATTOS [42] and of (b) processing steps for expressive co-verbal behavior generation algorithm. All steps are fused into a compact processing EVA engine. In this way, the expressive co-verbal behavior generation algorithm works with the verbal modules concurrently, by sharing data, machine-trained models, and other resources. The EVA engine takes into account the latest results of research on multimodal communication, goes beyond traditional computational speech processing techniques, and facilitates heuristic and psychological models of human interaction. The algorithm for the generation of expressive co-verbal behavior implemented within the engine in **Figure 5** generates the co-verbal behavior by considering
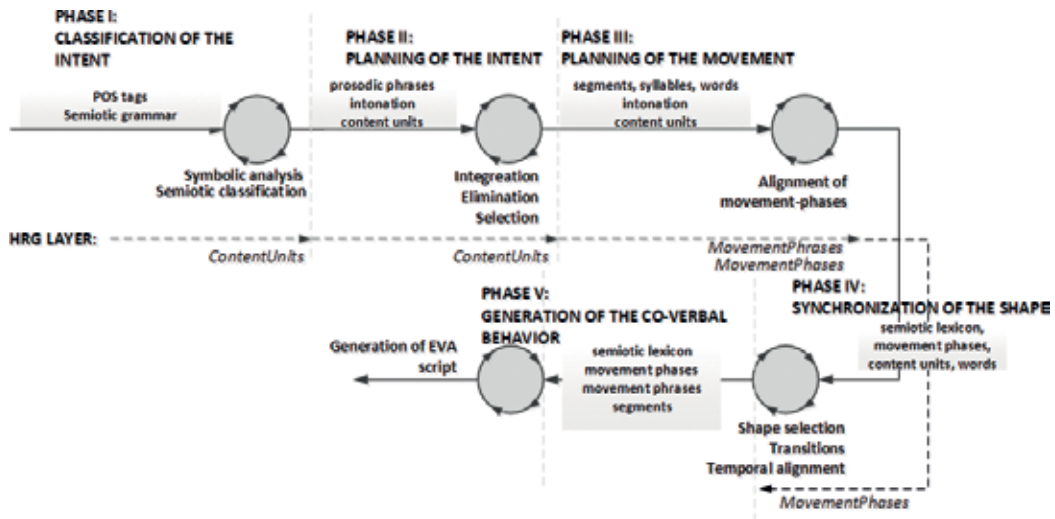
Figure 5. The algorithm for the generation of expressive co-verbal behavior.

the synchronization of intent and shape (the intent classification and planning), *the synchronization of several movements* (movement planning), and *prosodic and timing synchronization regarding speech and gestures* (synchronization of the form).

In phase I, named *classification of the intent*, the input is POS-tagged text and *semiotic grammar*. The *semiotic grammar* is used for mapping individual morphosyntactic sequence in the text onto a corresponding parametric description of the semiotic class (subclass) $Z = \{S, I, \omega_s, p_s\}$[1] In this way, we are able to perform the classification of the intent by defining the semiotic classes and corresponding cores of meaning. The algorithm searches for the longest morphosyntactic sequence x_{i-j} that can be found in the *semiotic grammar*, while the following two rules are implemented:

- *If at a specific word index the sequence x_A happens to be $x_A(S) \subseteq x_B(S)$, where both sequences belong to the same semiotic class, then sequence x_A must be discarded.*

- *If a sequence x at word index j is already contained in previous CU elements started at word index i and with the same semiotic intent (i < j), then it is discarded.*

In this way, for each such sequence, the content unit (*CU*) is created. The *CUs* are used to store the semiotic classification of the intent as well as those meaningful words that actually induce the shape in the stroke movement phase F_S of the gesture.

Nevertheless, sentences/statements can have multiple interpretations. Further, the classified *CU* interpretations can partly or fully overlap, therefore, introducing ambiguities and a number of inconsistencies. Thus, in phase II, named *planning of the intent*, these inconsistencies and

---

[1]Represents a unique set of features describing the semiotic class/subclass *S*, the index of the word(s) that represent a nucleus of meaning *I*, the distribution of the POS sequences for the specific semiotic class/subclass *ωs*, and the distribution of selecting the specific POS sequence among the semiotic classes/subclasses *ps.*

ambiguities have to be resolved by using *integration*, *elimination*, and *selection*. For this task the prosodic information (prominence, stress, prosodic phrases), as predicted by the TTS modules, is used. The prosodic information includes syllables labeled with *PA* (most prominent) and *NA* (stressed) tags that exist within minor phrases (*B2*) or major phrases (*B3*). The resolving is implanted by observing the following rules:

- *Each CU must include the most prominent syllable (PA) within a given prosodic phrase (B2 or B3), except in the case of enumeration.*

- *Each CU element must lie within the prosodic phrase (B2 or B3).*

- *Each prosodic phrase can be represented by no more than one concept of motion, i.e., with no more than one element CU.*

- *When the CU element contains the semiotic class enumeration, the CU boundaries must remain unchanged (the boundaries of prosodic phrases are not considered).*

At the end, the structure of the *intent* is uniquely defined, and the co-verbal behavior *G* can be represented as a sequence of co-verbal models *H* that are related with the *CU* as follows:

$$\hat{G} = T_m^{-1}\hat{H} = H[CU_1, t_1] \times H[CU_2, t_2] \times \ldots \times H[CU_n, t_n] \tag{1}$$

where $H[CU_i, t_i]$ describes the movement model that depends on semiotic classification and prosodic characteristics in each *CU* element. However, *internal overlapping* can still occur, when several *CU* elements contain one or more of the same words, while their boundaries lie within the same prosodic phrase. In this case we have to decide, which of the *CU* elements must be kept, since only one *CU* element is allowed within each prosodic phrase. Firstly, the algorithm removes all those *CU* elements that do not contain a word with a *PA* syllable, and if overlapping still exists, then on the remaining *CUs*, their normal distribution ω_s is considered, as calculated for the given semiotic class, when the *CU* with its maximum value is only kept. Therefore, common *CU* is created as:

$$CU_{S(CU_m)=S(CU_m)} = f(CU_m, \ldots, CU_n) = \\ \begin{cases} CU_m; \text{if } \omega_s(CU_m) > \omega_s(CU_n) \\ CU_n; \text{if } \omega_s(CU_m) < \omega_s(CU_n) \\ \max(len(CU_m), len(CU_n)); \text{if } \omega_s(CU_m) = \omega_s(CU_n) \end{cases} \tag{2}$$

Further, in the case of *ambiguity*, there are multiple *CUs* within a single prosodic phrase but without overlapping. When these *CUs* are consecutive and classify the same semiotic class *S* (e.g., represent the same communicative intent), they are merged into a single *CU* element as follows:

$$CU = f(CU_i, \ldots, CU_j) = \\ \begin{cases} \text{Join } CU_k, CU_{k+1}, \ldots; \text{ if} \\ \quad CU_k.POS = CU_{k+1}.POS \wedge S(CU_k) = S(CU_{k+1}) \\ \text{Add into a set; } \textit{otherwise} \end{cases} \tag{3}$$

The semiotic indicator *I* is defined as a set of corresponding semantic indicators contained within individual *CUs* in the sequence. Finally, *external overlapping* can occur, when some *CU* boundaries are stretched over the boundaries of the prosodic phrase. In this case the *CU* is kept only when the following rules are met:

- *The CU element includes the PA syllable, and this PA syllable lies within the boundaries of the prosodic phrase B2:* $PA \in B2 \wedge PA \in CU_i$.

- *The semiotic indicator I of the CU element lies within the boundaries of the prosodic phrase B2, i.e.,* $I \in B$.

And, only when both rules are met, the following two rules are implemented:

- *If* $CU \cap next(CU)$ *is not empty, the right boundary of the CU is set to the left boundary of the next (CU) element, in order that* $CU \cap next(CU) = \varnothing$ *is true.*

- *If* $CU \cap next(CU)$ *represents an empty set, the CU element is kept completely. Further, those words that lie outside the prosodic phrase that contains semiotic indicator I can only represent the holding phase and/or the retraction phase.*

In this case the common *CU* is created as:

$$CU = f(CU_i, CU_{i+1}) =$$
$$\begin{cases} keep\ \&\ reduce;\ if \\ \quad I \in B \wedge PA \in B \wedge PA \in CU_i \wedge CU_i \cap CU_{i+1} \neq \varnothing \\ keep\ \&\ extend;\ if \\ \quad I \in B \wedge PA \in B \wedge PA \in CU_i \wedge CU_i \cap CU_{i+1} = \varnothing \\ remove;\ if \\ \quad I \notin B \vee PA \notin B \vee PA \notin CU_i \end{cases} \quad (4)$$

In phase III, named *planning of the movement*, the movement models, which are based on *CU* units, are defined. A movement model is an animated sequence of shapes/poses that together represent a co-verbal expression. For each *H*, at least a stroke movement phase F_S has to be defined, which is aligned with the acoustic prosody information, as specified by the TTS engine. Therefore, the prosodic synchronization of movement phases is based on temporal information (regarding phoneme and pause durations). The following rule is used for the stroke movement phases F_S:

- *The stroke phase F_S is always performed on the PA word and is ended together with the corresponding PA syllable.*

The next step then represents the synchronization of all F_S with the gesticulation in case of *enumeration* and/or *search*, which are not directly related to the *PA* syllables, by using the following rule:

- *If the word that represents the semiotic indicator I for the specific CU does not contain the PA syllable, the NA syllable is considered in the same way instead.*

The most prominent words (with the *PA* syllable) do not necessarily represent the semiotic indicator *I* for the given *CU* element. If this is the case, the following rule is applied:

- *If the* semiotic *indicator I and the PA syllable are not represented by the same word, the stroke phase F_S must be defined by following the previous rules, but the hold movement phase F_{SH} must end with the NA syllable of the word that represents the semiotic indicator I within the given CU.*

Within the algorithm movement models, *H* is represented by movement phrase units (*MPHRs*), where each unit can contain several movement phases (*MPHs*). Further, each *MPHR* must contain at least stroke phase F_S. The syllables that occur before the stroke phase F_S are used for the preparation movement phase F_P, while the *sil* segment just before the first syllable of the F_S can be used for the hold movement phase F_{HS} (hold before stroke). And, those syllables after the F_S are used for the retraction movement phase F_R. In this way, the behavior structure is applied by the following rules:

- *The* preparation *phase F_P starts before the stroke phase F_S and lasts from the NA syllable to the beginning of the F_S.*

- *The sil segment, which can have a nonzero duration between the words with the preparation phase F_P and the stroke phase F_S, represents the so-called hold before stroke, which (if it exists) represents a ready-made idea regarding the content.*

Additionally, the created *MPHs* can be extended or merged by the following rules:

- The right boundary of movement phase (with the exception of the hold phase FH) must be a PA or NA syllable.

- The stroke phase FS and the preparation phase FP can be joined into the stroke phase FS, as this often occurs in multimodal communication (as observed in database annotations).

Nevertheless, the extensions are always limited by the boundaries of the specific *CU*. The structure of the movement models *H* is now synchronized with verbal content on the symbolic level. Further, temporal synchronization is performed by considering the durations of phonemes and pauses. The duration of individual movement phase is described by the following sum of the syllable durations that they may include:

$$t(MPH_i) = \left(\sum_{j=0}^{n-1} t\left(syl\, l_{j+k}\right) \right) \tag{5}$$

where *n* represents the number of syllables in each *MPH* unit and *k* its first syllable. F_P can be fused with the F_H phase, resulting in the following maximal duration:

$$t_{max}(F_P) = t(F_P) + \sum_{j=k}^{n-1} t\left(z_{k-j}\right) + t\left(z_r\right) \tag{6}$$

where *n* represents the number of syllables before the F_P, *k* its first syllable, and *r* the first syllable after the F_P. Further, F_S can be fused with F_P, resulting in the maximal duration:

$$t_{\max}(F_S) \ = \ t_{\max}(F_P) + t_{\min}(F_S) \tag{7}$$

Finally, F_R can be extended with subsequent syllables but only up to the last *NA* syllable:

$$t_{\max}(F_R) \ = \ t_{\min}(F_R) + \sum_{j=0}^{n-1} t(z_{k+1}) \tag{8}$$

where *n* represents the number of subsequent syllables, while *k* is the first syllable after the F_R. Temporal descriptions of movement phases define time instants when the individual shape must be fully manifested and also the time that is available for the transition between the shapes. Further, this time also restricts the set of suitable motion trajectories for the transition, as well as restricting the set of shapes.

In phase IV, named *synchronization of the shape*, the movement is then temporally aligned with the temporal features of verbal information (durations of phonemes and pauses). Based on morphosyntactic sequences, movement models, and durations of the movement phases, a lookup into *Gesticon* is carried out, in order to specify the best shapes *V* (or poses *P*) and trajectories *T* of the realization of the co-verbal behavior. Thus, a lookup for possible configurations of the embodiment within F_S phases is performed. It selects a set of probable poses *P* for each F_S. These poses are evaluated by using the *suitability functions* [43]. If there are no matched poses in the *Gesticon*, the set of most appropriate poses is selected by the CART (classification and regression tree) model, while the most appropriate pose *P* is assigned to each F_S. After defining the pose candidates on each stroke phase F_S, the poses for F_P, F_R, and F_H are also defined. At the end, the transition between the poses are also defined and aligned with given temporal and prosodic specifications. Namely, the trajectory describes the local space in which body part should move when traversing from the start to the end pose. The huge diversity of trajectories within the material demands restrictions, when describing them in the *Gesticon*. We are specifically interested in the trajectories of hands or the curve that the hand describes during the transition. The definition of the trajectory between two poses is performed by considering the temporal structure of the movement phase *MPH*, the semiotic class, the movement phase type, the morphosyntactic tags, prosodic features within the phase, and possible semantic relation. The lookup in the *Gesticon*, therefore, results in several possible trajectories. Therefore, only the most appropriate and closest to the temporal predictions on each sequence is used at the end.

In phase V, named *generation of the co-verbal behavior G*, the conversion of the defined movement models (stored within the heterogeneous relation graph (HRG)) into a procedural description can be understood as a parameterization of the animation. Each movement phrase is transformed into a symbolically, prosodically, and spatially coherent movement of an individual body part. Thus, it viably illustrates the communicative intent of the corresponding verbal segment. In order to be applied on an ECA and represented to the user, it has to be converted into a procedural description in the EVA-Script notation. Each model *H* represents simultaneous execution and is described within the block *<bgesture>*. The stroke movement phase F_S and the preparation movement phase F_P within the block *<bgesture >* represent sequences during which a change in configuration of embodiment

is actually requested. The hold movement phase F_H and the retraction movement phase F_R, however, do not require procedural description. Namely, F_H only represents a hold of the existing configuration, while F_R a retraction into a rest/neutral state. The transformation of movement model H into an EVA event (co-verbal behavior written in EVA-Script) is outlined in **Figure 6**. **Figure 6** also outlines how EVA event is applied to an ECA and then realized as a multimodal expression, which is built from the synchronized verbal and co-verbal sequences. As can be seen, the F_P is defined across the word "*bila*" (*was*), with predicted extension up to the word "*je*" (*is*). Further, the predicted duration of the F_P phase is 413 ms, while the maximum duration of the F_P is 593 ms. The F_S is defined across the *PA* word "*tako*" (*that*), with a predicted duration 300 ms and a maximal duration 893 ms. The post-stroke-hold phase F_{HS} is identified across the semiotic nucleus, the word "*velika*" (*big*), with a duration of 451 ms.
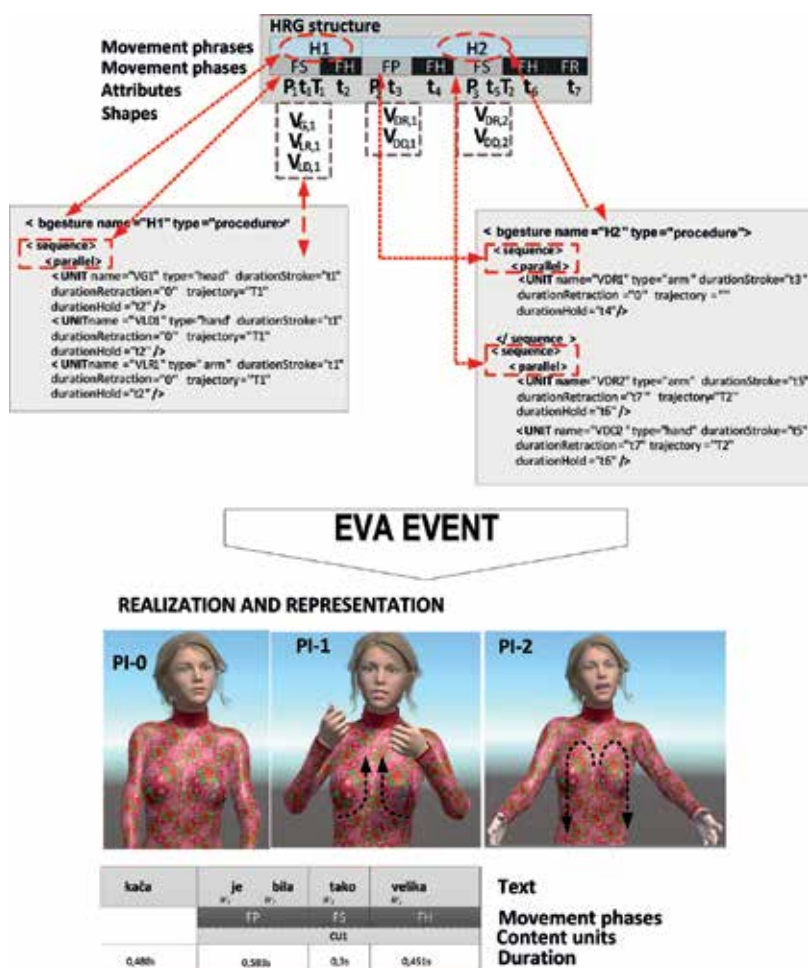
Figure 6. Realization of a sentence with the conversational agent EVA.

5. Realization of expressive conversational behavior on embodied conversational agents

The proprietary *EVA realization framework* proposed in [36–37, 44] has been developed in order to be able to evoke a social response in human-machine interaction based on expressive conversational behavior generated through the previous modules of the EVA model. The framework enables machines to engage with the user on the more personal level, namely, through humanlike entity realization of multimodal interaction models, which are based on the concept of conversation. Thus, this framework integrates ECAs as virtual bodies and generates responses via their embodiment. The ECA's artificial body and articulation capabilities (embodiment) are already close to those found on real humans. From the skin, face, hands, and body posture, these virtual entities tend to look and behave as realistically as possible. ECAs also tend to imitate as many features of human face-to-face dialogs as possible and integrate them into interaction as synchronized as possible. Although the humanlike equation is mostly defined via co-verbal behavior generation model and the corpus analysis, the framework actually represents the final component through which users actually come in contact with the response. Thus, one could say that the framework brings responses to "life." Further, diversity and capacity to handle highly dynamic and interchangeable contexts of human interaction are in addition to realism of appearance, one of the key challenges of the modern ECAs. 3D tools, such as Maya, Daz3D, Blender, Panda3D, and Unity, have opened up completely new possibilities to design virtual entities, which appear almost like real-life persons. The modern behavior generators open the possibility to plan and model responses almost completely to the context of situation and collocutor. The behavior realizer, therefore, represents the bridge between the two concepts. The *EVA realization framework* also creates an environment that is capable to deliver expressive and socially colored responses in the form of facial expressions, gaze, head, and hand movement. Its architecture is outlined in **Figure 7**.

It consists of animation-parameters builder, animation-realization engines, articulated 3D models, and created 3D resources. The *animation-parameters builder* is used to understand and transform the co-verbal events into animation parameters and integrate them onto the animation execution plan. The *animation-realization engines* then realize these animation parameters through their internal renderers and display them to the user. As outlined in **Figure 7**, two animation engines are implemented, one is based on Panda 3D[2] game engine and the other based on Unity 3D game engine[3]. Each of them incorporates its own set of articulated 3D models. However, all articulated 3D models support the same movement controllers (bones and morphed shapes). Thus, any EVA event can be used by either realizer, and the result will be still practically the same. The major difference between the supported animation engines is their implementation of frame-by-frame operations. Namely, in the Panda 3D engine, frame-by-frame operations are handled internally by the renderer, while in the Unity 3D engine, the renderer only renders each frame. This means that all

---

[2]Panda 3D: https://www.panda3d.org/.
[3]Unity 3D: https://unity3d.com/.
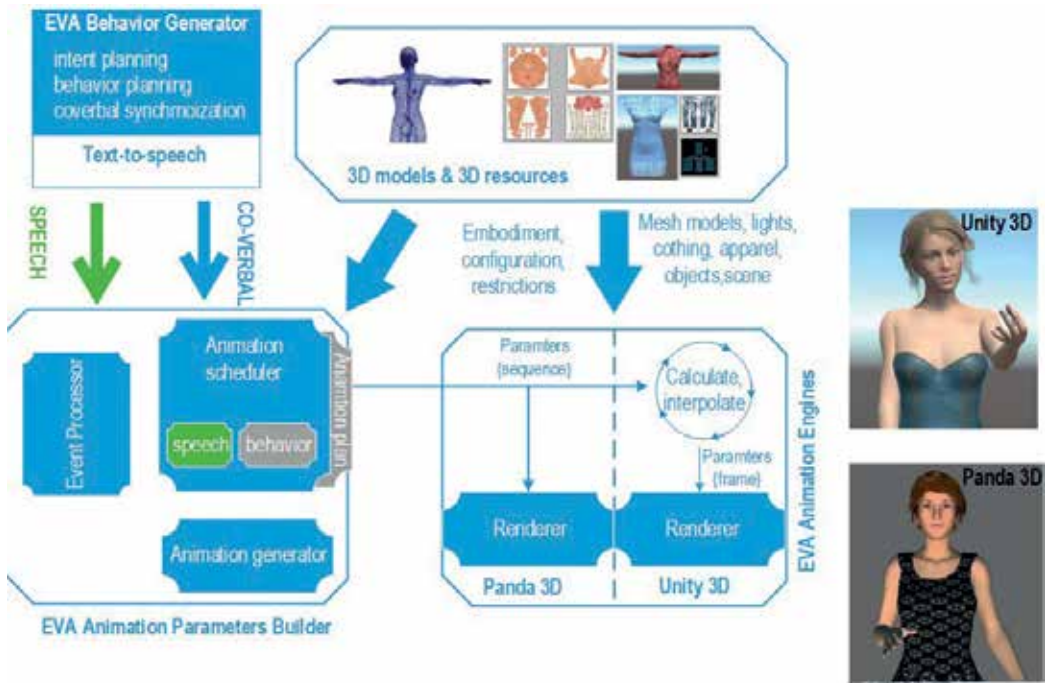
Figure 7. EVA realization framework.

calculations for the next "in-between" pose are calculated via the implanted algorithms. As a result, the Unity 3D implementation allows for controlling the scheduled animation and even animation already being executed. In order that the *EVA realization framework* realizes the generated synchronized behavior and represents it to the user, the *EVA behavior realizer* transforms conversational events into their physical representations. This is achieved by applying the co-verbal features described in the co-verbal events into the 3D resources available in the renderer. The *animation-parameters builder* translates the EVA-Script into animation parameters. This is achieved by interfacing each script's tag with the control unit or behavioral template and by extrapolating different groups of movements. Each group of movement is defined by semantic (which control units in which order), temporal (durations of stroke, hold, and retraction movement phases), and spatial features (ending position of the control unit). The main components of the *animation-parameters builder* are *event processor*, *animation generator*, and *animation scheduler*. The *event processor* intercepts and handles the conversational events. It parses event stream and checks event's type and priority. The *animation generator* then transforms the conversational behavior into animation sequences. As part of this process, the *animation generator* applies temporal and spatial constraints adjusted to the agents' articulated body. The a*nimation scheduler* then inserts the generated animation sequences into the execution plan. It handles the animation graph and feeds them to the rendering engines accordingly. Finally, after the realization of each animation sequence is completed, the *event processor* signals its status (conversational context) to the

behavior generator and optionally to the dialog handlers. Similarly, after the full realization of the behavior que, a change in conversational context event is raised, and the generation of inactive (rest) behavior is triggered.

The communication between processes within the *EVA realization framework* is implemented via event-oriented publish/subscribe model. Namely, when the *event processor* intercepts a conversational event, it firstly checks its type and priority. Afterward, it pushes it into the *animation scheduler*. When the *animation scheduler* receives the conversational event, it initiates internal interpreter in order to segment the behavior into three animation streams. The inter-preter transforms the EVA-Script behavior into body part-segmented schedule of parallel/consecutively executed behavior in a form of animation streams. At the same time, the *sched-uler* smoothly stops any idle behavior, destroys its handlers, and moves it to the rest pose. The overall result of the *animation-parameters builder* is, therefore, a set of animation parameters, which describe the execution of one or more animations representing the planned co-verbal act (multimodal response). The animation parameters are those features that specify how the animation engine should build its animation graph. The animation parameters are "fed" to the second component of the framework, the *EVA animation engine*. The *EVA animation engine* takes care for the transformation of animation parameters into animated sequences. The animation plan contains the co-verbal behavior que. After its que is emptied, the *anima-tion scheduler* signals that the animation stream has been completed and will destroy ani-mator objects, in order to release the reserved resources. After all animation segments are completed, the *animation scheduler* signals the end of the conversational event. As a result, the *event handler*, if no more co-verbal events arrive, triggers the manifestation of the idle behav-ior. Each animation engine transforms the animation parameters maintained in the anima-tion plans into corresponding sequential and/or parallel movements of control points (bones or morphed shapes). Further, both animation engines provide the forward kinematics (and inverse kinematics) and animation-blending techniques, which enable for animation param-eters to appear as viable behavior even on segments that have to be controlled by different gestures at the same time, e.g., simultaneously animating smile and viseme. Each gesture/expression/emotion is realized by combining different sequential/concurrent transformations of different movement controllers (embodiments) of the ECA. The EVA-Script events describe the facilitation of the movement controllers in a form of temporally defined end poses. Thus, each entry contains "next" configuration, which should be rendered over the specified tem-poral distribution. The in-between frames, which actually generate movement, however, are calculated and interpolated by animation engines. In case of Panda 3D engine, the render receives the required end pose and calculates the in-between frames automatically, while in the case of Unity 3D engine, the *animation handler* handles frame-by-frame operations, e.g., the render receives the next in-between configuration, which is calculated by the *animation handler* at each frame. In this way, any animation can be modified at any step, even during the execu-tion of some step/sequence. For a smooth transition, the scheduler does not have to wait and adjust its temporal scheduling. It just has to adjust its frame-by-frame schedule and replace it with new configurations. It can actually instantiate changes instantly as they occur. It can also insert new behavior between configurations, etc. As a result, the virtual character becomes more responsive and can react to changes of the conversational, environmental, and other

contexts instantly. The agent also "remembers" what it was gesturing prior to the excited state. Additionally, it can continue with the realization of that behavior after the excited state dissipates.

6. Realization of complex co-verbal acts

When the realization framework receives some co-verbal EVA event, it transforms it into a synchronized and fluid stream of movement, performed by the following independent body parts: hands, arms, face, and head. To retain naturalness (especially regarding visual speech coarticulation) and at the same time prevent "jerky" expressions, the animation-blending techniques are used. The same animation-blending techniques are also used to realize three different types of complex emotions: emotion masking, mixed expressions, and qualified emotions. These complex emotions further intensify the expressive factor of the framework and enable the implantation of highly humanlike representation of feelings in facial region (e.g., by modulating, falsifying, and qualifying an expression with one or more elemental expressions). The modulation of expression is realized by using animation-blending techniques, based on intensification or de-intensification. Both are similar to qualification of expression and implemented through the power (e.g., stress attribute) and temporal (durationUp, durationDown, delay, and persistence attributes) components of the domain of expressivity [36–37, 44]. **Figure 8** outlines output from the *realization framework* as interpretation and realization of EVA-Script events, including several layers of complexity, and EVA-Script attributes described through the EVA-Script language. In **Figure 8**, the behavior generator defines a co-verbal act that consists of two co-verbal events. The first one resembles the end of "searching idea" event (when some idea of a solution comes to our mind), and the second one reassembles the beginning of revelation of the idea (e.g., how one starts outlining the solution to collocutors). The co-verbal behavior is described in order to be performed via full embodiment (all co-verbal artifacts), namely, by using the arms, hands, face, and head. To add an additional layer of complexity, the inner synchronization and the temporal distribution are different for each co-verbal artifact.

During the first co-verbal event (e.g., revelation), the head, face, and right hand are the dominant artifacts. Thus, they appear to "move" with most significance and power. On the other hand, the left hand moves to its targeted position slightly delayed but as fast as possible. In the second act, however, the left hand is the dominant artifact. Thus, its movement will appear as most significant, e.g., the longest duration and with most power, while the face/head and right arm/hand are moved to the position as "quietly" as possible. The overall duration of the first co-verbal act (Act 1) act is 1.567 s. During this time period, the agent has to perform a pointing gesture, by pointing to the sky and by moving its left arm to a position that is relevant for the specified pointing gesture (e.g., almost touching the torso). Additionally, the agent should express a blend of happy/surprised emotion on his/her face. As outlined in **Figure 8**, head/gaze and facial expression started to appear first (delay = 0.0 s). The two co-verbal artifacts then moved to their final configuration in 0.5 s, while the right- and left-hand movements are delayed for 0.4 s. Thus, both configurations started to form 0.1 s interval,
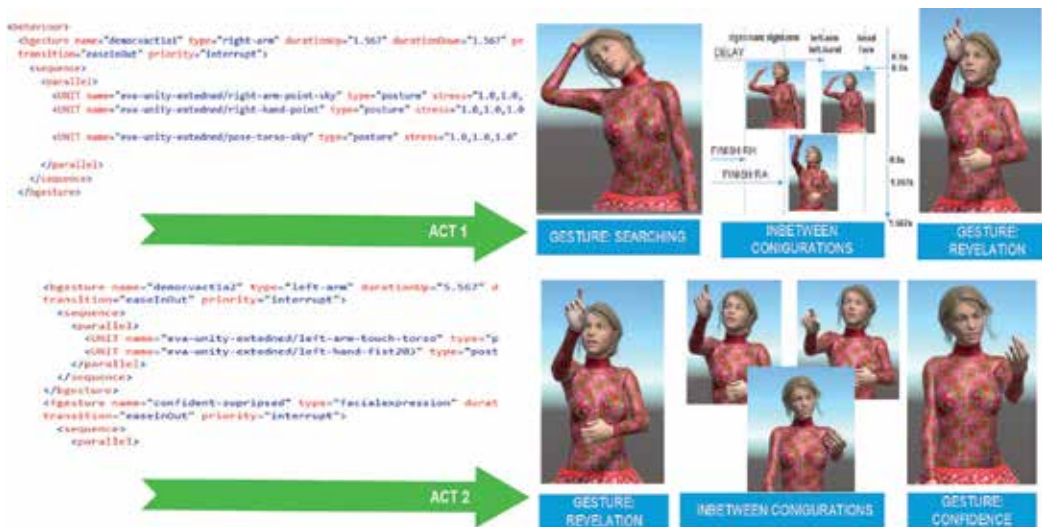
Figure 8. Realization of EVA event on ECA EVA rendered in unity-based realizer.

before the previous two co-verbal artifacts finished. The right arm finished with its animation after 0.567 s, while the right hand manifested the targeted hand shape in 0.3 s. The left hand and arm propagated to their end-configuration until the overall end of the event (e.g., for 1.167 s). Those co-verbal artifacts, which have already finished, just maintained their configuration. The second co-verbal act is targeted to last 5.567 s. During the realization of the second co-verbal act, the right arm (with hand) is regarded as less significant; therefore, it is moved to its intended rest position as slowly as possible. The left arm (with hand) is, in this situation, regarded as one of the significant co-verbal artifacts carrying some conversational meaning. The same holds true for the head and face. The left-hand movement in this case appeared as with most power in order to gain the most attention of the collocutor, and the face expressed confidence colored with excitement. Finally, by directing gaze to the collocutor, the ECA EVA prepares the conversational environment, which facilitates the full attention of the collocutor. Thus, it can start with the representation of the recently developed idea.

7. Conclusion

Natural interaction entails multiple behavior variations correlated in a dynamic and highly unpredictable setting. It also incorporates various social and interpersonal signals to "color" the final outcome. Furthermore, multimodality in interaction is not just an add-on or a style of information representation. It goes well beyond semantics and even semiotic artifacts. It significantly contributes to representation of information as well as in interpersonal and textual function of communication. In this chapter we have outlined approach to automatic synthesis of more natural humanlike responses generated based on EVA conversational model. The presented model consists of three interlinked and repetitive frameworks. The

first framework involves conversational analysis through which we analyze multiparty and spontaneous face-to-face dialogs in order to create various types of conversational resources (from rules and guidelines to complex multidimensional features). The second framework then involves an omni-comprehensive algorithm for the synthesis of affective co-verbal behavior based on the arbitrary and unannotated text. In contrast to the related research, the proposed algorithm allows for the conversational behavior to be driven simultaneously by prosody and text and modeled by various dimensions of situational, inter- and intrapersonal contexts. Finally, the predicted behavior well synchronized to its verbal counterparts has to be represented to a user in a most viable manner. Thus, the third framework in the proposed model involves implementation of co-verbal behavior realizer. In our case we have decided to fuse advantages of state-of-the-art 3D modeling tool and game engines with the latest concepts in behavior realization in order to deploy an efficient and highly responsive framework through which the generated co-verbal expressions may be represented to users via realistic and humanlike embodied conversational agents. Namely, modern behavior realizers have the capacity to support several parameters of believability of conversational behavior, such as diversity and multimodal planning, situational awareness, synthesis of verbal content, synchronization, etc. The game engines on the other hand are a powerful tool for rapid and high-quality design and rendering of virtual humanlike entities including ECAs. They enable the design and delivery of beautiful and highly realistic graphics and the efficient handling of hardware resources. To sum up, the ability to express information visually and emotionally plays a central role in human interaction and thus in defining ECA's personality, its emotional state, and can make such an agent an active participant in a conversation. However, in order to make him be perceived even more natural, the agent must be able to respond to situational triggers smoothly and almost instantly as they are perceived and by facilitating synchronized verbal and co-verbal channels. Thus, the presented model presents an important step toward generating more natural and humanlike companions and machine-generated responses.

Acknowledgements

Author details

Matej Rojc*, Zdravko Kačič and Izidor Mlakar

*Address all correspondence to: matej.rojc@um.si

Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia

References

[1] Luger E, Sellen A. Like having a really bad PA: The gulf between user expectation and experience of conversational agents. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM. 2016. pp. 5286-5297

[2] Feyaerts K, Brône G, Oben B. Multimodality in interaction. In: Dancygier B, editor. The Cambridge Handbook of Cognitive Linguistics. Cambridge: Cambridge University Press; 2017. pp. 135-156. DOI: 10.1017/9781316339732.010

[3] Li J, Galley M, Brockett C, Spithourakis GP, Gao J, Dolan B. A Persona-Based Neural Conversation Model. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany; 2016; pp. 994-1003

[4] Porcheron M, Fischer JE, McGregor M, Brown B, Luger E, Candello H, O'Hara K. Talking with conversational agents in collaborative action. In: Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. ACM. 2017. pp. 431-436

[5] Bonsignori V, Camiciottoli BC, editors. Multimodality Across Communicative Settings, Discourse Domains and Genres. Cambridge Scholars Publishing; 2016. Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK. ISBN (10): 1-4438-1107-6. ISBN (13): 978-1-4438-1107-1

[6] Kopp S, Bergmann K. Using cognitive models to understand multimodal processes: The case for speech and gesture production. In: The Handbook of Multimodal-Multisensor Interfaces. New York, NY, USA: Association for Computing Machinery and Morgan & Claypool; 2017. pp. 239-276

[7] McNeill D. Why We Gesture: The Surprising Role of Hand Movement in Communication. Cambridge: Cambridge University Press; 2016. ISBN-10: 1316502368. ISBN-13: 978-1316502365

[8] Davitti E, Pasquandrea S. Embodied participation: What multimodal analysis can tell us about interpreter-mediated encounters in pedagogical settings. Journal of Pragmatics. 2017;**107**:105-128

[9] Hazel S, Mortensen K. Embodying the institution—Object manipulation in developing interaction in study counselling meetings. Journal of Pragmatics. 2014;**65**:10-29

[10] Vannini P, Waskul D, editors. Body/Embodiment: Symbolic Interaction and The Sociology of the Body. New York, NY, USA: Ashgate Publishing, Ltd.; 2012. ISBN: 1409490610, 9781409490616

[11] Colletta JM, Guidetti M, Capirci O, Cristilli C, Demir OE, Kunene-Nicolas RN, Levine S. Effects of age and language on co-speech gesture production: An investigation of French, American, and Italian children's narratives. Journal of Child Language. 2015;**42**(1):122-145

[12] Esposito A, Vassallo J, Esposito AM, Bourbakis N. On the amount of semantic information conveyed by gestures. In: 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI); IEEE. 2015. pp. 660-667

[13] Kendon A. Gesture: Visible Action as Utterance. Cambridge University Press; 2004. ISBN 0 521 83525 9. ISBN 0 521 54293 6

[14] Zhao R, Sinha T, Black AW, Cassell J. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In International Conference on Intelligent Virtual Agents; Springer International Publishing. 2016. pp. 218-233

[15] Pejsa T, Gleicher M, Mutlu B. Who, me? How virtual agents can shape conversational footing in virtual reality. In: International Conference on Intelligent Virtual Agents; Cham: Springer. 2017. pp. 347-359

[16] Allwood J. A framework for studying human multimodal communication. In: Coverbal Synchrony in Human-Machine Interaction. Boca Raton; London; New York: CRC Press; 2013. cop. 2014. XIV, 420 str., ilustr. ISBN 1-4665-9825-5. ISBN 978-1-4665-9825-6

[17] Bozkurt E, Yemez Y, Erzin E. Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures. Speech Communication. 2016;**85**:29-42

[18] Chen CL, Herbst P. The interplay among gestures, discourse, and diagrams in students' geometrical reasoning. Educational Studies in Mathematics. 2013;**83**(2):285-307

[19] Holler J, Bavelas J. In: Breckinridge Church R, Alibali MW, Kelly SD, editors. Multimodal Communication of Common Ground. Why Gesture? How the Hands Function in Speaking, Thinking and Communicating. Vol. 7. 2017. pp. 213-240

[20] Poggi I. Hands, Mind, Face and Body: A Goal and Belief View of Multimodal Communication. Berlin: Weidler; 2007. ISBN (10): 3896932632. ISBN (13): 978-3896932631

[21] Yumak Z, Magnenat-Thalmann N. Multimodal and multi-party social interactions. In: Context Aware Human-Robot and Human-Agent Interaction. Switzerland: Springer International Publishing; 2016. pp. 275-298

[22] Kuhnke F, Ostermann J. Visual speech synthesis from 3D mesh sequences driven by combined speech features. In 2017 IEEE International Conference on Multimedia and Expo (ICME). IEEE. 2017. pp. 1075-1080

[23] Peng X, Chen H, Wang L, Wang H. Evaluating a 3-D virtual talking head on pronunciation learning. International Journal of Human-Computer Studies. 2018;**109**:26-40

[24] Wang N, Ahn J, Boulic R. Evaluating the sensitivity to virtual characters facial asymmetry in emotion synthesis. Applied Artificial Intelligence. 2017;**31**(2):103-118

[25] Gibet S, Carreno-Medrano P, Marteau PF. Challenges for the animation of expressive virtual characters: The standpoint of sign language and theatrical gestures. In: Dance Notations and Robot Motion. Switzerland: Springer International Publishing; 2016. pp. 169-186

[26] Tolins J, Liu K, Neff M, Walker MA, Tree JEF. A verbal and gestural corpus of story retellings to an expressive embodied virtual character. In LREC. 2016

[27] Ochs M, Pelachaud C, Mckeown G. A user perception-based approach to create smiling embodied conversational agents. ACM Transactions on Interactive Intelligent Systems. 2017;**7**(1):33. DOI: DOI: 10.1145/2925993, article 4 (January 2017)

[28] Bellamy RK, Andrist S, Bickmore T, Churchill EF, Erickson T. Human-agent collaboration: Can an agent be a partner? In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. ACM. 2017. pp. 1289-1294

[29] Neff M. Hand Gesture Synthesis for Conversational Characters. Handbook of Human Motion. Switzerland: Springer International Publishing; 2017. pp. 1-12. ISBN: 978-3-319-30808-1

[30] Provoost S, Lau HM, Ruwaard J, Riper H. Embodied conversational agents in clinical psychology: A scoping review. Journal of Medical Internet Research. 2017;**19**(5):e151, pp.1-17

[31] Rojc M, Presker M, Kačič Z, Mlakar I. TTS-driven expressive embodied conversation agent EVA for UMB-SmartTV. International Journal of Computers and Communications. 2014;**8**:57-66

[32] Shaked NA. Avatars and virtual agents—Relationship interfaces for the elderly. Healthcare Technology Letters. 2017;**4**(3):83-87

[33] Mlakar I, Kačič Z, Rojc M. A corpus for investigating the multimodal nature of multi-speaker spontaneous conversations–EVA corpus. WSEAS Transactions on Information Science and Applications. 2017;**14**:213-226. ISSN 1790-0832

[34] Mlakar I, Kačič Z, Rojc M. Describing and animating complex communicative verbal and nonverbal behavior using Eva-framework. Applied Artificial Intelligence. 2014;**28**(5): 470-503

[35] Shamekhi A, Czerwinski M, Mark G, Novotny M, Bennett GA. An exploratory study toward the preferred conversational style for compatible virtual agents. In: International Conference on Intelligent Virtual Agents. 2016. pp. 40-50

[36] Rojc M, Mlakar I, Kačič Z. The TTS-driven affective embodied conversational agent EVA, based on a novel conversational-behavior generation algorithm. Engineering Applications of Artificial Intelligence. 2017;**57**:80-104

[37] Rojc M, Mlakar I. An expressive conversational-behavior generation model for advanced interaction within multimodal user interfaces. In: Computer Science, Technology and Applications. New York: Nova Science Publishers, Inc.; 2016, cop. XIV, p. 234 str. ISBN 978-1-63482-955-7. ISBN 978-1-63484-084-2

[38] Pelachaud C. Greta: An interactive expressive embodied conversational agent. In: Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems; May 2015. pp. 5-5

[39] Mondada L. New challenges for conversation analysis: The situated and systematic organization of social interaction. Langage et Societe. 2017;**2**:181-197

[40] Velentzas J, Broni DG. Communication cycle: Definition, process, models and examples. In: Proceeding of the 5th International Conference on Finance, Accounting and Law (ICFA" 14); Vol. 17. 2014. pp. 117-131

[41] Mlakar I, Kačič Z, Rojc M. Form-Oriented Annotation for Building a Functionally Independent Dictionary of Synthetic Movement, Vol. 7403. Berlin; New York: Springer; 2012. pp. 251-265

[42] Rojc M, Mlakar I. Multilingual and multimodal corpus-based text-to-speech system PLATTOS. In: Ivo I, editor. Speech and Language Technologies. Rijeka: InTech; 2011. ISBN: 978-953-307-322-4

[43] Rojc M, Kačič Z. Gradient-descent based unit-selection optimization algorithm used for corpus-based text-to-speech synthesis. Applied Artificial Intelligence. 2011;**25**(7):635-668

[44] Mlakar I, Kačič Z, Borko M, Rojc M. A novel unity-based realizer for the realization of conversational behavior on embodied conversational agents. International Journal of Computers. 2017;**2**:205-213. ISSN: 2367-8895

High Performance Technology in Algorithmic Cryptography

Arturo Lezama-León, José Juan Zarate-Corona,
Evangelina Lezama-León,
José Angel Montes-Olguín,
Juan Ángel Rosales-Alba and
Ma. de la Luz Carrillo-González

Additional information is available at the end of the chapter

Abstract

Alan Turing's article, "Computation and intelligence", gives the preamble of the characteristics of guessing if it is a machine or another human being. Currently, the use of ubiquitous technologies, such as the use of firmware, allows direct access to analog data, however, we must find a way to secure the information. Analyzing cryptographic algorithms for the transfer of multimedia information. Raise the use of cryptarithmetic. Finite automata will be developed that will govern the logic of the cryptographic algorithms to be integrated into Firmware, performance tests and controls will be carried out to determine the best strategies for their performance and algorithmic complexity. Technologies are expressed that allow the creation of learning environments, such as neural networks, that support other processes as the recognition of patterns on images.

Keywords: cryptarithmetic, cryptographic algorithms, firmware, HPC

1. Introduction

In this research, it is revealed how algorithms are integrated into different representations of the data, information seen as signals, images, video and text.

1.1. Turing machines

Alan Turing's test, proposed in 1950, was designed to provide an operational and satisfactory definition of intelligence.

He suggested a test based on the inability to differentiate between indisputable intelligent entities and human beings.

The computer should have the following capabilities

- Natural language processing, which allows you to communicate in English.

- Representation of knowledge, to store what is known or felt.

- Automatic reasoning, to use stored information to answer questions and draw new conclusions.

- Machine learning, to adapt to new circumstances and to detect and extrapolate patterns.

- Computational vision, to perceive objects.

- Robotics, to manipulate and move objects.

These six disciplines cover most of the Artificial Intelligence (AI), in reference [1, 2].

1.2. Artificial intelligence and the cryptarithmetic

The Artificial Intelligence is the study of how to make computer do things which, at the moment, people do better, in Ref. [2].

1.2.1. cryptarithmetic

It considers high order constraints which can be represented as a collection dev binary options F $\Leftrightarrow$ T, in Ref. [3].

1.3. Automata design

An automaton is a graphic representation of a process, which simulates a sequential process, by means of a deterministic finite automaton, the use of a regular expression can be modeled, and this, in turn, can be evaluated by algorithmic complexity.

There is a counterpart which obeys different inputs at the beginning and consequently can represent several outputs, so we proceed by means of a mechanism to migrate them as a finite deterministic automaton, similarities with neural networks in terms of structure, but not functioning because they pursue different aspect.

Regarding its sequential form, it can be simulated by means of a flow diagram as a graphic form, however, in its parallel form, it must be used either an activity diagram or a Petri net, which can model the concurrent behavior, in reference [4].

Because the keys are usually registered by characters of the ASCII code, they can be validated by regular expressions and this in turn then recreating validation controls by deterministic finite automata.

When a possible entry is described under this form, then we are in the presence of an entry control but not access.

For access, we search for a match of a stored key to managing access to the data, if the password is treated by a cryptographic system then the entry and access controls do not change their form, but the time of computation to corroborate its degree of efficiency.

By having access control list (ACL), the restrictions are carried out through the management of routing tables and entry rules.

The use of ACL increases the level of security in information systems.

The use of Firewalls allows having an accurate strategy as if it were a blog.

Cryptographic systems have been implemented to treat various types of data and signals such as multimedia so that not only text can be treated.

In this way its study is of interest because in applications oriented to the control and automation of signals in greenhouses, home automation, automation, automation and control, information security in closed or public environments, the integration of microcontrollers is intended which can supply to a large extent, restrictive measures which make their assurance slow.

In this research, it is revealed how algorithms are integrated into the different representations of the data, information seen as signals, images, video, and text.

The signals are presented in the form of a 1D dimension, two 2D dimensions, and three 3D dimensions. It is important to understand that as the degree of complexity is advanced, it increases as the subject deserves or intends for an even more exhaustive study.

This research does not intend to venture to deny or violate systems, but as a form of prevention, to continue with research studies, still need to specify the cryptanalysis for each system, only the studies discussed so far are presented.

1.4. Computer legislation and regulations

The use of technology in today's world is inevitable. Whether it is making reservations on our smartphones, or checking emails, or checking in for flights, usage of technology is present. Whilst its benefits cannot be questioned, unfortunately, the increase of our reliance on technology implies that we are at higher risk of attack and breaches – cyber-attacks. Responding to this current scenario, current trends of governments protecting their critical infrastructures is the implementation of cybersecurity standards to their critical sectors, in reference [5]. The Global Risk 2014 by World Economic Forum (WEF) shows the 'cyber-attacks' listed in the top 5 global risks, highlighting that the dependency on technology by economies and societies is inevitable, in reference [6].

The implementation of cybersecurity standards is by no means a silver bullet in critical infrastructure protection. However, its implementation can establish a set of controls that contribute and build better resiliency. The cybersecurity standards may support the capabilities of preparing, protecting, responding and recovering from cyber-attacks. Some of the common cybersecurity-related standards being implemented globally include the following (not exhaustive): ISO/IEC 27032:2012. Information technology – Security techniques

– Guidelines for cybersecurity. ISO/IEC 27001 Information technology – Security techniques – Information security management systems-Requirements. ISO 22301 Societal security – Business continuity management systems-Requirements. ISO/IEC 15408 Information technology – Security techniques – Evaluation criteria for IT security. ISO/IEC 27035 Information technology – Security techniques – Information security incident management. ISO/IEC 27005 Information technology – Security techniques – Information security risk management. FIPS 140-1: Security Requirements for Cryptographic Modules. FIPS 186-3: Digital Signature Standard, in reference [5].

Countries take different approaches towards implementing cybersecurity standards in the efforts of protecting their critical infrastructures. Some countries implement cybersecurity standards through mandatory requirements, whilst others provide guidelines and frameworks, in reference [5].

The United Kingdom's priorities for action is to model the best practice on cyber security in the government systems which will set the standard for suppliers to government to raise the bar on cybersecurity requirement.

Due to this:

1. UK has enforced compliance to the Network Interoperability Consultative Committee (NICC) Minimum Security Standards (ND 1643) through the Communications Act 2003 to the Communications Sector,

2. UK has developed the cyber security and information assurance standards – Information Assurance Maturity Model (IAMM) incorporating the requirements from the Security Policy Framework (SPF),

3. SPF recognizes and has aligned its principles to the ISO/IEC 27001 and the Business Continuity Management (BCM) standards (BS 25999/ISO 22301), and.

4. The application of standards is promoted through the establishment of national-level certification schemes for the following standards such as the ISO/IEC 15408, ISO/IEC 27001, ISO/IEC 20000, BS 25999 and ISO 22301, in Ref. [5].

Australia's cybersecurity standards compliance implementation is supported by the country's Cybersecurity Strategy 2009 highlighting the need for a consistent and integrated framework of policies, procedures and standards to protect its government's systems, as well as the other interconnected systems. Australia's security measures are: (1) the development and enforcement of the Protective Security Policy Framework (PSPF) to the government agencies through a Directive by the Attorney-General Department (AGD), (2) Australian government has enforced the ISO/IEC 15408 for procurement of products with security functions in the Government Sector, and (3) standards implemented voluntarily and adopted by critical infrastructure organizations are the American National Standard Institute/International Society of Automation (ANSI/ISA)-99, Industrial Automation and Control Systems Security and ISO27799 Health Informatics – Information security management in health using ISO/IEC 27002.

United States of America (USA) has developed various national strategies on cyber security: (1) The Comprehensive National Cybersecurity Initiative (CNCI) will evolve to become the key

element of a broader updated national USA cyber security strategy, (2) standards mandated to the Energy and Dams sectors are the Reliability Standards Critical Infrastructure Protection (CIP) 002-009 through the Code of Laws of the United States of America (U.S.C)Title 16 – Conservation, Section 824o – Electric Reliability (16 U.S.C 824o), (3) the standards mandated for the Government Sector is the Federal Information Processing Standards (FIPS) through Federal Information Security Management Act 2002 (FISMA), (4) other standards in the critical sectors are ISO 27799, ISO/IEC 27010 Information technology – Security techniques – Information security management for inter-sector and inter-organizational communications as well as ISO/IEC 27011 Information technology – Security techniques – Information security management guidelines for telecommunications organizations based on ISO/IEC 27002 for Telecommunications and ISO/ IECTR 27015 Information technology – Security techniques – Information security management guidelines for financial services for Financial Services, and (5) the government promotes the application of cyber security standards via establishment of national-level certification schemes for the following standards such as ISO/IEC 27001, ISO/IEC 27005, ISO/IEC 27006, ISO/IEC 20000, and ISO 22301, in reference [5].

The different approaches that countries take to cybersecurity standard compliance show that cybersecurity standards whether implemented mandatorily or voluntarily is a measure to enhance the protection of the critical infrastructure. Enforcing cybersecurity standards compliance may bring about a positive outcome to the overall cybersecurity management of a country, and not just the organizations implementing them, in reference [5].

1.5. Mexico rules and regulation

In Mexico, computer legislation is distributed in different rules and laws because the way in which it is pursued depends on the nature of the crime, then a list of some of the most common: show in **Table 1**.

1.6. E-government (eGov)

The year 2015 marked a milestone in efforts to eradicate poverty and promote prosperity for all people on a safe planet. With the adoption of the 2030 Agenda for Sustainable Development and other major international commitments. The 2030 Agenda is centered on a set of far-reaching and people-centered universal Sustainable Development Goals (SDGs). Reaching these goals in all countries and creating peaceful, just and inclusive societies will be extremely difficult in the absence of effective, accountable and inclusive institutions. Institutions need to be capable and equipped to adapt the Agenda to the national situation.

They need to be able to mobilize the society and the private sector in implementing the SDGs. Capacities and innovation will be required to promote policy integration, enhance public accountability, promote participation for more inclusive societies as well as ensure equitable and effective public services for all, particularly for the poorest and most vulnerable groups. Information and Communication Technology (ICT) and e-government are important tools to realize these objectives, in reference [8].

The 2030 Agenda itself recognized that "the spread of information and communications technology and global interconnectedness has great potential to accelerate human progress, to

1. E-commerce	Procurement Law, leases and public sector services (Ley de adquisiciones, arrendamientos y servicios del sector público)
	Commercial code (Código de comercio)
	Federal civil code (Código civil federal)
	Federal Law for Protection of the Consumer (Ley federal de protección al consumidor)
	Federal Law of Administrative Procedure (Ley Federal de Procedimiento Administrativo)
	Law of Investment Funds (Ley de Fondos de Inversión)
	Federation fiscal Code (Código Fiscal de la Federación)
	Social Security Law (Ley del Seguro Social)
2. Electronic signature	Advanced Electronic Signature Law (Ley de Firma Electrónica Avanzada)
	Law of the Tax Administration Service (Ley del Servicio de Administración Tributaria)
3. Personal data protection	Law of Investment Funds (Ley de Fondos de Inversión)
	Federal Law on Protection of Personal Data in Possession of Individuals (Ley Federal de Protección de Datos Personales en Posesión de los Particulares)
4. Right to the information	Article 6 Constitutional (Artículo 6° Constitucional)
	General Law of Transparency (Ley General de Transparencia)
	Access to public information (Acceso a la Información Pública)
	Federal Transparency Law (Ley Federal de Transparencia)
	Access to public information (Acceso a la Información Pública)
5. Violation of correspondence	Article 173 to 177 of the Federal Criminal Code (Artículo 173 al 177 del Código Penal Federal)
6. Revelation of secrets	Article 210 to 211 Bis of the Federal Penal Code (Artículo 210 al 211 Bis del Código Penal Federal)
7. Illicit access to computer systems and equipment	Articles from 211 bis 1 to 211 bis 7 of the Federal Criminal Code. (Artículos del 211 bis 1 al 211 bis 7 del Código Penal Federal).
8. Copyright	Articles from 424 to 429 of the Federal Penal Code (Artículos del 424 al 429 del Código Penal Federal).
9. Industrial property	Industrial property law (Ley de la propiedad industrial).
10. Stock market	Law of the marketc of values (Ley del mercado de valores).
11. Telecommunication	Federal Law on Telecommunications and Broadcasting (Ley Federal de Telecomunicaciones y Radiodifusión)
	Federal Criminal Code and Political Constitution of the United Mexican States (Código Penal Federal y Constitución Política de los Estados Unidos Mexicanos)

In Ref. [7].

Table 1. Mexico rules and regulation.

bridge the digital divide and to develop knowledge societies, as does scientific and techno-logical innovation across areas as diverse as medicine and energy", in reference [8].

The General Assembly has recognized on several occasions the role of information and com-munications technology in promoting sustainable development and supporting public policies and service delivery. It has underscored that ICT have enabled breakthroughs "in Government and the provision of public services, education, healthcare and employment, as well as in busi-ness, agriculture and science, with greater numbers of people having access to services and data that might previously been out of reach or unaffordable". The General Assembly has also specifically affirmed the "potential of e-government in promoting transparency, accountability, efficiency and citizen engagement in public service delivery", in reference [8].

There are several definitions in circulation which differ as to the meaning or scope of the term "eGov". Next definitions illustrate different scope or stress in the understanding of what eGov is: (A) United Nations: the employment of the Internet and the world-wide-web for delivering government information and services to the citizens, (B) The World Wide Web Consortium says that is the use of the Web and other information technologies by governments to interact with the citizenry, between departments and divisions, and with other governments. (C) The Organization for Economic Co-operation and Development focuses on the use of new infor-mation and communication technologies (ICTs) by governments as applied to the full range of government functions. In particular, the networking potential offered by the Internet and related technologies has the potential to transform the structures and operation of government, in reference [9].

From the business perspective, eGov is a way of introducing new channels of interaction between government and consumers of its services, in order to make this interaction more convenient to the consumers and cheaper for the provider of the services, these business ben-efits, including the following

- Facilitating access to government data and processes to all types of consumers of these services, be it general public, business, government agencies or their employees, or other governments.

- Improving operational characteristics of government, including: (1) decreasing the cost to government of providing quality services to their consumers, (2) decreasing the load on the office workers by making at least some types of data or some of business processes directly available to service consumers.

- The reach of eGov services can be widened from initialized specialized groups to all con-sumers that have a need and a right for using the service without incurring substantial additional costs, in reference [9].

1.6.1. International and national standards in information security

The International Organization for Standardization (ISO) has the ISO/IEC 27000 family of standards, which aims to help in the management of asset security such as financial informa-tion, intellectual property and confidential information of employees or third parties.

Both ISO (the International Organization for Standardization) and IEC (International Electro-technical Commission) make up the specialized system for global standardization, in Ref. [10].

The implementation of this system aims to preserve the confidentiality, integrity and availability of the information, through the application of a risk management process, which gives certainty to the interested parties that the risk has been adequately managed. They are considered best practices and are not considered mandatory as mentioned on their website, in Ref. [11].

In North America, there is NIST (National Institute of Standard Technology) of the United States Department of Commerce. This Institute handles issues of cybersecurity and privacy through the application of standards and best practices, including the ISO/IEC 27000 standards, in order to help organizations, manage the risk of cybersecurity within of a framework, in Ref. [12]. This framework includes physical, cybernetic and people security, applicable to organizations dependent on technology, such as industrial control systems, Information Technologies, cyber-physical systems, or connected devices, this includes the IoT (Internet of Things), in Ref. [13].

NIST has the National Cybersecurity Improvement Commission that is looking for:

- Awareness and protection at all levels of government, business and society to protect privacy,

- Guarantee public security and national economic security.

- Empower the Americans to have better control of their digital security.

Within NIST is the NC Coe (National Cybersecurity Center of Excellence), which are agreements with industry and institutions that participate in projects related to the topic of cybersecurity, in Ref. [14]. When information security is implemented, and the rules related to the country and form of government are respected, then its regulation must be followed to concretize the integration of the strategies involved. Such as applicable international norms and standards for their protection. Within the IT Audit is made aware of the use of controls for each aspect that involves the transfer of information, such as access control, entry control, communication control, etc.

A control is a procedure that verifies another procedure.

Therefore, control must be implemented that has not been violated, so its internal study is important.

The use of aspects such as high computing performance, infrastructure characteristics, is an important reason. Because there is no supercomputer, we opt for the design of a group of nodes that allows its gradual growth until it arrives as an operation to count on the performance of the supercomputers.

A network of computers has precisely the vision of the distributed and parallel system, however, care must be taken in the construction and operation thereof.

2. High performance in technology

There is an international body that measures the capacity in supercomputing performance, within its schemes contemplated for most of its platforms the use of Linux, within the architecture is appreciated the use of MPP and the cluster, where you can appreciate the acceptable capabilities both work and performance. With this, it is possible to detect the speeds that are presented internationally. (e.g. **Table 2**, **Figure 1**).

The TOP500 project ranks and details the 500 most powerful don't distribute computer systems in the world, in reference [15].

The following graphs are presented in terms of performance and shared systems on an international scale, information retrieved from top500.org, in Ref. [15] (e.g. **Table 3**, **Figure 2**).

Mexico only occupies 0.2% of System share, it is for this reason that it is necessary to learn from the superpowers because it seems that the power of the supercomputing is wasted. (e.g. **Table 4**, **Figure 3**).

In which of the registered systems it can be noticed that Asia occupies the first place in the use of shared systems in High performance in computing.

2.1. High performance in computation

One of the first programs to verify the performance of the processes and programs was the task manager because it allows us to measure the performance by means of metrics to support the resources of a computer system such as our computer.

HPC stands for High-performance computing

By computation is processing, which has been attributed to thinking carefully about the hardware or a physical infrastructure, however, it should be distinguished in that the computation is also given by logic. It continued with the **Table 5**. Performance metric.

The use of cluster implies simultaneously using concurrent technology, in the concurrent model contains the following components

- Structure (static/dynamic)
- Granularity (nested/plane)
- Initialization (Thick/Fine Ending)
- Representation (parent/child or guardian/dependent)

Architecture	System share (%)
Cluster	87.4
MPP	12.6

Table 2. Architecture.

System share (%)
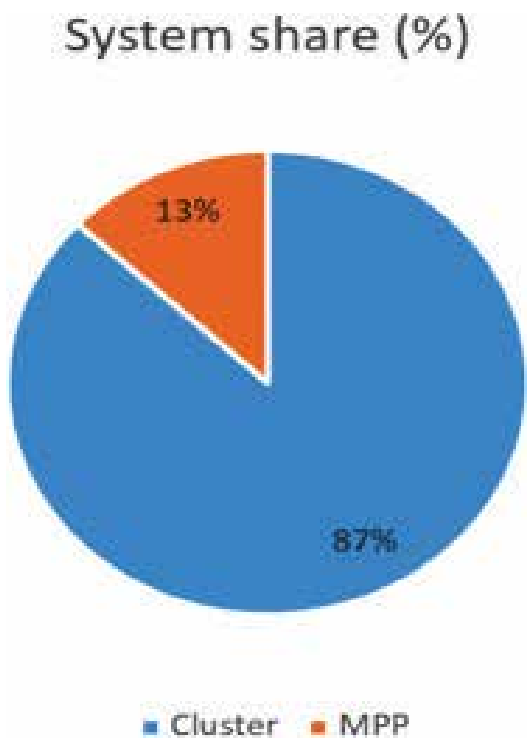
Figure 1. Architecture.

This systematic develop; consist in the use of the concurrent engineering by the design of components assembler. As show at the **Figure 4**.

By the moment these techniques have been studied at the final of Computer Sciences: the development of Systems in Real Time, the Recognition of Patterns, Semi-Intelligent Agents, the Computer Security as Cryptography and Crypto Analysis, the Information Systems based in knowledge, and Signs processing, are some cases for we appreciate this trouble, in reference [16].

Performance tests applied to cryptographic algorithms are abduction (statistics) and stress-based.

Continents	Performance
Other	0.70
Americas	30.20
Asia	48.70
Europe	20.40

Table 3. Continents.
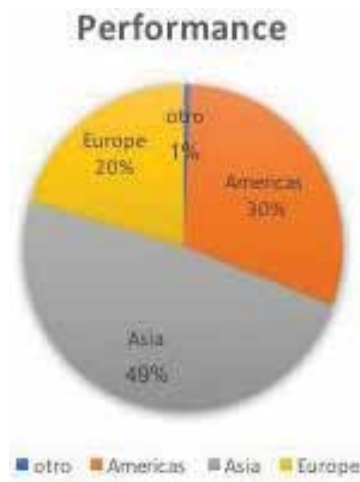
Performance

Figure 2. Performance.

Continents	System share (%)
Africa	0.2
Americas	29.8
Asia	50.4
Europe	18.6
Oceania	1

Table 4. Continents.
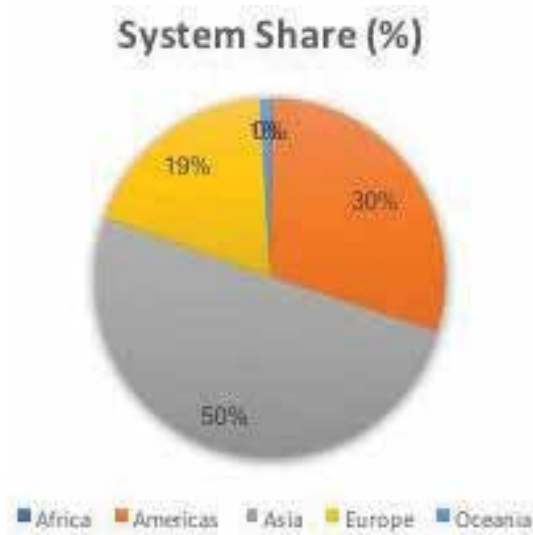
System Share (%)

Figure 3. System share.

Metric	Load balance
	Runtime
	Communication volume
	Computing time
Metrics for multidimensional optimization	Cluster vs. Supercomputer
	Fiedler vector or spectral clustering
	Minimum normalized cut
	Connectors to eigenvectors
	Clustering by k means
	Restrictions in two dimensions
	Fourier modal analysis

Table 5. Performance metric.
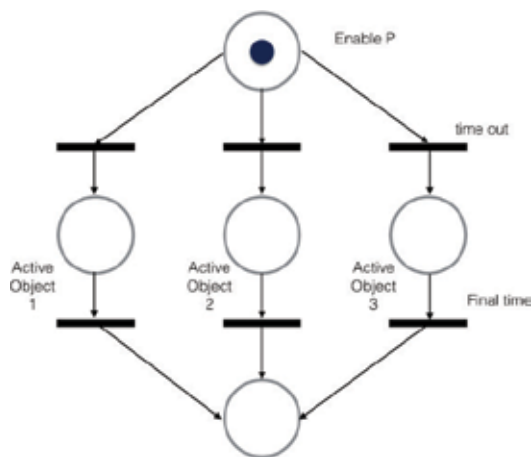
Figure 4. Petri net, model of concurrent active object.

3. Cryptographic algorithms

It can be implemented at any time, its operation involves calculating the key from the calculation of two keys that are presented with prime numbers.

The microchip company announced a microcontroller that has the ability to operate with most of the cryptographic algorithms, however, now it is worth thinking, what is it that is intended to ensure?

3.1. Cryptographic algorithms in IoT

Operating system being responsible for the management and control of all computer hardware resources, the security involves a multi-user and multi-programming environment where the

primary goals include the prevention of any unauthorized access to data stored within the system, the safeguarding of users from undesirable results triggered by their own actions, the assurance that various user programs will not interfere with one another, and the ability of different users to have different rights and abilities to cooperate with each other. The issues which deal with him are critical and far reaching, for this reason, they are of vital importance for both its own and its applications' sakes, in reference [17].

Achieving security in operating systems depends on the security goals one has. These goals will typically include goals related to confidentiality, integrity, and availability. In any given system, the more detailed particulars of these security goals vary, which implies that different systems will have different security policies intended to help them meet their specific security goals, in reference [18].

Network security threats are also an operating system responsibility since a computer network is usually a rather large, single system which has been created from numerous individual computer systems. The data processing functions must now be distributed among a set of distinct systems thus decentralizing the control of data storage and processing. In addition, information which must be transmitted between the various computers within the network is subject to exposure. Forged user identification and unauthorized access to stored data by legitimate users are also problems which plague a multi-user, multi-resource environment. Consequently, these factors combine to complicate the problem of ensuring a high degree of security within the network and may present formidable pitfalls, in Ref. [17].

A good cryptographic system must have several characteristics: Small variations of plain texts: large variations of encrypted texts, the sizes of the flat texts must be comparable with the ciphers, the ciphered texts must be calculated efficiently from the planes, and the relationship between plain and ciphered texts should be unpredictable. A bad cryptographic system is characterized because: It appears a random relationship between planes and ciphers, but in reality, it is not, it is susceptible to elementary cryptanalysis, the calculation of ciphers is inefficient in time and space, and is vulnerable to its own manufacturers, in Ref. [19].

In the past, protection against the monitoring of communication lines was guaranteed by the use of physically secure lines. This technique, however, proved to be extremely expensive and, often times, impractical. Since that time, it has been discovered that data encryption may be used as a viable alternative to secure lines, in Ref. [17]. Cryptographic technology is, therefore, a relatively inexpensive and highly effective process by which sensitive data may be protected against disclosure, in reference [17].

Android includes cryptography to protect their information, many of the applications that are installed in a cell phone with an operating system have already cryptographic features to protect the information that the cell phone itself has. Each of the applications that have a cell phone, both those that have the factory and those that are installed after having purchased the cell phone, are handled in layers, where each of them uses services offered by the previous ones and offers the same to layers superiors All applications use the services, application interfaces and libraries of the previous layers. This structure facilitates the development of cryptographic applications in this operating system, but there is the disadvantage that the execution of cryptographic algorithms becomes more expensive as their complexity and strength increase, in Ref. [20]. Linux has a kernel crypto API that offers a rich set of cryptographic ciphers as well

as other data transformation mechanisms and methods to invoke these. The kernel crypto API serves the following entity types: (a) consumers requesting cryptographic services and (b) data transformation implementations (typically ciphers) that can be called by consumers using the kernel crypto API, in Ref. [21].

Apple Inc. has a variety of cryptographic technologies around of all their products.

App Transport Security (ATS) ATS establishes best-practice policies for secure network communications using Apple platforms, employing Transport Layer Security (TLS) version.

3.2. Forward secrecy and strong cryptography

Secure Transport API. Use Apple's secure transport API to employ current versions of the Secure Sockets Layer (SSL), Transport Layer Security (TLS), and Datagram Transport Layer Security (DTLS) cryptographic protocols for network communications.

Supported Algorithms. With iOS 10 and macOS v10. 12, the RC4 cipher suite is now disabled by default. In addition, Apple recommends that you upgrade your servers to use certificates signed with the SHA-2 cryptographic function.

Cryptographic Signing. If distributing your Mac app outside of the Mac App Store, use cryptographic signing with Developer ID to certify that your app is genuine.

CryptoTokenKit for Smart Card Support. The CryptoTokenKit framework provides first-class access for working with smart cards and other cryptographic devices in macOS, in reference [22].

Microsoft shows in its web site the Cryptography API: Next Generation (CNG) that is the long-term replacement for the CryptoAPI. CNG is designed to be extensible at many levels and cryptography agnostic in behavior. Some of its features are listed next: Cryptographic Agility, Certification and Compliance, Suite B Support, Legacy Support, Kernel Mode Support, Auditing and Replaceable Random Number Generators, in reference [23].

According to the list dictated by top500.org, the best platforms in HPC are presented below (e.g. **Figure 5**).

3.3. Public key cryptography

The data that has been monitored by telepathic systems arises from information related to greenhouses, they are perhaps only data that support a part of a process in agronomy, however, they are important for their better interpretation.

JFLAP is software for experimenting with formal languages topics including nondeterministic finite automata, nondeterministic pushdown automata, multi-tape Turing machines, several types of grammars, parsing, and L-systems, by the University Duke, in Ref. [24] (e.g. **Figure 6**).
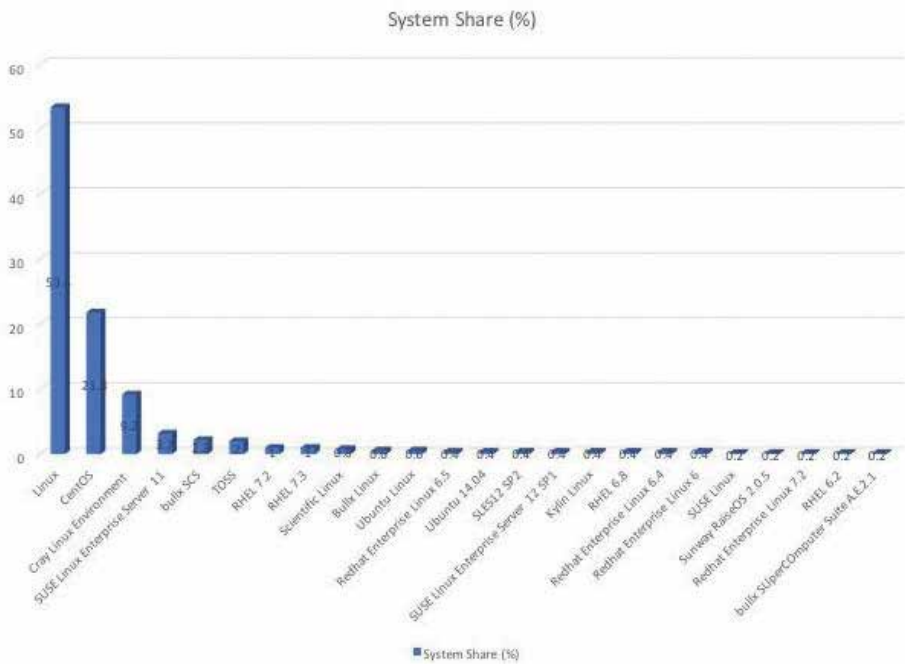
Figure 5. Histogram of platforms.

The type of applications that generally work with sophisticated equipment such as the supercomputer or cluster, is one that presents complexity in it is development, develop-ment, those that consume more massive storage capacity and require speeds appropriate to their way of being. While social networks currently involve more data consumption, it is important to recognize that so are the applications that generate scientific researchers by searching for new solutions to real-world problems or that are analogous to their area of expertise. The aim of the following scheme is precise to make known part of the advances that we as a research team have. We do not express all theories or investigations, it's just a
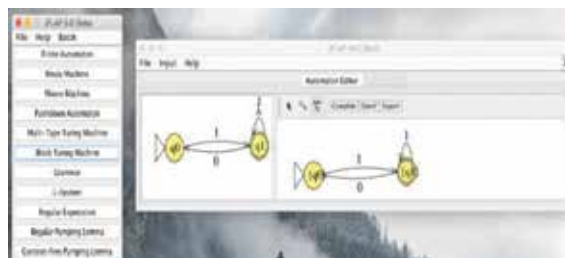
Figure 6. JFlap.

preamble of what these systems can be capable of doing in relation to what they are capable of doing. The use of state machines is important in the design of logical machines because they allow deducing the logic of their operation. That is to say, the following is the use of a simple machine that allows to deduce chains of 0 and 1, in such a way accept words of 0 and 1s, like prime numbers.

For the generation of keys, it is necessary to design an algorithm that allows the generation of random numbers, below is an algorithm that supports the generation of pseudorandom numbers, in reference [25].

Show in notation with Henon Map, this is a method chaotic map, and it is employed in dynamic systems to discrete process (e.g. **Figure 7**).

3.3.1. Encryption in one dimension (1D) flow

Used the RSA Algorithm to prove of text [26] (e.g. **Figure 8**).

3.3.2. Encryption in two dimension (2D) flow

The tent map, like various chaotic maps, as pseudorandom sequence generator, in Ref. [27] (e.g. **Figures 9** and **10**).
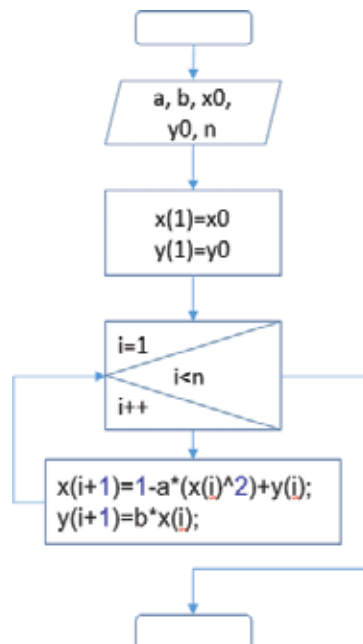
Figure 7. Pseudorandom by Henon map.
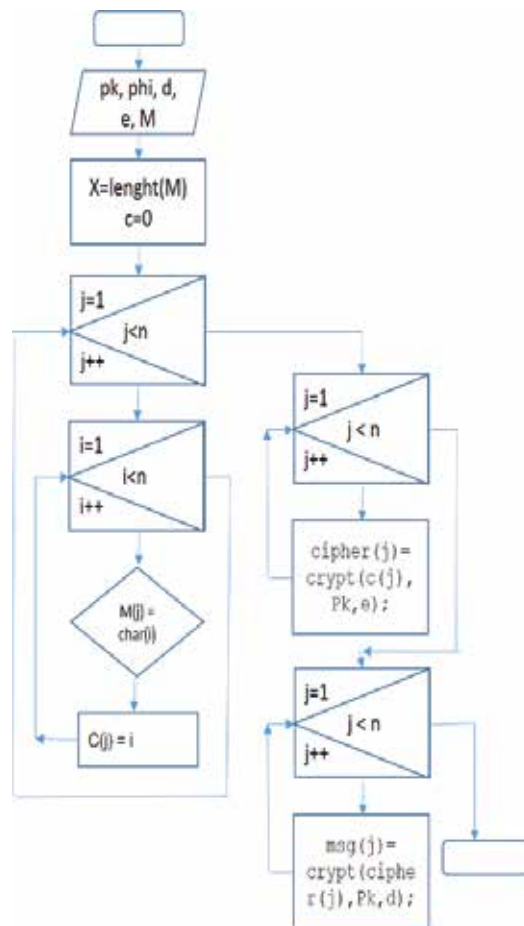
Figure 8. Algorithm by RSA.

These algorithms applied to images allow to know the range of operations necessary to work together, however, it is important to emphasize the moment in which the integration of the cryptographic algorithms must be perfected to the process involved.

3.3.3. Encryption in 3D flow

The reading of information generated by the stereo vision allow to get hard acquisition data, as the focal lens adjust. It proposes the tracking of systems at real time uses with the transformation Census. it is left for future work.

Real Time Stereo Vision with a modified Census transform and fast tracking in FPGA, in Ref. [28].

Figure 9. Algorithm by tent map.

```
clc; clear all;
a=0.9;
b=-0.6013;
c=2.0;
d=0.50;
x0=-0.72;
y0=-0.64;
n=input(max number of iteration = ');
x(1)=x0;
y(1)=y0;
fprintf('a= %i\nb= %i\n',a,b);
fprintf('c= %i\nd= %i\n',c,d);
fprintf('X0= %i\nY0= %i\n',x0,y0);
tic
for i=1:1:n
        x(i+1)=(x(i).^2)-
(y(i).^2)+(a*x(i))+(b*y(i));
        y(i+1)=2*(x(i)*y(i))+(c*x(i))+(d*y(i));
```

Figure 10. Algorithm by tent map.

4. Conclusions

4.1. Operative systems and cryptography

Security is a priority since the operative systems are responsible to manage a multi-user and multi-programing environment. Complexity around these tasks involve goals for preventing unauthorized access, safeguarding users from undesirable results by their own actions, managing rights and abilities to cooperate, they also implies confidentiality, integrity and availability. Another thing to take on account is the network security which is responsible to protect the information that is exposed when it is transmitted between various computers. A solution for these problems have been tackled through hardware solutions however these solutions have proved to be expensive and, often times, impractical. Data encryption is a viable alternative solution to secure the information because cryptographic technology is, a relatively inexpensive and highly effective process to protect sensitive data against disclosure. By default, popular operative systems like Android, Linux, Mac OS and Microsoft Windows have cryptography technologies for enhancing their security systems in the way to protect sensible data, even some of the share their APIs with the objective to give tools for reducing risks related with data security.
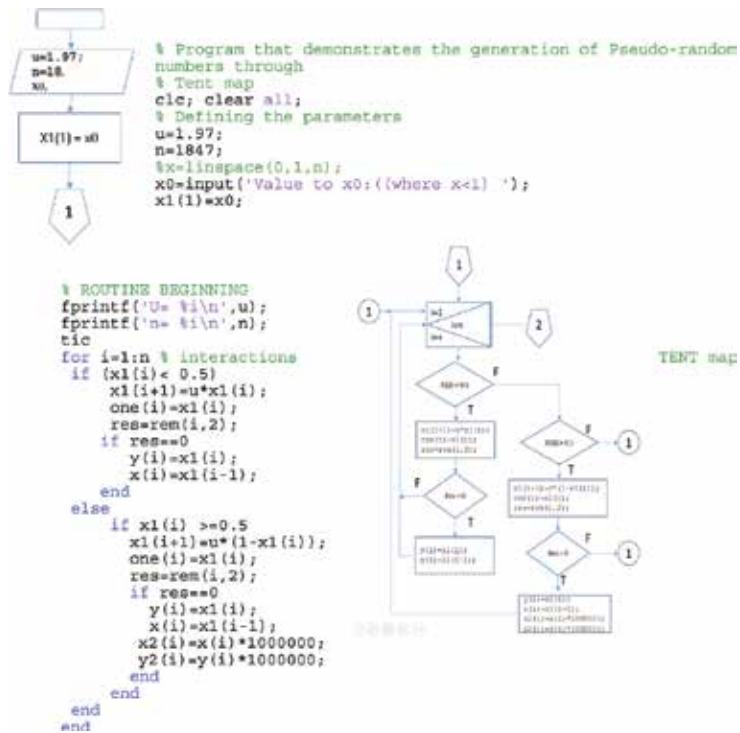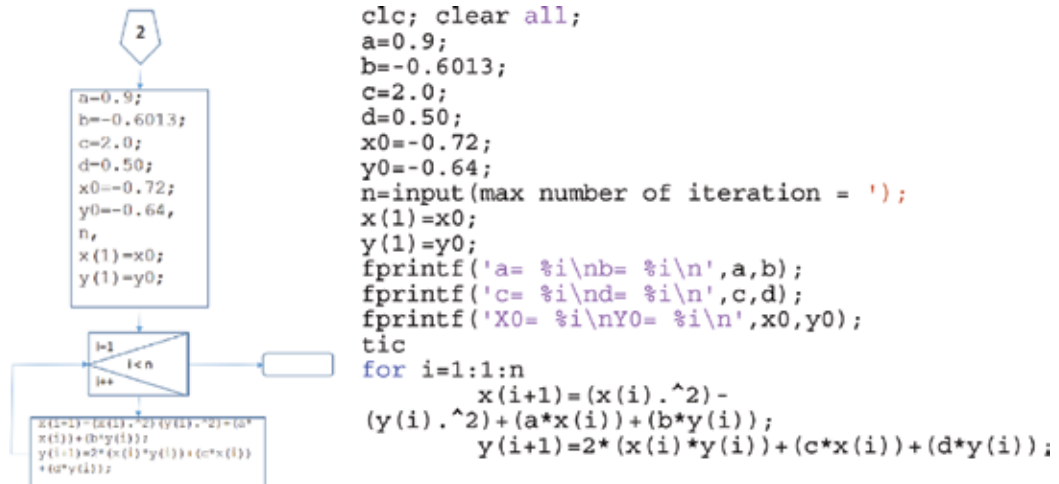
4.2. Operative systems

Countries around the world like United Kingdom, Australia and United States of America are making respectable efforts directed specifically in the cyber security infrastructures. Most of them adopted international security standards of which stand out ISO/IEC 27032:2012, ISO/IEC 27001, ISO 22301, ISO/IEC 15408, ISO27799, ISO/IEC 27010 and they also have its own intern standards and specialized organizations as is the case of United Kingdom with the Network Interoperability Consultative Committee (NICC) and the Information Assurance Maturity Model (IAMM).

In the case of Mexico, the government has not adopted international standards yet and the legislation is distributed in different laws and norms because the way in which it is pursued depends on the nature of the crime. Regulations consider crimes like: separated into different laws like electronic commerce, electronic signature, personal data protection, right to the information, revelation of secrets, industrial property among others. This weakens people, companies and the government itself when they are the victims of an attack which has as a target the security of their data.

It is necessary to prepare the relevant technology for the cryptographic algorithms that work with Linux, in the Cluster infrastructure, to make the most of the benefits related to a complete information processing. While it is necessary to have logical security, you only need to work on the integration of technologies in schemes at the level nodes of the net, as is the case with the Internet of things (IoT) thus ensuring all kinds of information such as multimedia can be operated.

Author details

Arturo Lezama-León[1]*, José Juan Zarate-Corona[1], Evangelina Lezama-León[2],
José Angel Montes-Olguín[3], Juan Ángel Rosales-Alba[3] and Ma. de la Luz Carrillo-González[3]

*Address all correspondence to: lezama@upp.edu.mx

1 Mathematics and Technology Sciences, Polytechnic University of Pachuca, Zempoala, Hidalgo, Mexico

2 Strategic Planning and Technology Management, Popular Autonomous University of the State of Puebla, Mexico

3 High Technological Institute Zacatecas Norte, Mexico

References

[1] Turing AM. Computing Machinery and Intelligence. Oxford Academic, Google Scholar Mind. 1 October 1950;**LIX**(236):433-460. DOI: 10.1093/mind/LIX.236.433. [Accessed: January 1, 2018]

[2] Russell S, Norvig P. Artificial Intelligence a Modern Approach. 2nd ed. Pearson Education Inc. Publishing as Prentice Hall; 2004. p. 1212. ISBN: 0-13-790395-2 [Accessed: January 2, 2018]

[3] Rich E, Knight K, Nair SB. Artificial Intelligence. 3rd ed. Mc Graw Hill. ISBN. 13: 978-0-07-0088770 [Accessed: January 3, 2018]

[4] Silva M. Las redes de Petri en la automática y la informática. AC; 1985. ISBN: 8472880451 [Accessed: January 4, 2018]

[5] KPMG International. Cyber Security Standards Compliance: A Vital Measure to Critical Infrastructure Protection. Printed in Malaysia. 2015 [Accessed: January 4, 2018]

[6] World Economic Forum. Insight Report Global Risks. 9th ed. 2014. ISBN 13: 92-95044-60-6. https://www.weforum.org/risks [Accessed: January 4, 2018]

[7] Torres & José. s.f. [Accessed: January 4, 2018]

[8] United Nations. 2016 [Accessed: January 4, 2018]

[9] California Department of Technology. 2014 [Accessed: January 4, 2018]

[10] https://www.iso.org/obp/ui/#iso:std:iso-iec:27001:ed-2:v1:en [Accessed: January 4, 2018]

[11] https://www.iso.org/isoiec-27001-information-security.html [Accessed: January 4, 2018]

[12] https://www.nist.gov/cyberframework [Accessed: January 4, 2018]

[13] https://www.nist.gov/sites/default/files/documents/2017/12/05/draft-2_framework-v1-1_without-markup.pdf [Accessed: January 5, 2018]

[14] https://nccoe.nist.gov/ [Accessed: January 6, 2018]

[15] top500.org, published on November 2017 [Accessed: January 7, 2018]

[16] Fúster A, Hernández L, Martín A, Montoya F, Muñoz J. Criptografía, protección de datos y aplicaciones, 2013. ALFAOMEGA. ISBN 978-607-707-469-4 [Accessed: January 7, 2018]

[17] Painchaud M. Cryptography and its application to operating system security [thesis]. Rochester Institute of Technology; 1981 [Accessed: January 8, 2018]

[18] Arpaci-Dusseau RH, Arpaci-Dusseau AC. Operating systems: Three easy pieces. 2015. [Accessed: January 8, 2018]

[19] Morales Luna G. Criptografía: Seguridad en la información. CINVESTAV-IPN. 2006 [Accessed: January 9, 2018]

[20] Núñez R. El uso de la criptografía en el sistema operativo Android. 2016 [Accessed: January 9, 2018]

[21] The Linux Kernel. (s.f.). The Linux Kernel. Recuperado el 05 de January de 2017, de The Linux Kernel: https://www.kernel.org/doc/html/v4.12/crypto/intro.html#introduction. [Accessed: January 10, 2018]

[22] Apple Inc. Security. Recuperado el 18 de January de 2018, de Apple developer: https://developer.apple.com/security/. 2018 [Accessed: January 11, 2018]

[23] Microsoft. CNG Features. Recuperado el 18 de January de 2018, de Microsoft: https://msdn.microsoft.com/es-es/library/windows/desktop/bb204775(v=vs.85).aspx.   2018 [Accessed: January 12, 2018]

[24] Jopcroft JE, Motwani R, Ullman JD. Introduction to Automata Theory; Languages and Computation. 3rd ed. Published by Pearson Education, Inc. Publishing as Addison-Wesley. 2007. ISBN: 978-84-7829-088-8 [Accessed: January 12, 2018]

[25] Lezama-león A, Liceaga-Ortíz-de-la-Peña JM, Zarate-Corona JJ. IoT Equipment Security. Advances in Digital Technologies. In: Mizera-Pietraszko J et al. editors. IOS Press, 2017 © 2017 The authors and IOS Press. All rights reserved. DOI: 10.3233/978-1-61499-773-3-126. Vol. 295. Publishing by Frontiers in Artificial Intelligence and Applications (FAIA). ISBN: 1879-8314 [Accessed: January 13, 2018]

[26] González RC, Woods R, Eddins SL. Digital Image Processing Using Matlab. Pearson Education, Inc, Publishing as Prentice Hall; 2004. ISBN: 0-13-008519-7 [Accessed: January 14, 2018]

[27] Liceaga-Ortiz-De-La-Peña JM, Lezama-León A, Zarate-Corona JJ, Hernández-Terrazas RO. Politechnique University of Pachuca. The tent map, like various chaotic maps, as pseudorandom sequence generator. SIMCI 2013. Simposio Ibero Americano Multidisciplinario de Ciencias e Ingenierías [Accessed: January 14, 2018]

[28] Pérez JMX, León AL, de la Peña JMLO, Hernández Terrazas RO. Real Time Stereo Vision with a modified Census transform in FPGA. In: Proceedings of the IEEE International Conference on Electronics, Robotics and Automotive Mechanics Conference (CERMA), 2012. DOI: 10.119/CERMA.2012.23 [Accessed: January 15, 2018]

A Deterministic Algorithm for Arabic Character Recognition Based on Letter Properties

Evon Abu-Taieh, Auhood Alfaries, Nabeel Zanoon,
Issam H. Al Hadid and Alia M. Abu-Tayeh

Additional information is available at the end of the chapter

Abstract

Handheld devices are flooding the market, and their use is becoming essential among people. Hence, the need for fast and accurate character recognition methods that ease the data entry process for users arises. There are many methods developed for handwriting character recognition especially for Latin-based languages. On the other hand, character recognition methods for Arabic language are lacking and rare. The Arabic language has many traits that differentiate it from other languages: first, the writing process is from right to left; second, the letter changes shape according to the position in the work; and third, the writing is cursive. Such traits compel to produce a special character recognition method that helps in producing applications for Arabic language. This research proposes a deterministic algorithm that recognizes Arabic alphabet letters. The algorithm is based on four categorizations of Arabic alphabet letters. Then, the research suggested a deterministic algorithm composed of 34 rules that can predict the character based on the use of all of categorizations as attributes assembled in a matrix for this purpose.

Keywords: conditional random field, rule-based mode, word prediction, virtual keyboard, Arabic text entry, enhancement, text entry system, theory of randomized search heuristics, structured prediction, theory of computation

1. Introduction

Arabic language is one of the top five languages spoken in the world. Arabic is used by more than 422 million native and non-native speakers in the world. Also, the letters of the Arabic alphabets are used in other languages like Urdu (65 million natives and 94 million

non-native) and Persian (110 million) languages. In addition, languages like Baluchi, Brahui, Pashto, Central Kurdish, Sindhi, Kashmiri, Punjabi, and Uyghur are using the Arabic letters. Hence, there is a need to develop an algorithm for character recognition for the Arabic language. Yet, there are major challenges that arise: first, Arabic is a cursive language. Unlike other languages written, Arabic alphabets change shape as written; hence, separate letters in Arabic are usually sub-word rather than stand-alone word. Second, Arabic is written from right to left unlike Latin languages. Third, Arabic has 28 alphabets, with some letters changing shapes based on the location of the letter in the word. Also, some letters are very similar in form yet have secondary marks to differentiate. Furthermore, Arabic is written from right to left cursively. Due to all the mentioned reasons, Arabic character recognition systems are under developed and lacking.

This research is composed of five sections. The first section presents 20 related works. Then, the research explains the letter shapes in Arabic language and the four categories used in the proposed algorithm. The four categorization methods will be employed to develop a deterministic algorithm method of categorization. The first categorization method depends on the number of dots used with each letter. The second categorization method depends on the shape of the letter, with classification to the letters. The third categorization is presented with the shape of the letter as used in the beginning, middle, and end of the word. The fourth categorization method relays on the proportion method, which is a method used in Arabic calligraphy that is based on rhombic dot. Then, the research suggested a deterministic algorithm composed of 34 rules that can predict the character based on the use of all of categorizations as attributes assembled in a matrix for this purpose.

2. Related work

Character recognition is an open-ended problem. A computer cannot recognize character or language alphabets. There is a great progress in character recognition that all can be seen in the different smart application used on smart phones and pad as well as notebooks and PCs. Problems that arise with non-Latin languages are well known, and many researchers have conducted research for their respective languages: Hindi language [1]; Chinese language [2–4]; and Arabic Language [5, 6]. Furthermore, many researchers have conducted research for the Arabic alphabets: Parvez and Mahmoud [7] conducted a survey for text recognition and published their work in a paper titled *Offline Arabic Handwritten Text Recognition: A Survey*. Three researchers [8] conducted a research and published their work in a paper titled *Robust Named Entity Detection Using an Arabic Offline Handwriting Recognition System*; the paper focus was "on extraction of a predefined set of Arabic named entities (NEs) in Arabic handwritten text." Another research by Fouad Slimane, Slim Kanoun, Adel M. Alimi, Jean Hennebert, and Rolf Ingold was conducted in 2010, yet the concentration was on printed character rather than handwritten one. The research conducted by Fard, Moghadam, Bidgoli, and Hussain [9] was very promising and was conducted on Persian language, yet the method used was neural

network based which is not deterministic. Another research conducted by Abu-Taieh [10] used an enhanced method of neural network. Another promising research that was studied by Aljarrah et al. [11] is the study concentrated on printed Arabic rather than hand written in order to produce Arabic optical character recognition system. The need for Arabic character recognition is evident according to Ali and Sagheer [12] which addresses the need of smart mobile phones and tablets and hence the need for Arabic character recognition.

Researchers Supriana and Nasution [13] cited nine works including their own research that all are non-deterministic. The research of Sarfraz, Ahmed, and Ghazi [14] developed a license plate recognition system. The research by Izakian, Monadjemi, Ladani, and Zamanifar [15] used chain codes, while Abandah, Khedher, and Mohammed [16] used selected feature extraction techniques. In their research Al-Taani and Al-Haj [17] used structural features, while Kapogiannopoulos and Kalouptsidis [18] used skew angle. The research of Zidouri [19] proposed a general method for Arabic letter segmentation, while Amin [20] used global features and decision tree technique on printed letters not handwritten. Cowell and Hussain [21] used extracting features.

3. Letters in Arabic language

To develop the proposed algorithm, the researchers studied and presented the different categorizations for Arabic letters. Next, each categorization will be explained accordingly. The first categorization method depends on the number of dots used with each letter. The second categorization method depends on the shape of the letter, with classification to the letters. The third categorization is presented with the shape of the letter as used in the beginning, middle, and end of the word. The fourth categorization method relays on the proportion method, which is a method used in Arabic calligraphy that is based on rhombic dot. Each categorization will be explained in Sections 3.1, 3.2, 3.3, and 3.4.

3.1. First categorization: number of dots in the letter

The use of dots to distinguish letters in Latin-based languages is familiar to people. In English, small letters I and J are distinguished by using a dot on top of the letter. In Arabic the use of a dot is used extensively; in fact, only 12 letters out of 28 letters are not doted. Furthermore, some letters use one, two, and three dots. Next, the concept of doted letters will be explained.

The first categorization is according to the number of dots used with each letter. This categorization splits the 28 letters (**Table 1**) into five branches, and from within it breeds two extra letters. The first branch is composed of 12 letters that has no dots whatsoever. The second branch is composed of ten letters: the eight letters have their dot above the body of the letter, and the other two letters have their dot below the letter body. The third branch is composed of four letters: the three letters have their two dots above the body of the letter, and the other one has two dots below its body. The fourth branch has two letters with three dots above the body.

First branch	No dots	12	ا و ه م ل ع ط ص س ر د ح
Second branch	One dot	10	ن ف غ ظ ض ذ خ ج ب
Third branch	Two dots	3	ة ق ت ي
Fourth branch	Three dots	2	ش ث
Fifth branch	With hamza	3	ك أ و

Table 1. Arabic alphabets (according to dots).

The fifth branch deals with hamza: there is one basic letter where the hamza is part of the letter "ك," and the other hamza is not part of the letter like the "أ" and "وْ." The categorization is summarized in **Table 1**.

3.2. Second categorization: letter shape

The second categorization is according to shape of the letter: this categorization splits the 28 letters into 15 branches based on the body of the letter rather than the dots on the letter (see **Table 2**). However, some increase the number of shapes to 18 shapes [22]. The first branch is made of four letters all very similar in shape: two of them are differentiated by one dot (one above the letter and below the letter), and the other two (one has two dots above it and one has three dots above it). The second branch has two letters very similar to each other: one can differentiate between them by the dot above one, while the other one has no dot; furthermore, the third branch and the fourth branch have the same idea similar in shape, yet one dot makes a difference. The same happens with the fifth and sixth branches. The seventh branch has three letters that are similar in shape: one without dot, one with dot above it, and one with dot below it. The eighth branch has two letters that are very similar in shape: one with one dot and the other with two dots. The ninth branch has two letters similar in shape: one with no dots and the other with three dots. The tenth has two letters: one with no dots and other with two dots. The 11th branch has two letters: one with no hamza and the other with hamza shape above the body of the letter. The 12th, 13th, 14, and 15th branches are not similar to each other nor to the rest of the letters.

One may add here a note about the shape of the letters; there are nine letters that have as part of them enclosed space that resembles a circle. These nine letters are (ض ص ظ ط ق ف ه و م). The enclosed circle property is an important aspect of the nine letters that will be used in the algorithm at a later stage.

3.3. Third categorization: letter location in a word

The third categorization of the Arabic alphabets is based on the location of the letter in a word. Generally, shapes of the Arabic alphabets change according to position of the letter in the word itself (beginning, middle, end); some letter can be connected (refers to the letter succeeding or preceding), and others cannot be connected. The shapes of the letters can be generated with ligature or character overlaps [23, 24]. When discussing the letters that start a word, these six letters when falling at the beginning of a word must stand alone; those letters are (و ز ر ذ د ا),

Branch	Count	Shape of letter	Proportion
1	4	ن ث ت ب	
2	2	ض ص	
3	2	ظ ط	
4	2	غ ع	
5	2	ذ د	
6	2	ز ر	
7	3	ج خ ح	
8	2	ق ف	
9	2	ش س	
10	2	ة ه	
11	2	ك ل	
12	1	م	
13	1	ا	
14	1	ي	
15	1	و	

Table 2. Arabic alphabets (according to shape).

and the rest of the letters do change form as seen in **Table 3**. Using the same six letters in the middle or end of the word, these letters are only connected and they do not change form. All letters when used at the end of the word have two states: connected and stand-alone.

From the previous one can notice that six letters have special characteristics, namely, (ا, و, د, ذ, ر, ز). These characters when used in the beginning of a word must stand alone, and also when they end a group of characters, they must be followed by independent character. Hence, they only connect to the predecessor not the successor.

Letter	Beginning	Middle	End
ا	ا	‍ا‍	‍ا
ب	بـ	ـبـ	ب بـ
ت	تـ	ـتـ	ت تـ
ث	ثـ	ـثـ	ث ثـ
ج	جـ	ـجـ	ج جـ
ح	حـ	ـحـ	ح حـ
خ	خـ	ـخـ	خ خـ
د	د	ـد	د د
ذ	ذ	ـذ	ذ ذ
ر	ر	ـر	ر ر
ز	ز	ـز	ز ز
س	سـ	ـسـ	س سـ
ش	شـ	ـشـ	ش شـ
ص	صـ	ـصـ	ص صـ
ض	ضـ	ـضـ	ض ضـ
ط	طـ	ـطـ	ط طـ
ظ	ظـ	ـظـ	ظ ظـ
ع	عـ	ـعـ	ع عـ
غ	غـ	ـغـ	غ غـ
ف	فـ	ـفـ	ف فـ
ق	قـ	ـقـ	ق قـ
ك	كـ	ـكـ	ك كـ
ل	لـ	ـلـ	ل لـ
م	مـ	ـمـ	م مـ
ن	نـ	ـنـ	ن نـ
ه	هـ	ـهـ	ه هـ
و	و	ـو و	ـو و
ي	يـ	ـيـ	ي يـ

Table 3. Arabic alphabets: stand alone, beginning, middle, and end of a word.

The matrix, seen in **Table 4**, represents the different combination between all 28 letters. The first column in the matrix is the letter coming at the beginning of the order, and the first row is all the letters coming second. Each cell in the matrix shows the two letter shapes and how they change as the order differs. The highlighted letters are the previously mentioned six letters, namely, (ز، ر، ذ، د، و،ا), which if appears at the beginning of the word, then they stand alone. When these letters appear consecutively within a word, they will both be written as stand-alone independent letters.

3.4. Fourth categorization: letter proportion

To keep letters proportional to each other, two ways were used by calligraphers: rhombic dot and circles. Arabic calligraphy was used in mosques and castells as decoration since Islam

forbids pictures and statues [22]. Hence, there is a need to decorate with words. Proportion is an essential part of the written word. The circle proportion was suggested by "Ibn Mugla," a well-known calligrapher from the eleventh century [25]. Three elements are the bases of proportion in Arabic calligraphy [26, 27]:

- The height of the *alif*, which is a straight and vertical stroke (3–12) rhombic dots.

- The width of the alif, (the rhombic dot) which is the square impression formed by pressing the tip of the calligrapher's reed pen to paper (see **Figures 1** and **3**).

- An imaginary circle with alif as its diameter, within which all Arabic letters could fit and be written (see **Figure 2**).

The circle is halved vertically and horizontally, with diameter equals the height of the first letter in Arabic alphabets called *alif*. Looking back at **Table 2** and **Figure 4** that represent the shape and form of the letter, the first branch, the shape of the letter, is in the lower half of the circle. The second branch, according to the circle, takes up the first quarter and the third

Table 4. Matrix of the different combinations for all 28 letters.

Figure 1. Example of measuring the letter by using rhomboid dots [28].
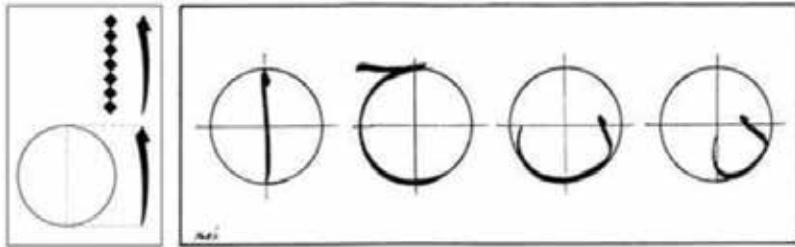
Figure 2. Example of measuring the letter by using circle [28].

Figure 3. The rhombic dot as a guide to proportions [22].

quarter. In the third branch, the letter falls in the first quarter of the circle with the upper half diameter aligned with the half alif of the letter. In the fourth branch, the letter lies on the left half of the circle. In the fifth branch, two letters are located in the fourth quarter of the circle. In the sixth branch, two letters also fall in the fourth quarter of the circle. In the seventh branch, the letter lies on the left half of the circle. In the eighth branch, two letters are both parted in the first and second quarter of the circle with a circular part above the horizontal diameter. In the tenth branch, the letter takes the first quarter of the circle. The eleventh branch takes the second and third quarter of the circle. In the twelfth branch, the letter is at the center of the circle and uses the bottom half the alif. The thirteen branch is the alif itself, which is the diameter of the circle. The fourteenth branch is taking the third quarter of the circle. The fifteenth branch takes the first and fourth quarter of the circle.
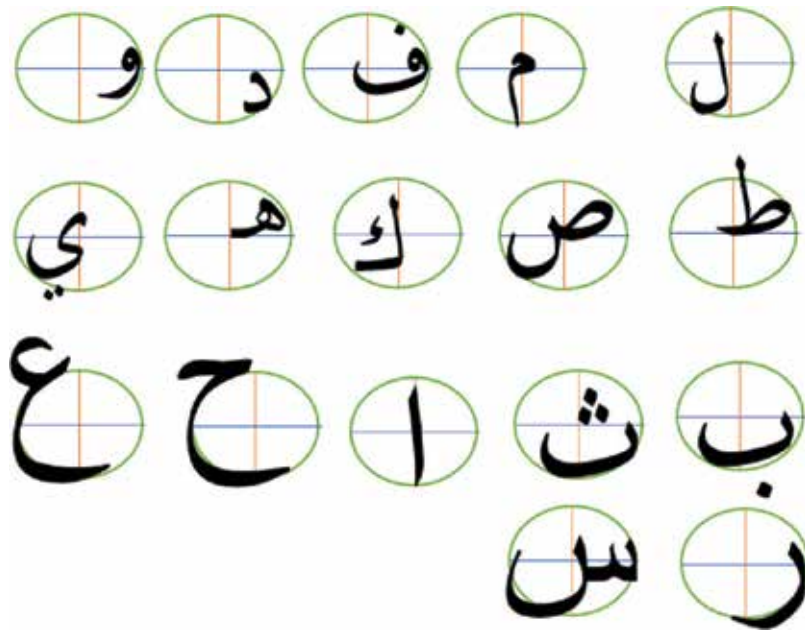
Figure 4. Proportions in Arabic calligraphy.

One can conclude by studying the second categorization and the proportion categorization through the following:

• First, the second branch and ninth branch both (four letters) take same area of the circle.

• Second, the third branch and tenth branch (four letters) both use the first quarter of the circle.

• Third, the fourth branch and seventh branch use the left edge of the circle, yet the differentiation between the two is that one letter is written from right to left and one letter is written from left to right as seen in **Figure 5**.

• Fourth, the fifth and sixth branches (four letters) use the fourth quarter of the circle.
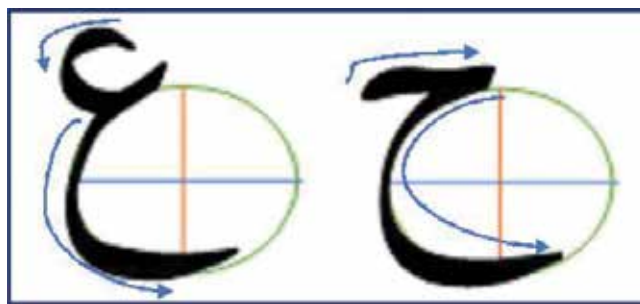
Figure 5. Direction of writing with two circle edge letters.

Hence, give an insight to further classify the letters and manage them into groups. The previous sections explained in details the four categorizations used in the proposed algorithm. Each categorization was an essential in the building blocks and rules of the algorithm.

3.5. Findings of the four categorizations

Based on the four categorizations explained above, a tree of rules can be built as seen in **Figure 6**. The rule tree has five branches: the first branch is for Arabic alphabets that contain *no dots*. The second branch is for letters with *one dot*. The third branch is for letters with *two dots*. The fourth branch includes all letters with *three dots*. The fifth branch is for letters with *hamza*.

For the first branch including 12 letters and in order to distinguish among the letters, the fourth categorization logic was used. Each letter in this branch was located in the quarters of the circle suggested in the fourth categorization. Two letters used the same quarters (س ص); both fall in the first and third quarters of the imaginary circle, which explain the fourth categorization. Still, letter (ص) has an enclosed space, while letter (س) has no enclosed space. Hence, differentiating between the two letters depends on the enclosed space. The enclosed space property is explained previously in the third categorization. The edge of the circle from the fourth categorization was used to differentiate between the ten letters and the letters (ع ح). Furthermore, to differentiate between the two letters, the direction of writing was used. The direction of writing was explained in **Figure 5** previously.

The second branch consisting of all letters with *one dot* included ten letters. The branch spliced further to *dot below* and *dot above* the body of the letter. Again, in this branch the imaginary circle from the fourth category was used. The location of the letters according to the quarters of the imaginary circle was used as seen in **Figure 6**. Also, the distinguishing feature of the letter falling on the edge of the imaginary circle is used, and the property of writing direction seen in **Figure 5** is also used.

The third branch consisting of all letters with *two dots* included three letters. The branch spliced further to *dots below* and *dots above* the body of the letter. There is only one letter in all the alphabets that has two dots below it (ي). And, there are three letters with two dots above the letter body. To distinguish between the three letters, the imaginary circle from the fourth categorization and again none shared the same quarters of the circle.

The fourth branch included all letters with *three dots*. The branch included only two letters both have the dots above their body. Hence, the quarters of the imaginary circle were used to distinguish between them.

The fifth branch is the *hamza* (ء) branch which included three letters, and the distinguishing features were the imaginary circle quarters: letter (ك) is in the second and third quarters of the circle, letter (ﺅ) is in the first and fourth quarters, and letter (أ) falls on the diameter of the circle.

The *hamza* (ء) can be seen on top of the letters (أ, ﺅ, ك); the hamza sometimes is considered an independent letter when used in some words like (عالء) and is used as part of the letter in other words. The hamza is a distinguishing character between the two letters (ك, ل). Hence, it is treated as semi-letter and is not listed in the alphabets.
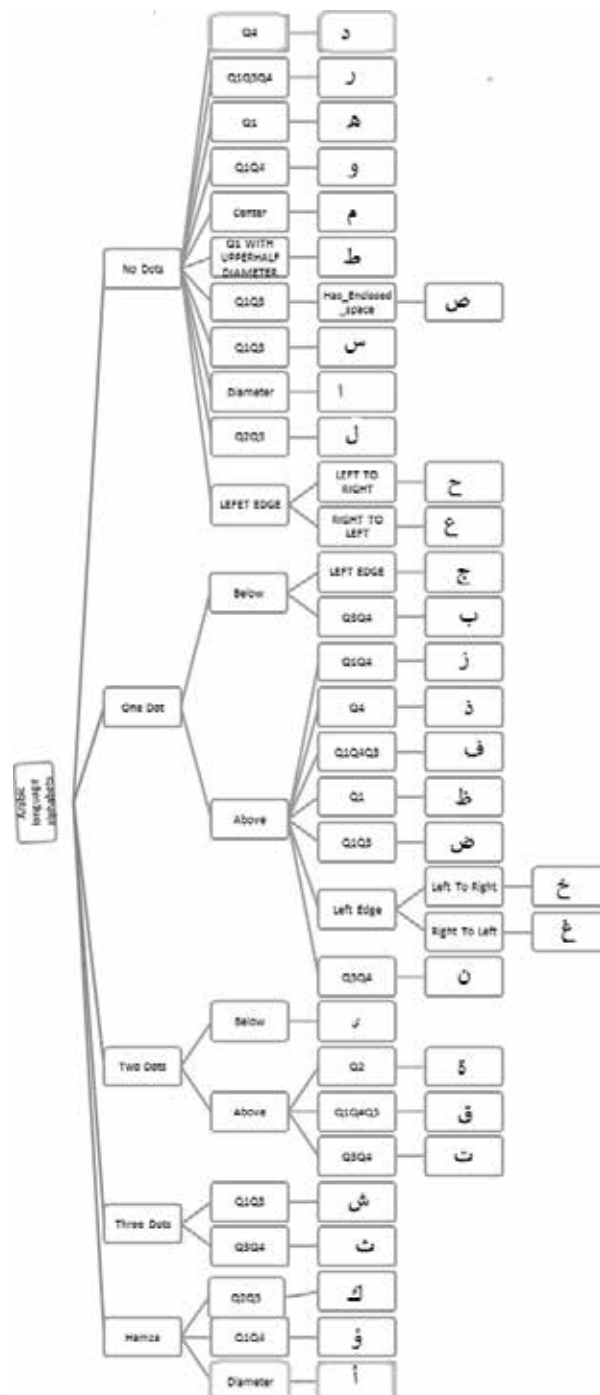
Figure 6. Categorization of the tree drawn based on the four categorizations.

4. Suggested algorithm

After studying all the previously mentioned categorizations, one can reach the conclusion that a deterministic algorithm can predict the character being drawn based on the following matrix in **Figure 7** and along with the matrix is the suggested algorithm in **Figure 8**, hence reducing the determination of a letter to 38 rules.

The suggested algorithm shown in **Figure 7** is composed of five major if-then statements which are based on the first categorization explained above and later summarized in **Figure 7**. The first if-then statement runs from line 1 to 12 in **Figure 8**. The if-then statement really deals with all cases of the letters which have no dots, and their location in the circle is mentioned in proportion categorization. The enclosed space property mentioned earlier was very important to distinguish letter "س" and letter "ص"; both letters fall in the same location in the circle Q1 and Q2, yet the latter has an enclosed space. Also, notice that both "ح" and "ع" have the same properties, yet to differentiate them, the direction of writing is used [9].

The second major if-then statement started at line 13 and dealt with letters with one dot. As shown in **Figure 6**, the dot can be either above or below the body of the letter, i.e., both letters "ن" and "ب" fall in the lower part of the circle quarters 3 and 4. Yet, to make the differentiation, the dot was essential here, the latter had the dot below as seen in lines 23 and 20 in **Figure 8**. Also, line 21, in the same figure, dealt with two letters that are essential falling in the same location, and both had the dot above them, distinguished by the writing direction left to right or right to left. The algorithm can be improved by eliminating line 24, hence reducing the number of rules to 37, since one can use one statement.

The third major if-then statement starts at line 25 and ends at line 31 in **Figure 8**. The if-then statement deals with letters that have two dots according to the first categorization and the matrix seen in **Figure 7**. The nested if-statement deals with the two dots whether above or below the letter. Three letters have two dots above them, yet their location on the circle is very distinguishable; hence, using the attribute "enclosed space" was not necessary. Furthermore, one can eliminate line 31 since this is the only letter in the alphabet that has two dots below it. Still, for the purpose of clarity, line 31 was left in the suggested algorithm. If line 31 was eliminated, the number of rules will be again reduced to 36 rules.

The fourth major if-then statement deals with letters that have three dots; there are only two of them. Both letters can be distinguished based on their respective location according to the proportion categorization. Again, line 34 can be eliminated but was left for the purpose of clarity. If line 34 was eliminated, the number of rules will be again reduced to 35 rules.

The last major if-then statement starts at line 35; the statement deals with case of "hamza." The hamza is an essential part in letter "ك" and is used with other letters like "أ ؤ." The three letters are distinguished by their location within the circle according to proportion categorization. Line 38 can be eliminated but was left to clarify the algorithm, hence reducing the number of rules to 34.

		Dot position	Circle quarters	Pen movement	Has enclosed space		Results Letter
First branch	No dots					12	
	No dots		Q4				د
	No dots		Q1 Q4 Q3				ر
	No dots		Q1		x		ه
	No dots		Q1Q4		x		و
	No dots		center		x		م
	No dots		Q1 with upper half diameter		x		ط
	No dots		Q1, Q3		x		ص
	No dots		Q1, Q3				س
	No dots		diameter				ا
	No dots		Q2 Q3				ل
	No dots		Left edge	Left to right			ع
	No dots		Left edge	Right to left			ح
Second branch	One dot	above				10	
	One dot	above	Q4 & Q1				ز
	One dot	above	Q4				ذ
	One dot	above	Q1, Q4, Q3		x		ف
	One dot	above	Q1		x		ظ
	One dot	above	Q1, Q3		x		ض
	One dot	above	Left edge	Left to right			خ
	One dot	above	Left edge	Right to left			غ
	One dot	above	Q3 Q4				ن
	One dot	below	Left edge				ج
	One dot	below	Q3 Q4				ب
Third branch	Two dots					3	
	Two dots	Above	Q2		x		ة
	Two dots	Above	Q1, Q4, Q3		x		ق
	Two dots	Above	Q3 Q4				ت
	Two dots	below	Q3				ي
Fourth branch	Three dots		Q1, Q3			2	ش
			Q3 Q4				ث
Fifth branch	With hamza					3	
			Q2 Q3				ك
			Q1Q4		x		ؤ
			diameter				أ

Figure 7. The property rules to define each letter in the Arabic alphabets.

```
1          If Number_of_Dots=0
2                  If circle_location=Q4  then "ﺩ"
3                  If circle_location= Q1&Q3  then "ﺏ"
4                  If circle_location= Q2&Q3  then "ﻝ"
5                  If circle_location=Q1&Q3&Q4  then "ﺭ "
6                  If circle_location=Q1 & Has_Enclosed_space=true then "ﻩ"
7                  If circle_location= Q1&Q4 & Has_Enclosed_space=true then "ﻭ"
8                  If circle_location= Center & Has_Enclosed_space=true then "ﻩ"
9                  If circle_location=Q1 & with upper half diameter then "ﻁ"
10                 If circle_location=Q1&Q3 & Has_Enclosed_space=true then "ﺹ"
11                 If circle_location=diameter then "ﺍ"
12                 If circle_location= Left_edge then if writing_Direction= left_to_right then "ﺡ"
13         else "ﻉ"
14         Else If Number of dots=1
15                 IF dot_location=Above_letter then
16                         If circle_location= Q1&Q4 then "ﺯ"
17                         else If circle_location= Q4 then "ﺫ"
18                         else If circle_location= Q1&Q3&Q4 then "ﺕ"
19                         else If circle_location= Q1 then "ﻅ"
20                         else If circle_location= Q1&Q3 then "ﺽ"
21                         else If circle_location= Q3&Q4 then "ﻥ"
22                         else If circle_location= Left_edge then if writing_Direction=
23         left_to_right then "ﺥ" else "ﻍ"
24                 Elseif dot_location =Below_letter
25                         If circle_location= Q3&Q4 then "ﺏ"
26                         else If circle_location= Left_edge then "ﺝ"
27         Else If Number_of_dots=2
28                 IF dot_location=Above_letter then
29                         If circle_location= Q2 then "ﻱ"
30                         else If circle_location= Q1&Q3&Q4 then "ﻕ"
31                         else If circle_location= Q3&Q4 then "ﺕ"
32                 Elseif dot_location =Below_letter
33                         If circle_location= Q3 then "ﻱ"
34         Else If Number_of_Dots=3
35                 If circle_location= Q1&Q3 then "ﺵ"
36                 else If circle_location= Q3&Q4 then "ﺙ"
37         Else (Hamza case)
38                 If circle_location= Q2&Q3 then "ﺀ"
                   else If circle_location= Q1&Q4 then "ﺅ"
                   else If circle_location= diameter then "ﺇ"
```

Figure 8. Determine_Character (input:one_character).

5. Conclusion

The proposed algorithm stems from many needs that are more apparent today. First, there is a rise in the use of handheld devices, which use character recognition methods that serves mainly Latin-based languages. Arabic language is one of the top five languages spoken in the world. Arabic is used by more than 422 million native and non-native speakers in the

world. Arabic language is different from other languages: Arabic is a cursive language, written from right to left, and letters change shape according to the position of the word. Hence, there is a dire need to develop an algorithm for character recognition for the Arabic language. However, many algorithms have used artificial intelligent methods to recognize characters that make their algorithms non-deterministic, while the proposed algorithm is deterministic. This research presents four categorization methods that will be employed to develop a deterministic algorithm method of categorization. The first categorization method depends on the number of dots used with each letter. The second categorization method depends on the shape of the letter, with classification to the letters. The third categorization is presented with the shape of the letter as used in the beginning, middle, and end of the word. The fourth categorization method relays on the proportion method, which is a method used in Arabic calligraphy that is based on rhombic dot. Then, the research suggested a deterministic algorithm composed of 34 rules that can predict the character based on the use of all of categorizations as attributes assembled in a matrix for this purpose [29].

The proposed algorithm is only one piece in the whole puzzle. There are many parts that need to be developed. One major part is the input section of the algorithm. Such part needs to exist in order for the puzzle to be complete. The input section needs to parse the word into segments that can detect the shape of the letters, the dots, and the hamza. Furthermore, this research will be a building block for further research and development.

Biography

Evon M. O. Abu-Taieh, PhD, is an associate professor and an author/editor of four scholar books, contributed in more than eight scholar books. She has more than 40 published papers. She was previously the acting dean in the University of Jordan (Aqaba) for 3 years. Dr. Evon is an editorial board member in five renounced journals. She has more than 29 years of experience in education, computers, aviation, transport, AI, ciphering, routing algorithms, compression algorithms, multimedia, and simulation.

Auhood Abdullah Alfaries, PhD, is as assistant professor in the IT Department in King Saud University (KSU). Dr. Auhood received her PhD degree in Semantic Web and Web Services from the School of Computing and Information Systems, Brunel University, UK. She held a number of IT-related academic and administrative positions both in KSU and princess Noura bint Abdulrahman University (PNU). She has experience in quality and program accreditation by serving in a number of quality-related roles since 2011. Auhood is associated with a number of important bodies such as an associate of the UK Higher Education Academy, a member of the Institute of Electrical and Electronics Engineers (IEEE), and a member of the Saudi Computer Society. She is also an ABET program evaluator. She participated as a conference and journal reviewer and a member of a number of national and international workshops and conference program committees. She has served as the vice dean and dean of E-Learning and Distance Learning Deanship in KSU and then in PNU for 2 years and has also served as the assistant general director and then the director for the General Directorate of Information and Communications Technology (ITC) in PNU. Currently, she is the dean of the

College of Computer and Information Sciences. Auhood's research interest includes semantic web, ontology engineering, natural language processing, machine learning, and cloud computing. She is a member of IWAN Research Group.

Dr. Nabeel Mohammed Zanoon received his PhD degree in Computer Systems Engineering, from the South-West State University, Kursk, Russia, in 2011. He is a faculty member with Al-Balqa' Applied University since 2011, where he is currently an assistant professor and the head of the Department of Applied Sciences as well as the director of the ICDL Computer Centre and Cisco Academy Branch of Aqaba University College. He has published several researches in several areas: security of e-banking, algorithm scheduling in grid and cloud, meta-grammar, hardware and architecture, fiber optical, and mobile ad hoc networks.

Issam Hamad Al Hadid is a lecturer at the University of Jordan. He completed his PhD degree at the University of Banking and Financial Sciences (Jordan) in 2010, obtained his MSc degree in Computer Science at Amman Arab University (Jordan) in 2005, and earned his BSc degree in Computer Science at Al-Zaytoonah University (Jordan) in 2002. He has published many research papers in different fields of science in refereed journal and international conference proceedings. His researches focus on self-healing architecture; also, his research interests include AI, knowledge-based systems, security systems, compression techniques and algorithms, and information retrieval.

Alia Abu-Tayeh earned her PhD degree in 1995. She is a lecturer in the University of Jordan (Aqaba) and an ex-lecturer in King Hussein University. She published many scientific articles in renowned journals. Her interest ranges from linguistics to applied mathematics in computer and languages.

Author details

Evon Abu-Taieh[1]*, Auhood Alfaries[2], Nabeel Zanoon[3], Issam H. Al Hadid[1] and
Alia M. Abu-Tayeh[1]

*Address all correspondence to: abutaieh@gmail.com

1 The University of Jordan, Aqaba, Jordan

2 King Saud University, Riyadh, Saudi Arabia

3 Al-Balqa' Applied University, Aqaba, Jordan

References

[1] Sharma MK, Samanta D. Word prediction system for text entry in Hindi. 13, 2, article 8 (June 2014). 2014. 29 pages. DOI=http://dx.doi.org/10.1145/2617590

[2] Xue N. Automatic inference of the temporal location of situations in Chinese text. Conference on Empirical Methods in Natural Language Processing (EMNLP '08). Stroudsburg, PA, USA: Association for Computational Linguistics. 2008. pp. 707-714

[3] Xue N, Converse SP. Combining classifiers for Chinese word segmentation. first SIGHAN workshop on Chinese language processing. 18. Stroudsburg, PA, USA: Association for Computational Linguistics. 2002. pp. 1-7. DOI: dx.doi.org/10.3115/1118824.1118839

[4] Xue N, Chen J, Palmer M. Aligning Features with Sense Distinction Dimensions. COLING/ACL on Main conference poster sessions (COLING-ACL '06). Stroudsburg, PA, USA: Association for Computational Linguistics; 2006. pp. 921-928

[5] Green S, Manning CD. Better Arabic parsing: Baselines, evaluations, and analysis. 23rd International Conference on Computational Linguistics. Beijing, China. 2010. pp. 394-402

[6] Marton Y, Habash N, Rambow O. Dependency parsing of modern standard Arabic with lexical and inflectional features. Computational Linguistics. 2013;**39**(1):161-194. DOI: 10.1162/COLI_a_00138

[7] Parvez MT, Mahmoud SA. Offline Arabic handwritten text recognition: A Survey. ACM Computer Survey. 2013;**45**(2):35. DOI: dx.doi.org/10.1145/2431211.2431222

[8] Subramanian K, Prasad R, Natarajan P. Robust named entity detection using an Arabic offline handwriting recognition system. The Third Workshop on Analytics for Noisy Unstructured Text Data (AND '09). New York, NY, USA: ACM. 2009. pp. 63-68. DOI: dx.doi.org/10.1145/1568296.1568308

[9] Fard MM, Moghadam M, Bidgoli BM, Hussain M. Persian on-line handwritten character recognition by RCE spatio-temporal neural network. 5th international conference on Soft computing as transdisciplinary science and technology (CSTST '08). New York, NY, USA: ACM. 2008. pp. 90-94. DOI: dx.doi.org/10.1145/1456223.1456246

[10] Abu-Taieh E. Artificial neural networks: Enhanced back propagation in character recognition. In: Proceedings of Information Technology and Organizations: Trends, Issues, Challenges and Solutions. Volume 1. USA: IGI Global; 2003. pp. 263

[11] Aljarrah I, Al-Khalee O, Mhaidat K, Alrefai M, Alzu'bi A, Rabab'ah M. Automated system for Arabic optical character recognition. In: The 3rd International Conference on Information and Communication Systems (ICICS '12). 5, p. 6. New York, NY, USA: ACM. 2012. DOI: http://dx.doi.org/10.1145/2222444.2222449

[12] Ali A, Sagheer AM. Design and implementation of secure chatting application with end to end encryption. Journal of Engineering and Applied Sciences. 2017;**12**:156-160

[13] Supriana I, Nasution A. Arabic Character Recognition System Development. Juhana Salim MI, editor. Procedia Technology. 2013;**11**:334-341. DOI: 10.1016/j.protcy.2013.12.199

[14] Sarfraz M, Ahmed MJ, Ghazi SA. Saudi Arabian License Plate Recognition System. 2003 International Conference on Geometric Modeling and Graphics (GMAG'03). London, UK: IEEE Computer Society Press. 2003. pp. 36-41. DOI: 10.1109/GMAG.2003.1219663

[15] Izakian H, Monadjemi SA, Ladani BT, Zamanifar K. Multi-font Farsi/Arabic iso-lated character recognition using chain codes. International Journal of Computer and Information Engineering. 2008;**2**(7):67-70. Retrieved from http://waset.org/publications/301/multi-font-farsi-arabic-isolated-character-recognition-using-chain-codes

[16] Abandah GA, Khedher, Mohammed Z. Analysis of handwritten Arabic letters using selected feature extraction techniques. International Journal of Computer Processing of Languages. 2009;**22**(1):1-25. DOI: doi.org/10.1142/S1793840609001981

[17] Al-Taani AT, Al-Haj S. Recognition of on-line Arabic handwritten characters using struc-tural features. Journal of Pattern Recognition Research. 2010:23-37

[18] Kapogiannopoulos G, Kalouptsidis N. A fast high precision algorithm for the estima-tion of skew angle using moments. IASTED, International Conference Signal Processing, Pattern Recognition and Application. Crete, Grece: SPPRA. 2002. pp. 275-279

[19] Zidouri A. On multiple typeface Arabic script recognition. Research Journal of Applied Sciences Engineering and Technology. 2010;**2**(5):428-435. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.4044&rep=rep1&type=pdf

[20] Amin A. Recognition of printed Arabic text based on global features and decision tree learn-ing techniques. Pattern Recognition. 2000;**33**:1309-1323. DOI: 10.1016/S0031-3203(99)00114-4

[21] Cowell J, Hussain F. Extracting features from Arabic characters. IASTED International Conference on COMPUTER GRAPHICS AND IMAGING (pp. 201-206). Honolulu, Hawai, USA: ACTA Press. 2001

[22] Sood S, Fitzgerald E. Arabic Script and the Art of Calligraphy. Retrieved from The Metropol itan Museum of Art. 2012: www.metmuseum.org/learn/for-educators/publications-for-educators/art-of-the-islamic-world/unit-two/proportional-scripts

[23] Slimane F, Kanoun S, Alim AM, Henneber J, Ingold R. Comparison of global and cas-cading recognition systems applied to multi-font Arabic text. In Proceedings of the 10th ACM symposium on Document engineering (DocEng '10). ACM, New York, NY, USA, 161-164. 2010a. DOI=http://dx.doi.org/10.1145/1860559.1860591

[24] Slimane F, Kanoun S, Alimi AM, Hennebert J, Ingold R. Comparison of global and cascading recognition systems applied to multi-font Arabic text. In: Of the 10th ACM symposium on Document engineering (DocEng '10). New York, NY, USA: ACM. 2010b. pp. 161-164. DOI: http://dx.doi.org/10.1145/1860559.1860591

[25] Sourdel D. Ibn Muḳla. In: Bearman P, Bianquis Th, Bosworth CE, van Donzel E, Heinrichs WP, editors. Encyclopaedia of Islam. 2nd ed. Edited by: Consulted online on 01 July 2017 http://dx.doi.org/10.1163/1573-3912_islam_SIM_3306. First published online: 2012. First print edition: ISBN: 9789004161214, 1960-2007

[26] Grabar O. The Mediation of Ornament. The A.W. Mellon Lectures in the Fine Arts. Princeton, NJ: Princeton University Press; 1992. p. 38

[27] Osborn J. Narratives of Arabic script: Calligraphic design and modern spaces. The Journal of the Design Studies Forum. 2009;**1**(3). DOI: doi.org/10.1080/17547075.2009.1 1643292

[28] Schimmel A. (10 14). Styles of Calligraphy, Islamic Art & Architecture. Retrieved from Islamic Art and Architecture. 2011: http://islamic-arts.org/2011/styles-of-calligraphy

[29] Shah NN, Bhatt N, Ganatra A. A unique word prediction system for text entry in Hindi. In: Second International Conference on Information and Communication Technology for Competitive Strategies (p. Article No. 118). ACM. 2016. DOI: 10.1145/2905055.2905334

Human-AI Synergy in Creativity and Innovation

Tony McCaffrey

Additional information is available at the end of the chapter

Abstract

In order to maximize creative behavior, humans and computers need to collaborate in a manner that will leverage the strengths of both. A 2017 mathematical proof shows two limits to how innovative a computer can be. Humans can help counteract these demonstrated limits. Humans possess many mental blind spots to innovating (e.g., *functional fixedness, design fixation, analogy blindness*, etc.), and particular algorithms can help counteract these shortcomings. Further, since humans produce the corpora used by AI technology, human blind spots to innovation are implicit within the text processed by AI technology. Known algorithms that query humans in particular ways can effectively counter these text-based blind spots. Working together, a human-computer partnership can achieve higher degrees of innovation than either working alone. To become an effective partnership, however, a special interface is needed that is both human- and computer-friendly. This interface called *BrainSwarming* possesses a linguistic component, which is a formal grammar that is also natural for humans to use and a visual component that is easily represented by standard data structures. Further, the interface breaks down innovative problem solving into its essential components: a goal, sub-goals, resources, features, interactions, and effects. The resulting human-AI synergy has the potential to achieve innovative breakthroughs that either partner working alone may never achieve.

Keywords: creativity, innovation, human-computer interface, artificial intelligence, intelligence augmentation

1. Introduction

Recent critiques of IBM Watson in the business world (*Forbes* and *Fortune*) and technical world (*MIT Technical Review* and *Wired*) suggest concerns about Watson's abilities and potential [1–5]. One critique is based on Watson's inability to draw conclusions beyond the corpus it has been trained on. Another critique is that it cannot make connections, or draw

analogies, between different domains of data—such as between oncology and heart disease [1]. Yet another critique argues that AI, including IBM Watson, is still several breakthroughs away from being really intelligent, but this critique does not specify the particulars of these needed breakthroughs [5]. Our theory is that IBM Watson is exhibiting the limitations demonstrated in a 2017 mathematical proof of the limits to a computational approach to innovation and analogical thought [6]. No matter how many academic journal articles Watson processes about its topic of choice, it cannot overcome the proven limits. Further, human blind spots, such as *functional fixedness*, are implicit within the human-produced text and data used by Watson. Unless IBM Watson and other AI technologies face these limits and address them properly, they will continue to experience frustration over the uneven results that these technologies produce.

More generally, any computational approach to innovation and creativity (e.g., Machine Learning, Deep Learning, AI in general) has limits to how creative or innovative it can be. The 2017 mathematical proof details two of these limitations [6]. Humans can help counter these limits. On the other hand, humans have many known mental blind spots to innovation, including *functional fixedness* [7], *design fixation* [8], and *analogy blindness* [9, 10]—to name a few. For every known mental obstacle to innovation, there now exists an effective counter-technique, which can be implemented in software [11]. These counter-techniques help humans be more innovative as well as improve the AI technologies that operate on the text and data produced by humans.

From these findings, it makes sense to create a human-computer interface for innovation that is both human- and computer-friendly so that the computer can help humans be more innovative and humans can return the favor for the computer. The overall result thus far has been a human-computer partnership that has already found novel solutions to such tough problems as how to significantly reduce concussions in American football players and how to adhere a coating to the non-stick surface Teflon [11, 12]. This human-computer synergy has the potential to achieve even greater innovative breakthroughs.

This chapter first articulates new definitions of creativity, innovation, feature, and effect. These definitions permit quantified arguments about the innovation process. Next, the main points of the proof will be presented. All the proof's details are contained in Ref. [6]. The main conclusion is that no computational approach can fully take over the creative or innovative process.

Then, several of the weaknesses to human innovation will be presented along with their effective algorithmic counter-techniques. A full description of human weaknesses and programmable counter-techniques are contained in Ref. [11]. How these human weaknesses become computer weaknesses is explained with an emphasis on how the programmable counter-techniques can also improve the innovation of any AI technology. Finally, the human-computer interface that permits humans to counter computer limits and the computer to counter human weaknesses will be presented. This interface called *BrainSwarming* has the potential to result in greater innovative behavior than either humans or computers can achieve working alone.

2. Proven computer limits to innovation

Section 2.1 articulates the new definitions for creativity/innovation, feature, and effect, which then permit the quantification of the size of the space of innovation for physical objects. The

space of innovation for a given object is shown to consist of all possible effects that the object could produce when interacting with every other possible object, material, force, energy, and condition (e.g., barometric pressure and gravity strength). Section 2.2 quantifies the number of interactions that are possible between an object of interest and all other objects, materials, forces, and energies in the world. Section 2.3 builds on Section 2.2 by exploring all the ways that two given objects could interact within various conditions to produce interesting effects. The number of possible interactions and possible effects is so astronomically large that the fastest supercomputer today could not examine them even it started working from the invention of the first computer. Section 2.4 articulates other reasons why a computer could not predict certain effects of an interaction. Finally, Section 2.5 shows how humans can help counteract the challenges that computers have when innovating.

2.1. New definitions

Any creative/innovative solution to a problem is built upon at least one commonly over-looked or new feature of the problem. A feature that is commonly overlooked or new is called obscure. The above description is called the *Obscure Features Hypothesis of Innovation* [13].

If the solution was based upon a commonly noticed feature, then it would get a low rating on originality and a high rating on obviousness [11]. For example, if a scented jar candle company came out with a new scent called *Bag of Halloween Candy*, people might enjoy it and buy this new scent but it still might not receive very high creativity scores. The feature of scent has been explored a great deal in the context of candles. In contrast, if a candle company devised a self-snuffing candle that could be set to extinguish itself after a desired amount of time, then this would receive high creativity scores [11]. The feature of self-snuffing in the world of candles has been under-explored.

Given that features are a crucial aspect of creativity, a definition adapted from the philosopher Nietzsche permits the number of features of an object to be quantified [6]. Nietzsche states: "The features of a 'thing' are its effects on other 'things': if one removes other 'things,' then a thing has no features" [14]. From this perspective, every feature emerges from interactions and is not intrinsic to the object itself. Certainly, color is not intrinsic to an object, but results from light interacting with the object and our retinas, which results in processing in the human visual cortex. Change the circumstances of the interaction and the color changes. Change the lighting. Put on sunglasses. Experience trauma in the visual cortex. These and other changes can result in a change in color.

As a further example, the mass of an object now appears to be the result of the object interacting with Higgs bosons [15]. Mass and length of an object change as its speed increases as it nears the speed of light [16]. Even the size of an object depends on the gravitational field that it is experiencing. A table of a certain size might be stable within one gravitational field but collapse in another gravitational field because its legs cannot hold up the weight of its tabletop. Any feature of an object, in fact, can be described as the effect of interactions.

Given these definitions of creativity and feature, we are able to quantify the number of features, interactions, and effects by defining a feature as an effect that results from interacting the object of interest with other objects, materials, forces (e.g., centripetal and centrifugal), and energies (acoustic, magnetic, chemical, biological, human, thermal, electrical, hydraulic,

pneumatic, mechanical, electromagnetic, and radioactive: [17]). Given that some amount of a material (e.g., a patch of velvet or a chunk of steel) can be considered an object, we can leave out material from the definition of feature above. Also, the lists of forces and energies may increase someday, especially as we better understand dark matter and dark energy, but these lists are currently stable but potentially dynamic in the future.

2.2. Interactions

For our calculations, let us estimate that there are 10 million objects in the world. In April 2015, the US Patent Office issued its nine millionth patent [18], and this number does not include the patents unique to patent offices of other countries or the trade secrets contained in no patent databases. Further, this estimate leaves out natural objects (e.g., stone) and common objects (e.g., ball) that are also excluded from all patent databases. Further, the number of patented objects grows everyday as new patent applications are submitted on a daily basis. However, 10 million is a reasonable estimate for the present time, and it is an easy number with which to do calculations.

Given an object of interest, how many interactions are possible with 10 million objects? Strictly speaking, there are $2^{10,000,000}$ possible subsets of 10 million things, which is approximately 10^{80}, so our object of interest could interact with every possible subset of objects. More realistically, however, an engineer might interact their object of interest with between one and five other objects, which would result in on the order of 10^{27} subsets. Computers have existed for on the order of 10^9 seconds, so to examine all subsets of five or fewer objects would require examining $10^{27}/10^9 = 10^{18}$ subsets per second since the 1950s. The fastest supercomputer as of June 2015, the Tianhe-2, computes on the order of 10^{16} floating-point operation per second [19]. So, if the Tianhe-2 existed since the first computer existed, it could still not examine all the possible interactions of our object of interest with a reasonable number of subsets of possible objects. This calculation only allows one floating-point operation to process each subset. Further, it does not take into account all the possible conditions that these interactions could take place in: differing barometric pressures, humidity, temperature, lighting intensity, radiation, magnetic fields, strength of gravitational field, and so on.

In sum, even with our conservative estimates, the current fastest supercomputer could not fully explore the space of possible interactions for our object of interest in a reasonable amount of time.

2.3. Many ways to interact

The assumption made in the previous section is that, given two or more objects, it is obvious how they should interact. A spoon is used to stir the contents of a coffee cup, for example. That is what *functional fixedness* would dictate, which is the tendency to fixate on the designed use of an object, including when it is interacting with another object, and ignore the plethora of other possible uses [7]. But the spoon and the coffee cup could interact in an almost incalculable number of ways to achieve many different effects. For example, rest the spoon across the opening of a steaming cup of coffee. Place an ice cube in the spoon and watch the ice cube melt. Or, place a marble in the spoon as it rests across the cup's opening. Slap down on the handle part of the spoon hanging over the edge of the cup and launch the marble across the

room. Or, place the coffee cup on one side of the room and take the spoon to the other side of the room. Make the coffee cup into a target by trying to throw the spoon into the cup. Or, play golf with the spoon as the putter, a marble as the golf ball, a table top as the putting green, and a sideways coffee cup as the hole.

Or, again set the coffee cup upright on the counter and place the spoon horizontally so it rests across the opening of the cup. Turn the spoon over so the curved part is facing upward and play a game of trying to balance various objects on the curved surface so they do not fall into the cup. Or, shake a spoon around in an empty cup to make a rattling sound. Or, turn a coffee cup over so that the open end is facing down. Place a spoon into the open end of the empty coffee cup and set the contraption on the counter. The spoon will force one side of the coffee cup to elevate a bit, forming a trap. When the spoon is disturbed by a mouse, for example, the coffee cup will fall and flatten, possibly trapping the mouse.

These are just a few of the ways to interact the spoon and the coffee cup to achieve an interesting effect. To consider all the ways that these two objects could interact, we would have to take into account every possible spatial relation between the two objects; every possible speed, acceleration, and deceleration of the two objects with respect to each other; every possible type of movement (linear, nonlinear, spinning at various angles and speeds); every possible surface that they may rest upon; every possible lighting condition, wind condition, heat condition, radiation level, magnetic field strength, electrical current flow, barometric pressure, humidity, earthquake or turbulence condition, and gravity strength; every possible extra object involved in the interaction (e.g., ice cube, marble, liquid coffee, and a human); as well as other conditions that we are probably overlooking.

If any of these conditions is actually measured by a continuous variable, then the number of different interactions between the spoon and coffee cup is truly computably nonenumerable. Even if all these conditions are measured by discrete variables that extend to a finite number of decimal places, then the number of possible interactions is outlandishly large. All these digits of precision on a variable are probably unnecessary in most cases, but when one is approaching a phase transition (e.g., liquid coffee approaching gas or the ceramic coffee cup possibly becoming superconductive), then many decimal places might be necessary to understand the onset of the transition. If one is approaching a previously unknown phase transition, then the slightest change in one condition, as measured by a change in the 100th decimal place for that variable, for example, could produce a radically different effect.

In sum, although we calculated that there may be about 10^{18} possible interactions between one object and up to five other objects out of 10 million possible objects, taking into account the incredible number of ways that any two objects can interact with each other plus all the possible conditions that those interactions could take place in, raises our number of interactions at least several orders of magnitude and quite possibly many orders of magnitude [6]. The overall result is a number of possible interactions that becomes increasingly beyond the ability of current and projected supercomputers to explore even if they were running since the invention of the first computer.

When quantum computers come fully into being, then all the above calculations will need to be redone. There has been work showing how quantum computers could handle certain computably enumerable sets [20]. However, if any of the conditions (e.g., heating, humidity,

radiation, etc.) actually requires a continuous variable for its measurement, then the number of possible interactions is truly continuous and thus not computably enumerable. If all the conditions can be measured with discrete variables, then it is possible yet unclear whether the set of interactions is the type of set that is computably enumerable by a quantum computer, according to the specifications in Ref. [20]. Even if the set of possible interactions were computably enumerable, however, any gaps in the theories involving those interactions—as described in the next section—would make the set of derived effects from the set of interactions uncomputable.

2.4. Predicting effects computationally

Can a computer compute the effects of a set of objects or entities that are interacting? It depends on whether a theory exists that derives the particular effects under consideration. Sometimes, theory is ahead of empirical measurement and sometimes empirical measurement is ahead of theory. For the former, Einstein's General Relativity, developed between 1907 and 1915, predicted that light would bend around massive objects such as our Sun [16]. It took until 1919, however, until Arthur Eddington verified this prediction by measuring starlight that moved around a total solar eclipse [21]. For the latter, empirical measurement determined that galactic clusters did not have sufficient mass to account for their rotational speeds, so the existence of dark matter was posited as a way to increase gravitational effects present in galactic clusters [22].

If no theory exists to predict a particular effect of an interaction, then no algorithm exists to compute that effect. Given our previous example of a coffee cup interacting with a spoon, if there are gaps in the theories for how the interaction would proceed in a possible condition (e.g., lighting, wind, heat, radiation level, magnetic field strength, electrical current flow, barometric pressure, humidity, earthquake or turbulence condition, and gravity strength), then no computer could predict the effects within that particular configuration. That particular combination of conditions would have to be empirically measured. Thus, a computer's ability to list out a particular combination of conditions does not mean that the computer could successfully predict the effects of the interaction taking place within that amalgam of conditions.

2.5. Humans countering computer limits

Humans are needed to carry out the empirical measurements that neither a computer can carry out nor a robot has not been set up to execute. Further, with our vast experience of interacting with the physical world, humans already know many effects of interactions but have yet to encode them for a computer. If humans have not yet experienced the interaction, often we can comfortably predict the main effects of that interaction after running a mental simulation in the sensorimotor cortices of our brain [23, 24].

In this way, humans can help flesh out and teach the computer many effects that the computer does not currently know and is presently unable to derive. Further, humans are good at crafting theories that make predictions of effects that then can be empirically tested. So, humans can encode their theories that a computer can then use to derive effects. Although a computer will continue to learn new effects taught to it by humans and derive effects based on new

theories, given the computable nonenumerability of effects, humans will continue to maintain their rightful place in innovation—even with the onset of quantum computers (see previous Section 2.3 *Many Ways to Interact*).

3. Human weaknesses to innovation and counter-techniques

In this section, we present five human blind spots to being creative and innovative (i.e., *functional fixedness* [7], *design fixation* [8]; *analogy blindness* [9, 10], *goal fixedness* [11], *assumption blindness* [11]) as well as the effective algorithms, several of them patented, that can guide humans out of these creative dead ends. For a more complete discussion, we refer the reader to [11], which includes the treatment of other human creative blind spots.

3.1. Functional fixedness

Functional fixedness is the tendency to fixate on the common use of an object or one of its parts [7]. In 2012, the first highly effective counter-technique, the *generic parts technique* (GPT), was developed [13]. Consider the *Two Rings Problem* [26], in which you have to fasten two steel rings together in a figure-eight configuration. The rings are each about 6 inches in diameter and weigh about three pounds. All you have to work with is a long candle, a strike-anywhere match, and a two-inch cube of steel.

Most people first try to light the candle and drip wax around the rings. However, the rings are too heavy to be fastened securely with a wax bond. The key is to notice that the candle's wick is a string. Remove the string by scraping the wax away on the steel cube and tie the rings together.

People who used the GPT solved 67% more problems than a control group [13]. The idea is to break an object into its parts while you ask two questions. First, can the object be broken down further into smaller parts? For example, in **Figure 1**, *candle* can be broken down into *wax* and *wick*. Second, does the description imply a use? If so, re-describe it generically in terms of its shape, material, or size. For example, a wick implies burning to give off light. When re-described in terms of its material (i.e., string), this new more generic description opens up new possibilities for uses—especially, tying things together. In this case, this use is sufficient to solve the *Two Rings Problem*, but if an even more generic description is needed then perhaps *long, interwoven fibrous strands* would be about the next level of generic description.

Software that exists can find the solution to the *Two Rings Problem* and other problems requiring the discovery of obscure features if the key obscure features are either already known in a corpus or dataset or they have been articulated by a human in a targeted query [27]. For example, if a corpus/dataset "knows" that the verb *tie* is a synonym of *fasten, string* is a material for *wick*, and *tie* is a use of *string*, then software has generated the following solution to the *Two Rings Problem*: "a candle's wick is made of string, which might be able to tie ring to ring" [27]. If any of these connections is missing in the text/data, then the solution is difficult to reach. For example, if your data source is ConceptNet 5.5 [28], for example, then *tie* is a way to *fasten* things, *wick* is a part of *candle*, but *wick* is not a type of *string*, so it is difficult to
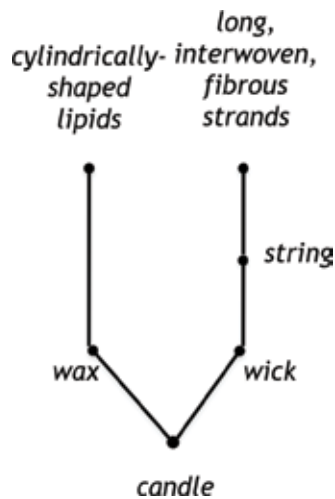
Figure 1. Generic parts technique.

reach the conclusion that you can use the wick to tie things together. In order to obtain this information, either a human must intervene to answer the questions of the *generic parts technique*, another text or data source is used to obtain the crucial information, or you find another possible route through ConceptNet 5.5 such as *wick* being a type of *cord*, which can be used to *tie* things. I chose ConceptNet 5.5 as an example because its information comes from multiple good sources [28]: including Open Mind Common Sense contributors, DBPedia 2015, OpenCyc 2012, and Open Multilingual WordNet.

3.2. Design fixation

Design fixation occurs when a designer attempts to create a novel design but fixates on the features of known designs they have seen [8, 29, 30]. For example, people instructed to create a novel candle might think up a new scent for a candle or add multiple wicks to the candle. In the context of candles, however, scent and number of wicks have been frequently explored. To be truly innovative, you need to manipulate a commonly overlooked (or perhaps new) feature of a candle. But how does one notice something that is rarely noticed?

Although the number of features of a candle (or any object) is intractable and not computably enumerable, classifying the types of features that any object could possess into an extensive category system has been a highly effective method for overcoming *design fixation* [11].

We initially listed 32 categories of feature types for objects, but now use a 50-category system [11]. We asked people to list as many features as they could for many common objects (e.g., candle, umbrella, etc.). We then categorized their answers based on our 32-category system. On average, people overlooked 20.7 of the 32 categories (67.4%) for each of the objects [11]. For each object, they overlooked different types of features. For example, for a rocking chair, they would notice motion—that the chair was designed to move in a certain way. For a candle, however, no one we tested ever noticed that a candle is motionless when it burns. Its flame flickers, but the candle itself does not move.

To be innovative, you need to build upon a feature that has been commonly overlooked and, based upon our findings, the majority of feature types of common objects are overlooked. Therefore, there is plenty of room to create novel variations for even the most common of objects.

For example, a candle that moves from its own dynamics is an under-explored type of candle. Examining the other overlooked features for a candle, we found that no one noticed that a candle loses weight when it burns. Thus, we leveraged weight loss to produce a candle in motion. By placing a candle on one side of a scale-like object and a counterweight on the other side, the candle moved upward slowly as it burned down. For fun, we placed a snuffer above the candle so that it eventually moved into the snuffer and extinguished itself. The *self-snuffing candle* was born [11]. A constructed prototype revealed that the *self-snuffing candle* works as described.

Computationally, in ConceptNet 5.5 [28], a candle has no connection with being motionless or losing weight while burning, while a rocking chair has many connections related to motion. ConceptNet 5.5, as an example of many textual and data sources, would not be a good source for noticing overlooked features that could become the basis of a novel design. The over-looked features need to be uncovered through another method such as using the extensive category system of feature types discussed above.

3.3. Goal fixedness

Goal fixedness occurs when a solver stays close to the original phrasing of the problem's goal and does not notice the various ways to phrase the goal in synonymous ways [11]. Any goal can be phrased in the following form: *verb noun-phrase prepositional-phrases*. The verb describes the change that is desired (e.g., *increase profits in the New England area during the holiday season*). In some cases, you want something to stay the same that is trying to change (e.g., *maintain altitude with one damaged engine*). The *noun-phrase* names what needs changing (e.g., *profits*) or maintaining (e.g., *altitude*). The prepositional phrases describe the important constraints and relations that need to hold true (e.g., *in the New England area during the holiday season*).

Focusing on the verb, people are able to list between 5 and 11 synonyms of a verb [11]. Humans drastically underperform when compared to the synonyms that are present in a good thesaurus. In *WordNet* [31], for example, the number of synonyms for the verbs we tested ranged between 24 and 172. Each synonym has nuances that may lead to new solutions.

For example, suppose a person was working on fastening the rings together in the *Two Rings Problem* [26] and used *WordNet* [31] to explore the synonyms of *fasten*. *WordNet* has a hierarchical structure to its synonyms. More specific synonyms are called *hyponyms*, while more general synonyms are called *hypernyms*. The hyponyms of a verb often name specific ways to achieve the change. There are 61 hyponyms of *fasten*, and they describe many ways to fasten things together, including *tie, weld, staple, velcro, clip, glue, buckle, pin, sew, clamp, chain, garter, clinch, strap, grout, lodge, cement, hasp, bind, button, latch,* and *rivet*. In this case, *tie* is the verb that names how to solve the *Two Rings Problem*. A computer program can easily solve the *Two Rings Problem* when it is revealed that *tie* is a hyponym of *fasten* and *string* is the material from which a wick is composed [27]. For countering *goal fixedness*, both *WordNet* and ConceptNet 5.5 are effective datasets.

3.4. Assumption blindness

Any phrasing of the goal belies many assumptions [11]. For example, a company was stuck on trying to adhere a coating to the nonstick surface Teflon. Everything they tried failed. However, some analysis of the verb *adhere* revealed some of its assumptions.

The verb *adhere* assumes a chemical solution often involving some type of adhesive. The verb *adhere* also assumes that two things are being adhered to each other, that the adherence is probably meant to be permanent, that the two things being adhered are in direct contact with each other, that the direct contact is playing an important causal role in the adherence, and so on.

Noticing three of the assumptions was crucial to a solution: (1) using a chemical process between (2) two surfaces where (3) contact is crucial to the solution. Exploring alternatives to these assumptions led to a novel solution: (1) using a magnetic process among (2) three surfaces where (3) contact is not crucial to the solution. Specifically, a magnetic surface is placed behind the Teflon surface, while the coating with some ferrous content is placed in front of the Teflon surface. The coating sticks through the Teflon to the magnetic surface and forms a kind of *Teflon sandwich*.

In general, there is a master list of 50 types of features that any physical solution might possess [11]: including size, shape, material, quantity, type of energy used (e.g., chemical, magnetic, etc.), spatial relations among the parts, symmetry, and motion. To uncover some important assumptions, simply proceed through the list and ask if the verb under consideration assumes anything about each of these feature types.

These types of assumptions are contained in neither ConceptNet 5.5 nor, most likely, any other current text or data source. These assumptions need to be unearthed carefully through a method such as the one described above.

3.5. Analogy blindness

Gick and Holyoak [9, 10] were the first to show experimentally how difficult it is for humans to notice by themselves how an idea from one area could be adapted as a solution in another area. For example, they had participants read a brief military story that held the crucial idea for solving a surgery problem [9, 10]. Thirty percent solved the surgery problem after mere exposure to the military problem, but 80% solved it after being told to use the military problem to help solve the surgery problem.

Building upon the work of Julie Linsey and colleagues [32–36], who focused upon looking at synonyms of the main verb expressing the goal of the problem, McCaffrey and Krishnamurty [11] went one step further to explore the synonyms of both the verb and noun phrase of the goal. For example, consider the goal *reduce concussions in American football players.* The goal is in the form *verb noun-phrase prepositional-phrase.* Focusing upon *reduce concussions*, we explored the synonyms of both words and took into consideration some basic engineering knowledge. This process led to an extensive list of alternative phrasings:

including *diminish trauma, lessen impact, reduce energy, soften collision, minimize force, decrease momentum,* and *repel energy.*

Next, we entered each of the phrases into Google in the form "concussions diminish trauma" [12]. This step helped us determine which phrases were under-explored in the context of concussions. We found that *repel energy* was almost completely ignored. The word *repel* is closely associated with magnets, and this connection quickly triggered the creation of a possible solution. Magnetize all football helmets with the same pole so they do not want to be near each other. Tests with models showed that potential head-on collisions were turned into glancing blows as the helmets slowed down and slightly veered when approaching each other at high speeds [12].

In the *BrainSwarming* graph for *reduce concussions* (**Figure 2**), the goal was placed at the top and the alternative goal phrasings grew downward from the top. The resources were divided into two types: *objects* and *energies,* and placed across the bottom. A solution was constructed in the middle showing an interaction between the two helmets and magnetic energy. This interaction satisfied the subgoal *repel energy,* which satisfied the main goal *reduce concussions.*

Obtaining synonyms from *WordNet*, ConceptNet 5.5, or other sources can definitely uncover nuanced phrasings of the goal that may illuminate novel solutions to a problem. These data sources and the process of creating alternative goal phrasings could potentially help any AI technology that is focused on the task of problem solving.
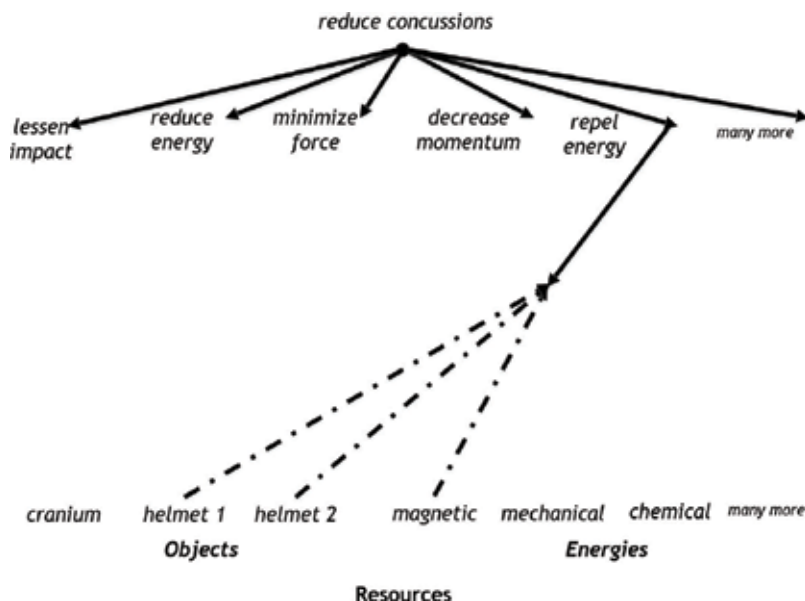
Figure 2. BrainSwarming graph for reduce concussions.

4. Human-computer interface to achieve synergy

In order for humans to counter computer limits and computers to counter human weaknesses, an interface is needed, which is comprised of data structures that both humans and computers can easily populate. In order to make the human-computer interaction efficient, the interface needs to be both human- and computer-friendly. Building upon our new definitions, we can define a problem as a set of desired effects and define a solution as a sequence of interactions that ultimately produces the desired effects named in the problem.

We define the problem solving grammar for innovation in Extended Backus-Naur Form (EBNF: [37]), which is a compact notation mostly used for defining the syntax of computer programming languages. For our grammar, we only need a few of EBNF's symbols: "::=" means "is defined as," a superscripted "+" means there can be one or more of the preceding item, and a superscripted "*" means there can be zero or more of the preceding item.

The bidirectional *BrainSwarming* graph has been tested with various age groups and has been found to be easy to understand [25]. As illustrated in **Figure 3**, a goal is placed at the top, and the refinements of the original goal grow downward below it. The resources to solve the problem are placed across the bottom, and the parts and features grow upward above the resources. The two directions grow toward each other until they connect, at which point you have your first candidate solution. The solution is comprised of a sequence of interactions between resources, features, and parts until the goal's effects are satisfied. Humans find the graph intuitive, and the computer can easily represent the bidirectional graph as a set of trees.

A goal is a set of desired effects. Any effect can be described as an action verb that describes a change (or a nonchange), a noun phrase to name that which needs changing (or should be kept from changing), and a list of prepositional phrases that describe important constraints and
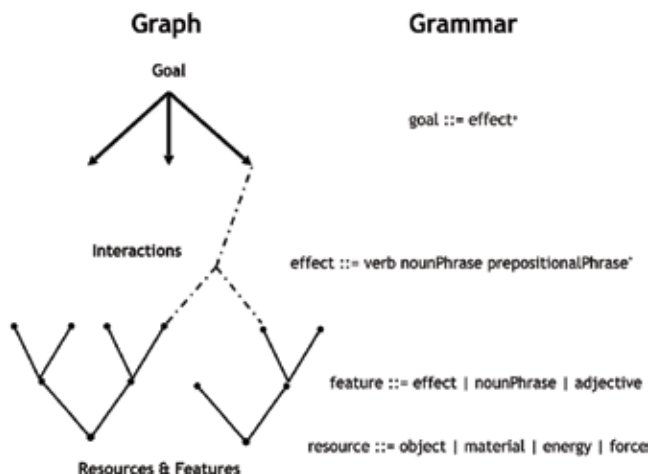
Figure 3. BrainSwarming graph and grammar.

relations. A feature is synonymous with an effect, but sometimes a shorthand can be used: an adjective (e.g., heavy) or a noun phrase (e.g., a heavy, metal rectangle). A resource is either an object (e.g., hammer), a material (e.g., velvet), an energy (e.g., magnetic), or a force (e.g., centrifugal).

Each of these grammatical forms is both human- and computer-friendly. Each phrase has a natural English form, and each phrase is regular so that it is easy for a computer to parse.

5. Current implementation and applications

The current implementation of *BrainSwarming* (brainswarming.io) includes the visual graph and grammar of **Figure 3**, as well as software to counteract *functional fixedness*, *design fixation*, *goal fixedness*, and *analogy blindness* [11]. The counter-technique for *assumption blindness* has not been implemented at this time. *BrainSwarming's* visual graph is simultaneously updatable online by multiple users working from different locations. *Analogy Finder* [11, 38], the counter-technique to *analogy blindness*, currently searches the U.S. Patent database for analogous solutions from different fields (i.e., solutions that achieve the same basic effect and can be adapted to the particulars of the problem at hand). The user of *Analogy Finder* enters the verb and noun-phrase of a goal (e.g., *reduce concussions*). *Analogy Finder* then explodes the phrase into many synonymous phrases. The user selects which phrases to use in the search and *Analogy Finder* finds patents that achieve the desired effect expressed in many different ways (jargons) across all domains of the U.S. Patent database. *Analogy Finder* could easily be made to search any dataset or corpus (e.g., academic journals, *Wikipedia, etc.*).

In *BrainSwarming's* current implementation, it has helped humans create a novel, magnetic solution to reduce concussions [12, 39], an original way to stick a coating to Teflon [11], as well as several other new solutions to proprietary problems. For the Teflon problem, the *assumption blindness* technique was executed by hand, as it has not yet been implemented, although the synonym generation and analogous search were conducted by the software.

Because *BrainSwarming* allows multiple users to remotely access it, one of *BrainSwarming's* users could be a software program—specifically, an AI program such as IBM Watson or a machine learning program. Of course, the AI program would need to be able to communicate through an API to join the *BrainSwarming* team that is working on a particular problem. In this way, the humans on the team could help the AI program overcome the proven limits that hold for any computational approach to innovation. While *BrainSwarming's* software explicitly counteracts certain human obstacles to creativity (e.g., *functional fixedness*), the contributing AI program could help the team of human users overcome such things as noticing patterns that humans have trouble noticing (e.g., machine learning programs), revealing relevant information that a user has not yet read (e.g., IBM Watson), revealing information that is outside a user's expertise, or making connections that a user has yet to make. Finally, *BrainSwarming's* software can help counteract the obstacles to innovation (e.g., *functional fixedness* or *analogy blindness*) that are implicit in the corpora and datasets being used because humans produced these sources of information.

In this way, *BrainSwarming* counteracts the known cognitive obstacles to innovation in the human users as well as in the human-produced corpora and datasets. It also provides a highly visual interface that humans and any contributing AI program can access. The interaction among the human users, the visual interface, *BrainSwarming's* software that counteracts obstacles to creativity, and any contributing AI programs has the potential to achieve great innovative breakthroughs. Much testing is required to test the innovative power of the *BrainSwarming* platform when an AI program such as IBM Watson interacts with it as one of its users. The crucial comparison would be the innovativeness of IBM Watson (or another AI program) on its own compared to IBM Watson interacting with the *BrainSwarming* platform and some human users. Even without a contributing AI program, *BrainSwarming* helped human users become more innovative by coming up with novel solutions to some very difficult problems [11, 12, 39].

6. Conclusions

Every innovative solution is built upon an obscure (i.e., commonly overlooked or new) feature of the problem [13, 27]. Both computers and humans tend to overlook different sets of obscure features based on their differing search biases. These differences are somewhat complementary so that computers and humans can help each other uncover obscure features that the other partner would miss [6, 11]. The result is significantly more unearthed features of the problem, resulting in a higher chance of unearthing the key obscure features required for a novel solution to the problem. Further, computers cannot completely take over the creative and innovative process due to the fact that the set of features of any object is not computably enumerable, so it cannot be fully explored by a computational device [6]. Working together through a computer- and human-friendly interface called *BrainSwarming* permits computers and humans to easily innovate together. This human-computer synergy has already produced innovative solutions to some difficult industry problems [11, 12, 39]. Adding an AI program such as IBM Watson to the team innovating together through *BrainSwarming* has the potential to produce innovative breakthroughs that neither IBM Watson nor the human team could achieve on its own.

Specifically, imagine IBM Watson plugged into the *BrainSwarming* interface. It populates the bi-directional graph and reads information placed in the graph by the human users. In this way, it dynamically informs the human users of its insights as well as learns of the insights from the human users. Specifically, humans' implicit knowledge of many features and effects that have yet to be encoded can be communicated to Watson. New empirical results from tests conducted by humans can be entered for Watson to use. Further, Watson can learn obscure features from the *generic parts technique* in order to overcome *functional fixedness*. It can use overlooked features uncovered from the 50-category feature type list in order to overcome *design fixation*. It can capitalize on the 50-category feature type list to unearth assumptions hidden behind the main goal verb in order to counter *assumption blindness*. Watson can also leverage the synonyms of *WordNet* and ConceptNet 5.5 in order to overcome *goal fixedness* and *analogy blindness.* In turn, IBM Watson can offer its considerable knowledge and insight to the *BrainSwarming* graph. Overall, IBM Watson, or any AI technology, can interact with humans

through the *BrainSwarming* interface to uncover crucial obscure features for the problem at hand. Because innovative solutions are built upon obscure features, any AI technology using the *BrainSwarming* interface can potentially achieve higher levels of innovativeness than either the human users or the AI technology can achieve on its own. We look forward to testing this exciting hypothesis.

Acknowledgements

Conflict of interest

This research is associated with my company, Innovation Accelerator, Inc. (www.innovationaccelerator.com), and may lead to the development of software products, in which I have a business and/or financial interest. I have in place an approved plan for managing any potential conflicts arising from this arrangement.

Author details

Tony McCaffrey

Address all correspondence to: tony@innovationaccelerator.com

Innovation Accelerator, Inc., West Brookfield, Massachusetts, United States

References

[1] Bloomberg J. Is IBM Watson a 'Joke'? [Internet]. 2017. Available from: https://www.forbes.com/sites/jasonbloomberg/2017/07/02/is-ibm-watson-a-joke/#76e55bb8da20 [Accessed January 5, 2018]

[2] Freedman D. A Reality Check on IBM's AI Ambitions [Internet]. 2017. Available from: https://www.technologyreview.com/s/607965/a-reality-check-for-ibms-ai-ambitions/ [Accessed: January 5, 2018]

[3] Schank R. The Fraudulent Claims Made by IBM about Watson and IBM [Internet]. 2017. Available from: http://www.rogerschank.com/fraudulent-claims-made-by-IBM-about-Watson-and-AI [Accessed: January 5, 2018]

[4]   Darrow B. Has IBM Watson's AI Technology Fallen Victim to Hype? [Internet]. 2017. Available from: http://fortune.com/2017/06/28/ibm-watson-ai-healthcare/ [Accessed: January 5, 2018]

[5]   Frank B. AI is Still Several Breakthroughs Away from Reality [Internet]. 2017. Available from: https://venturebeat.com/2017/06/23/ai-is-still-several-breakthroughs-away-from-reality/ [Accessed: January 5, 2018]

[6]   McCaffrey T, Spector L. An approach to human–machine collaboration in innovation. Artificial Intelligence for Engineering Design, Analysis and Manufacturing. 2017:**32**:1-15. DOI: 10.1017/S0890060416000524

[7]   Duncker K. On problem-solving. Psychological Monographs. 1945;**58**(5):1-113

[8]   Jansson D, Smith S. Design fixation. DES Studies. 1991;**12**(1):3-11

[9]   Gick M, Holyoak K. Analogical problem solving. Cognitive Psychology. 1980;**12**:306-355

[10]   Gick M, Holyoak K. Schema induction and analogical transfer. Cognitive Psychology. 1983;**15**(1):1-38

[11]   McCaffrey T, Krishnamurty S. The obscure features hypothesis in design innovation. International Journal of Design Creativity and Innovation. 2014:**3**(1):1-28

[12]   McCaffrey T, Pearson J. Find innovation where you least expect it. Harvard Business Review. 2015;**93**(12):82-89

[13]   McCaffrey T. Innovation relies on the obscure: A key to overcoming the classic *functional fixedness* problem. Psychological Science. 2012;**23**(3):215-218

[14]   Nietzsche F. Will to Power (Translated by Kaufmann W, Hollingdale R). New York: Random House; 1968 (Original work published 1901)

[15]   Ellis J, Gaillard M, Nanopoulos D. An Updated Historical Profile of the Higgs Boson [Internet]. 2015. Available from: http://arxiv.org/pdf/1504.07217.pdf [Accessed: January 5, 2018]

[16]   Einstein A. Relativity: The Special and the General Theory (Translated by Lawson R). New York: Barnes & Noble; 2004 (Original work published 1920)

[17]   Hirtz J, Stone R, McAdams D, Szykman S, Wood K. A functional basis for engineering design: Reconciling and evolving previous efforts. Research in Engineering Design. 2002; **13**(2):65-82

[18]   USPTO Patent Full-Text and Image Database [Internet]. Washington, DC: US PatentDatabase; 2016. Available from http://patft.uspto.gov/netacgi/nph-Par-ser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=9000000.PN.&OS=PN/9000000&RS=PN/9000000 [Accessed: February 11, 2016]

[19]   Top500 Lists. Top 500 [Internet]. 2015. Available from: http://www.top500.org/lists/2015/06/ [Accessed: February 11, 2016]

[20] Calude C, Tadaki K. Spectral Representation of Some Computably Enumerable Sets with an Application to Quantum Provability [Internet]. 2013. Available from: https://arxiv.org/abs/1303.5502 [Accessed: January 5, 2018]

[21] Kennefick D. Testing relativity from the 1919 eclipse—A question of bias. Physics Today. 2009;**62**(3):37-42

[22] Zwicky F. On the masses of nebulae and of clusters of nebulae. Astrophysical Journal. 1937;**86**:217

[23] Battaglia P, Hamrich J, Tennenbaum J. Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences of the United States of America. 2013;**110**(45):18327-18332 Available from: http://www.pnas.org/content/110/45/18327.full

[24] Hegarty M. Mechanical reasoning by mental simulation. Trends in Cognitive Sciences. 2004;**8**(6):280-285

[25] McCaffrey T. Overcome any Obstacle to Creativity. Washington, DC: Rowman & Littlefield; 2018

[26] McCaffrey T. The obscure features hypothesis for innovation: one key to improving human innovation [unpublished doctoral dissertation]. University of Massachusetts, Amherst. 2011

[27] McCaffrey T, Spector L. How the obscure features hypothesis leads to innovation assistant software. In: Proceedings of the 2nd International Conference of Comput Creativity (ICCC). Mexico City, Mexico: ICCC; 2011. pp. 120-122

[28] ConceptNet 5.5 [Internet]. 2017. Available from: conceptnet.io [Accessed: January 5, 2018]

[29] Smith S. Getting into and out of mental ruts: A theory of fixation, incubation and insight. In: Sternberg R, Davidson J. The Nature of Insight. Cambridge, MA: MIT Press; 1995. pp. 229-251

[30] Smith S, Ward T, Schumacher J. Constraining effects of examples in a creative generation task. Memory & Cognition. 1993;**21**:837-845

[31] Miller G. WordNet: A lexical database for English. Communications of the ACM. 1995; **38**:39-41

[32] Linsey J. Design-by-analogy and representation in innovative engineering concept generation [dissertation]. The University of Texas at Austin, Austin, TX. 2007

[33] Linsey J, Laux J, Clauss E, Wood K, Markman A. Increasing innovation: A trilogy of experiments towards a design-by-analogy method. In: Proceedings of ASME Des Engineering and Technical Conference (IDETC). Las Vegas, NV: IDETC; 2007. pp. 1-15

[34] Linsey J, Markman A, Wood K. WordTrees: A method for design-by-analogy. In: Proceeding of the 2008 Amer Soc Eng Educ Ann Conf (ASEEAC). Pittsburg, PA: ASEEAC; 2008. pp. 1-13

[35] Linsey J, Wood K, Markman A. Increasing innovation: Presentation and evaluation of the WordTree design-by-analogy method. In: Proceedings of the ASME 2008 International

Design Engineering Technical Conferences & Computers and Information in Engineering Conferences. New York, NY: ASME; 2008. pp. 21-32

[36] Linsey J, Markman A, Wood K. Design by analogy: A study of the WordTree method for problem re-representation. ASME Transactions, Journal of Mechanical Design. 2012;**134**:041009-041012

[37] Aho A, Sethi R, Ullman J. Compilers: Principles, Techniques, and Tools. New York: Addison Wesley; 1986

[38] McCaffrey A. Analogy Finder. U.S. Patent US9501469B2. Washington, DC: USPTO; November 22, 2016

[39] Marks P. Eureka machines. New Scientist. 2015;**227**(3036):32-35

Min *k*-Cut for Asset Selection in Risk-Based Portfolio Strategies

Saejoon Kim and Soong Kim

Additional information is available at the end of the chapter

Abstract

Risk-based portfolio strategies such as the equal-weighted, the minimum variance, and the risk parity portfolios vie to find portfolios that are well diversified according to their respective measures. In this chapter, we propose asset-selected risk-based portfolio strategies that aim to reduce the two known weaknesses of these strategies, namely the large portfolio size and poor diversification with respect to other measures. We formulate this task as a minimum k-cut problem through which we establish asset selection from all assets in the investable universe before the risk-based strategy is applied. Empirical results on the data sets of the S&P 500 and the KOSPI 200 indicate that our asset-selected risk-based portfolio strategies possess superior properties across extensive performance measures compared to the baseline risk-based strategies.

Keywords: alternative portfolio management, smart beta strategy, risk-based portfolio, minimum *k*-cut, portfolio optimization

1. Introduction

Portfolio selection has been a main research topic in finance for over 60 years, dating back to the seminal paper by Markowitz [1] which laid the foundation of modern portfolio theory. Markowitz's analysis has established the mean–variance-efficient portfolios which achieve optimal tradeoff between return and risk. Extensive efforts have been made into various directions since then including the development of the capital asset-pricing model (CAPM) [2]. Most recent studies on this topic of portfolio selection have focused on alternative or "smart" beta strategies which exploit risk premia other than the systematic risk, or seek a better diversification of risk. These are quantitative approaches to portfolio selection and have played an important role in the field of portfolio management recently. These lie in between an

active management and the passive management manifested by the perennial market capitalization-weighted portfolio. A major category in the class of alternative beta strategies is the risk-based portfolio strategies [3–6] whose main objective is to manage risk more effectively than the market capitalization-weighted portfolio.

Currently, the existing representative risk-based portfolio strategies include the equal-weighted portfolio in which every asset has an equal weight, the minimum variance portfolio that achieves the smallest volatility, and the risk parity portfolio of [7, 8] in which every asset in the portfolio is exposed to equal risk. Clearly, diversification benefits of the three risk-based portfolios do not perfectly coincide. An important and favorable characteristic of these strategies is that they do not require the estimation of the expected returns, which is very error-prone, in their formulations. A somewhat comprehensive description of these strategies in terms of risk factors was presented in [3], and it was shown that the equal-weighted and the risk parity portfolios are special cases of the constrained minimum variance portfolio in [7]. A general framework of the quantitative asset allocation models of the three risk-based portfolio strategies has been presented in [5], and a detailed comparison of the strategies has also been provided in [6].

In this chapter, we propose to improve the characteristic and the performance of the risk-based portfolio strategies. Firstly, we address the presence of an inherent problem in the exact implementations of the equal-weighted and the risk parity portfolios that arise from the large cardinality of the respective portfolios. By construction, each of these portfolios has cardinality equal to that of the investable universe which can be too large to be implemented exactly in practice for many investors. To this end, we investigate a preselection of assets from the set of investable universe prior to implementing the risk-based portfolio strategies in order to reduce the portfolio cardinality to a more manageable size. This part relates to the improvement in the "characteristic" of the risk-based strategies.

Secondly, we address the relative weakness of a risk-based strategy with respect to some diversification measures. For example, the minimum variance portfolio produces the portfolio with the smallest variance, however, also one that is poorly diversified with respect to weight and to risk. Similarly, the equal-weighted portfolio produces the portfolio with a perfect weight diversification but one that also has a relatively high variance. To this end, as in the first case, we consider a preselection of assets and, in particular, the selection of "diversified" assets that will endow each risk-based strategy a "better" assets pool from which the portfolios are constructed. Consequently, the risk-based portfolio strategies defined on our diversified assets pool will perform superior to those defined on the original investable universe across different diversification measures and also with respect to other more popular performance measures such as the return and the Sharpe ratio as well; our results will be shown later. This part relates to the improvement in the "performance" of the risk-based strategies.

The described improvements are achieved simultaneously by formulating the problem as a minimum k-cut problem with assets represented as vertices in the graph. As the risk-based strategies are applied only after the assets have been selected for inclusion in the assets pool, we call our proposed strategies "asset-selected risk-based portfolio strategies." Furthermore, our asset-selected risk-based portfolio strategies require modest additional computational cost to the respective baseline risk-based ones.

We tested and compared the risk-based strategies with the proposed asset-selected risk-based strategies on the data sets of the S&P 500 from 1990 to 2016 and the KOSPI 200 from 2002 to 2016 where the latter is the set of sector-diversified largest 200 companies by market capitalization listed in the Korea Stock Exchange. Therefore, the assets that we refer to in this chapter are all stocks.

The organization of this chapter is as follows. In the next section, we present the formulation of the three risk-based portfolio strategies along with the associated diversification measures. In Section 3, we present the minimum *k*-cut-based asset selection method that forms the basis from which our contributions of this chapter come. It is followed by the presentation of the asset-selected risk-based portfolio strategies in Section 4. Extensive empirical results for the strategies are presented in Section 5, and the conclusion of the chapter is provided in Section 6.

2. Risk-based portfolio strategies

We consider the following three types of risk-based portfolio strategies from which we aim to make asset selection in an effective fashion: equal-weighted, minimum variance, and risk parity. We also make comparison with the market capitalization-weighted portfolio. While this portfolio does not generate competitive returns, it has a nice property of automatic rebalancing and serves as a common benchmark against which other alternative investment strategies can assess their performances. We explore these portfolios under the long-only constraint which will guarantee a unique solution and provide more realistic investing environment for many investors. We list the four portfolio strategies subsequently where the corresponding abbreviations are shown inside the parentheses:

1. Market capitalization-weighted portfolio (M)

2. Equal-weighted portfolio (EW)

3. Minimum variance portfolio (MV)

4. Risk parity portfolio (RP)

To formulate the portfolio strategies, let us introduce some notations that will be used throughout this chapter:

N		number of assets in the investable universe
Σ		the covariance matrix
σ_i		volatility of asset i
$\boldsymbol{\sigma}$	$=$	$(\sigma_1, \cdots, \sigma_N)^t$
x_i		weight of asset i in the portfolio
$\mathbf{x}$	$=$	$(x_1, \cdots, x_N)^t$, portfolio weight vector
$\sigma(\mathbf{x})$		portfolio volatility using $\mathbf{x}$ as the portfolio weight vector

Therefore, in our notations, σ and $\mathbf{x}$ are length$-N$ vectors where t denotes the transpose operator. Note that $\sigma(\mathbf{x}) = \sqrt{\mathbf{x}^t \Sigma \mathbf{x}}$. As each portfolio strategy is completely defined by its

portfolio weight vector, let us describe each strategy by its weight vector. EW is the strategy in which all assets are given equal weights. The weight vector $\mathbf{x}_{EW}$ for EW is given by the following:

$$\mathbf{x}_{EW} = \frac{1}{N}\mathbf{1} \tag{1}$$

where $\mathbf{1}$ is the all-ones vector. While the strategy looks too simple to generate consistent positive excess returns, it has shown to outperform M considerably, and no other alternative investment strategies were consistently better than EW for a wide selection of markets and holding periods [3, 4]. MV [1] is the strategy in which the portfolio volatility is minimized. The weight vector $\mathbf{x}_{MV}$ for MV is given by the following:

$$
\begin{aligned}
\mathbf{x}_{MV} \quad &= \quad \arg\min_{\mathbf{x}} \left(\mathbf{x}^t \Sigma \mathbf{x}\right) \\
\text{s.t.} \quad &\mathbf{1}^t\mathbf{x} = 1, \\
&\mathbf{x} \geq \mathbf{0}.
\end{aligned} \tag{2}
$$

MV provides an optimal return-risk tradeoff, and in particular, it lies on the leftmost tip of the efficient frontier curve. MV is perhaps the most stable portfolio among the ones along the efficient frontier curve as most of the estimation errors come from that of the returns, and this fact adds to the appeal of MV. RP [7, 8] is the strategy in which the risk associated with each asset is the same across all assets in the portfolio. Specifically, let the risk contribution of asset i, RC_i, be defined as

$$RC_i = x_i \cdot \frac{\delta\sigma(\mathbf{x})}{\delta x_i} = \frac{x_i^2 \sigma_i^2 + x_i \sum_{j \neq i} x_j \sigma_{ij}}{\sigma(\mathbf{x})} \tag{3}$$

where $\frac{\delta\sigma(\mathbf{x})}{\delta x_i}$ is the *marginal* risk contribution of asset i, and σ_{ij} is the covariance of assets i and j. RP requires that.

$$RC_i = RC_j \quad \text{for all } i, j.$$

Note that $\sum_{i=1}^{N} RC_i = \sigma(\mathbf{x})$, and thus the risk contribution from each asset adds up to the portfolio risk or volatility. We also note that in MV, the marginal risk contributions are all equal for all assets, that is, $\frac{\delta\sigma(\mathbf{x})}{\delta x_i} = \frac{\delta\sigma(\mathbf{x})}{\delta x_j}$ for all i, j. It is known that RP possesses a unique solution in long-only investment environment [9] which is our case under study. The weight vector $\mathbf{x}_{RP}$ for RP is given by the following:

$$
\begin{aligned}
\mathbf{x}_{RP} \quad &= \quad \arg\min_{\mathbf{x}} \left(\frac{1}{2}\mathbf{x}^t \Sigma \mathbf{x} - \sum_{i=1}^{n}\ln x_i\right) \\
\text{s.t.} \quad &\mathbf{1}^t\mathbf{x} = 1, \\
&\mathbf{x} > \mathbf{0}
\end{aligned} \tag{4}
$$

which is a convex optimization formulation that can be computed efficiently [9]. One can obtain an equivalent, but computationally less efficient, optimization problem as earlier and show that RP functions as a tradeoff between EW, which maximizes weight diversification with perfect disregard for volatility or variance reduction, and MV, which minimizes variance with perfect disregard for weight diversification [7]. As a consequence, it can be deduced that.

$$\sigma(\mathbf{x}_{MV}) \le \sigma(\mathbf{x}_{RP}) \le \sigma(\mathbf{x}_{EW}).$$

To quantify the amount of diversification attained by the three risk-based portfolio strategies, in the following, we list the three diversification profile measures [6] for the weight vector **x** of a portfolio strategy. They are (1) weight diversification

$$\frac{1}{N \sum\limits_{i=1}^{N} x_i^2}, \tag{5}$$

(2) volatility reduction

$$\frac{\mathbf{x}_{MV}^t \Sigma \mathbf{x}_{MV}}{\mathbf{x}^t \Sigma \mathbf{x}} \tag{6}$$

and (3) risk diversification

$$\frac{1}{N \sum\limits_{i=1}^{N} (RC_i / \sigma(\mathbf{x}))^2}. \tag{7}$$

These equations show that the weight vector **x** for EW, MV, and RP achieves the highest weight diversification, volatility reduction, and risk diversification, respectively. Note that each of the measures assumes values between 0 and 1, which represent the lowest and the highest levels of diversification attained, respectively.

To get an idea of how the three risk-based strategies fare with respect to the three diversification profile measures, in **Table 1**, we list the values of the risk-based portfolio strategy-diversification profile measure pairs for the S&P 500 and the KOSPI 200 data. The values shown are the averages of the 105 and 56 periods we used in our experiments for the S&P 500 and the KOSPI 200 data, respectively. The table shows a similar pattern for the two data sets with respect to the strong and weak profile measures for each risk-based portfolio strategy, suggesting that this is a characteristic of the strategies. In particular, EW showed a relatively high-risk diversification but a relatively weak volatility reduction as intuition might suggest. MV showed a very low-weight diversification and risk diversification. RP showed a relatively high-weight diversification but relatively low volatility reduction.

Before we further proceed to examine our asset-selected portfolio strategies, let us describe our experiment setting which is as follows. For the S&P 500 data, the investing time horizon spans from February 1, 1990, to May 2, 2016, that constitutes a total of 105 quarters, and for the

	Weight diversification	Volatility reduction	Risk diversification
S&P 500			
EW	1.000	0.229	0.838
MV	0.047	1.000	0.047
RP	0.765	0.329	1.000
KOSPI 200			
EW	1.000	0.290	0.873
MV	0.095	1.000	0.095
RP	0.801	0.397	1.000

Table 1. Diversification profile measure summary.

KOSPI 200 data, it spans from May 2, 2002, to 2016, for a total of 56 quarters. Both sets of data were collected on May 2, 2016, and the stock prices have been adjusted for dividends and splits before experiment.

The daily closing prices of stocks in the formation or the look-back period of a length of 252 days were used to execute the risk-based portfolio strategies whose resulting portfolios were put into effect the following trading day. There were no missing data in the closing prices of stocks, and no preprocessing of data was made as is common in the finance literature. This portfolio is held for one quarter after which portfolio rebalancing is performed. By portfolio rebalancing, we mean an independent and new execution of the portfolio strategy using data in the formation period of the most recent 252 days to update or rebalance the portfolio. This process is iterated throughout the investing time horizon, and specifically, portfolio rebalancing was made after market close on the last trading day of January, April, July, and October of each year. To clarify the terms used in this chapter, in Ref. to the date of portfolio formation (for the first portfolio) or rebalancing (for portfolios thereafter), let us call the preceding 252 days the "formation period" and the subsequent quarter, typically 58–60 days, the "holding quarter." The sequence of consecutive holding quarters is termed the "holding period" in this chapter which is the entire investing time horizon of our experiment.

3. Min k-cut for asset selection

The risk-based portfolio strategies EW, MV, and RP described in the previous section provide substantial return–risk tradeoff advantage compared to M [3]. Moreover, the qualitative feature that is important about these strategies is that each achieves the maximum of the associated diversification profile measure. Building on top of these risk-based strategies, we present an improvement in two directions in this chapter. Firstly, we consider the size of the risk-based portfolio that is completely determined by the strategy that defines it. For example, strategies EW and RP, by definition, generate portfolios whose size is equal to that of the

assets pool which is the universe of all investable assets, specifically, N. Therefore, to accu-rately implement EW and RP, the investor has to hold all assets that exist in the investable universe (normally, in the order of hundreds) in the portfolio. This may be too difficult to achieve in practice, and furthermore, holding all assets in the investable universe may not be an investor's idea of a portfolio. Thus in reality, the implementation of EW and RP is sometimes vastly approximated by various heuristic approaches created by the investor. In this chapter, we present a systematic way to reduce and control the size of the portfolios generated by these strategies. This characteristic of our proposed strategy may be very beneficial from the practical point of view. Strategy MV, on the other hand, normally gener-ates a very concentrated set of assets whose size may be larger or, typically, smaller than an investor's preferred value. In the case of the former, our contribution will provide a system-atic way to reduce the portfolio size for this strategy. In the case of the latter, our proposed strategy may have a slight adverse effect in this respect. As a side note, for the MV strategy, one may add cardinality constraint to Eq. (2) to match the size of the portfolio with investor's investment constraints; however, this will result in a mixed integer quadratic programming that is proven to be computationally hard.

Secondly, we address the possibility of performance enhancement of the risk-based portfolio strategies. Specifically, we have witnessed that risk-based strategies show, sometimes serious, weakness in some of the diversification profile measures. As this phenomenon is attributed to the disregard for one measure of diversification by a strategy that optimizes a different measure of diversification, we consider an assets pool that is well diversified on which the risk-based strategies are executed. For this matter, we present a systematic way to execute assets selection from the pool of all investable assets such that its effect will be an improved performance across all diversification profile measures and return-risk tradeoffs.

Relating the above two directions of improvement, in this chapter, we present an assets selection method that realizes the two improvements simultaneously. To describe our method, consider the correlation matrix $R = \left(\rho_{ij}\right)_{i,j=1}^{N}$ of the set of all investable assets, where $\rho_{ij}\left(\overset{\Delta}{=} \frac{\sigma_{ij}}{\sigma_i \sigma_j}\right)$ is the correlation coefficient between assets i and j. To make the matrix to have nonnegative entries, let $\widetilde{\rho}_{ij}\overset{\Delta}{=}e^{\rho_{ij}}$ for all $i,j = 1, \cdots, N$, and let us consider the new matrix $\widetilde{R} = \left(\widetilde{\rho}_{ij}\right)_{i,j=1}^{N}$. Now, form the weighted graph G whose adjacency matrix is $\widetilde{R}$. This complete graph G has the property that more (less) correlated pair of assets has higher (lower) weighted edge. Therefore, to obtain a set of k ($\ll N$) assets that are least correlated with each other, one may partition the graph G into k-connected subgraphs so that the edges removed to obtain the partition has a minimum weight and then pick an asset in each partition according to some rule. This approach to obtaining a set of k assets with the described property complements the main objective of some of the risk-based portfolios. Specifically, recall that the main objectives of EW and RP are the maximizations of weight and risk diversifications, respectively, with perfect disregard for other measures. This suggests the implication of the presence of, possibly highly, correlated assets in the obtained portfolios. Thus, for these strategies, the preselection of assets seems beneficial. On the other hand, the main objective of MV is the maximization of volatility reduction in which lesser correlated assets are selected to some degree. Therefore for this

strategy, the benefit of the preselection of assets with the described property seems not to be as large as in the other cases.

In all cases, by reducing the assets pool from the universe of all investable assets of size N to a set of k-diversified assets with respect to correlation with other assets, we have effectively executed a *diversified assets selection.* Therefore, it remains to describe (1) how to partition the graph into k-connected subgraphs, satisfying the constraints mentioned above, and (2) how to pick an asset in each partition.

The first part of this is precisely the minimum k-cut problem of which finding for the exact solution is well known to be NP-hard [10]. To this end, we use an efficient approximation algorithm to this problem that finds a minimum k-cut within a factor of $2\left(1 - \frac{1}{k}\right)$ of the optimal due to [11], which is as follows:

Min k-cut approximation algorithm [11]:

1. For each edge $\widetilde{\rho}_{ij}$, pick a minimum weight cut that separates the end points of $\widetilde{\rho}_{ij}$.

2. Sort these cuts by increasing weight, obtaining the list $\rho'_1, \rho'_2, \cdots, \rho'_{N(N-1)/2}$.

3. Greedily, pick cuts from this list until their union is a k-cut; cut ρ'_i is picked only if it is not contained in $\rho'_1 \cup \cdots \cup \rho'_{i-1}$.

We note that the factor of $2\left(1 - \frac{1}{k}\right)$ of the optimal is still known as the best approximation factor for tractable algorithms for the minimum k-cut problem [12]. The complexity of this algorithm is dominated by that of finding the cuts $\rho'_1, \rho'_2, \cdots, \rho'_{N(N-1)/2}$ which can efficiently be calculated through the use of Gomory-Hu tree representation. Specifically, $N - 1$ max flow computations suffice to implement the above Min k-Cut Approximation Algorithm. Moreover, using the Gomory-Hu tree representation, all partitions but one in the k-cut contain exactly one vertex each with the remaining $N - k + 1$ vertices being contained in the last partition. This characteristic of the algorithm when used with Gomory-Hu tree representation almost eliminates the need for the second part of our diversified assets selection as we need to pick only one vertex in the only partition that contains more than one vertex. Nevertheless, it remains to describe how to pick an asset in this last partition with more than one asset. To this end, we define the affinity of asset i, $a(i)$, as

$$a(i) = \sum_{j \neq i} \widetilde{\rho}_{ij} \tag{8}$$

from the matrix $\tilde{R}$. Therefore, the affinity of asset i gives a measure of how the asset i is correlated with all other assets. Eq. (8) shows that the larger the value of $a(i)$, the more correlated the asset i is with other assets. To pick the one vertex in the last partition with $N - k + 1$ vertices, we picked the vertex i with the highest $a(i)$ in the partition as this vertex would appropriately serve as the "representative" of this partition. To see the effect of using affinity as the criterion for asset selection, we also tried picking the one vertex with the smallest $a(i)$ in the partition. We labeled the former asset selection method as "Max," and the latter one as "Min." We note that this second part adds negligible computational burden on the overall

diversified assets selection process as the correlation matrix R and thereby $\tilde{R}$ is easily obtained from the covariance matrix Σ.

In the execution of the overall asset-selected portfolio selection, this part of asset selection is executed first and then the risk-based portfolio strategy is conducted. We have presented an "efficient" method to obtain a reduction in the size of the assets pool to any number of choices in this section. In the next two sections, we empirically demonstrate that this method also proves to be very "effective," characterized by superior return–risk tradeoff performances compared to the baseline risk-based portfolio strategies.

4. Asset-selected risk-based portfolio strategies

In this section, we formally present our proposed asset-selected risk-based portfolio strategies. As mentioned before, the purpose of our proposed asset selection-based strategy is twofold. The first is endowing the investor the option to choose the exact size of the portfolio when the risk-based strategy is either EW or RP. The second purpose is obtaining a superior return-risk tradeoff through effective subset selection of the assets prior to applying the risk-based portfolio strategy. We empirically demonstrate that our proposed strategies can generate returns that are sufficiently higher than the pure, or the baseline, risk-based strategies.

Now, to formally describe the strategies, let us denote the baseline risk-based strategy which does not employ asset selection by S and the asset-selected risk-based strategy by

$$S\_\mu$$

where S is the name of the risk-based portfolio strategy, that is,

$$S \in \{EW, MV, RP\},$$

and μ is the name of the asset selection method. For our asset selection-based strategies, we picked the number of partitions k equal to approximately 25% of the total number of assets N in all of the experiments conducted in this chapter. In addition to the two methods of μ described in the previous section, we also tried selecting the asset with the largest market capitalization in the last partition with $N - k + 1$ assets. We denote this method as "MC," that is,

$$\mu \in \{Max, Min, MC\}.$$

To illustrate the performance of these risk-based portfolio strategies, the next three figures, **Figures 1–3**, exhibit the annualized return versus the annualized volatility plots of the three strategies, respectively, for the S&P 500 (marked by ∘) and the KOSPI 200 (marked by Δ) data. **Figure 1** shows the plots for the EW-based strategies for the two data sets. Any point to the left of and/or above the point of the baseline strategy can be interpreted as improvement over the baseline strategy. We observe that EW_Max, EW_Min, and EW_MC all lie above and/or to the left of EW for both data sets. The defining favorable characteristic of the asset selection-based
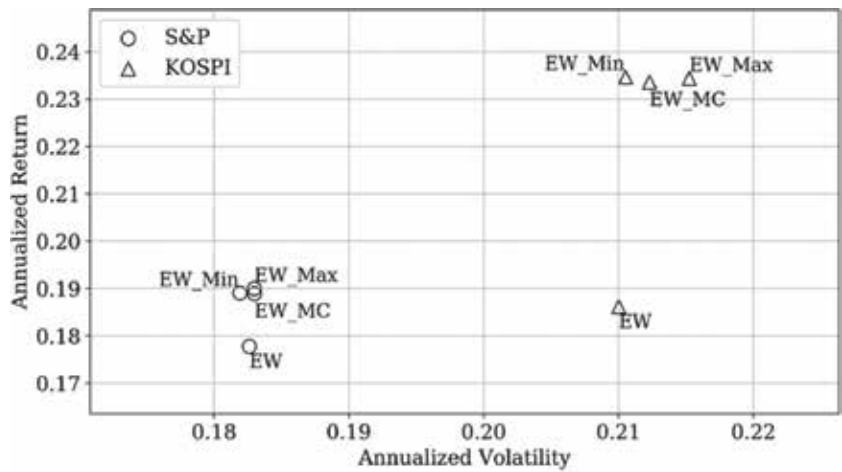
Figure 1. Annualized return versus volatility for EW-based strategies.
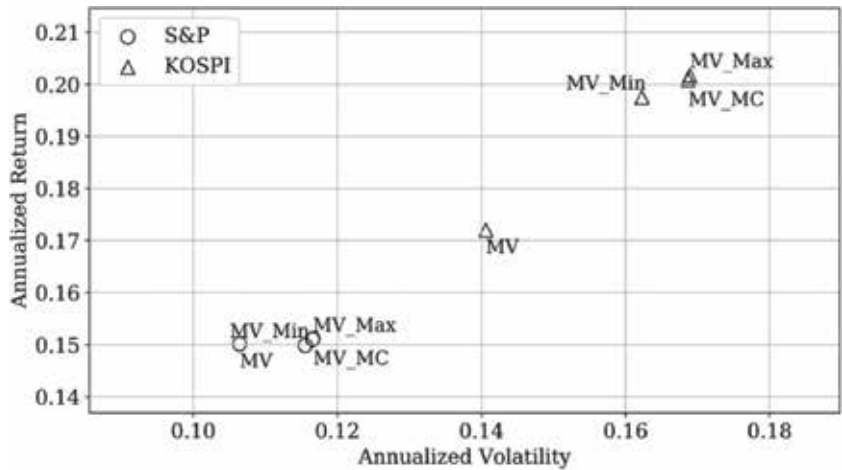
Figure 2. Annualized return versus volatility for MV-based strategies.

strategies that can be deduced from the performances of EW_Max, EW_Min, and EW_MC is that the number of assets in the portfolio can be reduced significantly to any number of investor's choice, facilitating easier portfolio management while, at the same time, improving the return–risk performance!

Figure 2 shows the similar plots for the MV-based strategies. As one can expect, since the baseline strategy achieves the minimum volatility among all strategy types, improvement cannot be and was not made with respect to annualized volatility from the asset selection-based strategies. On the other hand, for the KOSPI 200 data set in particular, asset selection-based strategies were able to produce higher returns than the baseline strategy which may be attributable to the diversified assets selection. **Figure 3** shows the plots for the RP-based
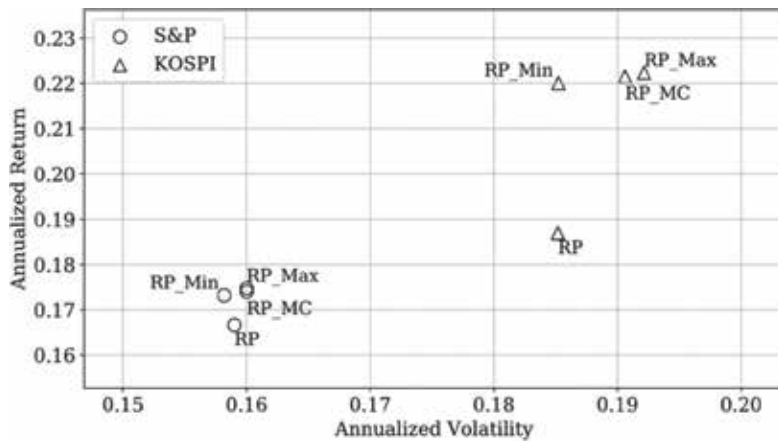
Figure 3. Annualized return versus volatility for RP-based strategies.

	Weight diversification	Volatility reduction	Risk diversification
S&P 500			
EW_Max/Min/MC	1.000	0.349/0.344/0.350	0.835/0.829/0.835
MV_Max/Min/MC	0.131/0.136/0.131	1.000	0.131/0.137/0.131
RP_Max/Min/MC	0.842/0.826/0.843	0.479/0.484/0.480	1.000
KOSPI 200			
EW_Max/Min/MC	1.000	0.456/0.440/0.470	0.863/0.857/0.869
MV_Max/Min/MC	0.253/0.256/0.257	1.000	0.253/0.257/0.257
RP_Max/Min/MC	0.859/0.835/0.864	0.603/0.595/0.612	1.000

Table 2. Diversification profile measure summary for asset selection-based strategies.

strategies, and in particular, the plots resemble those of the EW-based strategies, that is, RP_Max, RP_Min, and RP_MC all lie above and/or to the left of RP. As RP produces the portfolio only with equal-risk exposure to all assets, it is clearly benefitted by making diversified assets selection.

In **Table 2**, we list the values of the three diversification profile measures of S_μ, $S \in \{$ EW, MV, RP $\}$ and $\mu \in \{$ Max, Min, MC $\}$ for the S&P 500 and the KOSPI 200 data analogous to **Table 1**. The diversification profile measures were calculated so that S_μ, $S \in \{$ EW, MV, RP $\}$ achieves 1 for weight diversification, volatility reduction, and risk diversification, respectively, for every $\mu \in \{$ Max, Min, MC $\}$. **Table 2** shows a similar trend with respect to the stronger and weaker profile measures for each of the asset-selected risk-based portfolio strategies as in **Table 1**. However, a noteworthy finding from **Table 2** is that, compared to **Table 1**, the values of the profile measures have increased significantly across all profile measures for all portfolio

strategies. Furthermore, such an improvement is consistent across all asset selection methods and both data sets. This result serves as good evidence that effective subset selection of assets prior to applying the risk-based portfolio strategy improves the "quality" of the risk-based strategies in all of the diversification profile measures considered. In summary, **Figures 1**–**3** and **Tables 1** and **2** indicate that our proposed asset-selected risk-based portfolio strategies provide clear superior return-risk tradeoff and diversification profile measure performances to the baseline risk-based portfolio strategies.

5. Results

In this section, we present a broader set of empirical results for the asset-selected risk-based portfolio strategies presented in this chapter. This allows us to better understand the advantages and the disadvantages of the asset-selected portfolio strategies. For this matter, we examine the following set of performance measures: (A) cumulative return, (B) annual return average, (C) annual return standard deviation, (D) annualized return, (E) annualized volatility, (F) Sharpe ratio, (G) beta, (H) portfolio size, (I) maximum drawdown, and (J) one-way turnover. In this set of performance measures, all return measures are simple returns except the "cumulative return (A)" whose value is set to 1 on the first day of the holding period to explicitly show the increase in the value of the initial asset throughout the entire investing time horizon. The "portfolio size (H)" is the ratio of the size of the portfolio to that of the investable universe which is equal to N. We recall that the performances of all strategies considered in this chapter except the market capitalization-weighted portfolio M reflect survivorship bias of the same degree as all data pertaining to assets were collected on the same day. For the M strategy, we used the index to calculate for the performance measures.

5.1. S&P 500 data

Table 3 shows the results of M, EW-based, MV-based, and RP-based strategies for the S&P 500 data. For the EW-based and the RP-based strategies, our asset-selected strategies showed a clear superior performance in terms of all types of returns considered, a comparable or slightly worse maximum drawdown performance, and an inferior one-way turnover performance to the respective performances of the baseline strategies. It seems that this single drawback of higher one-way turnover for the asset-selected strategies is an intrinsic characteristic of the strategies that stems from the construction and can be viewed as an implementation cost for the improved return-risk tradeoff performances gain. For the MV-based strategies, essentially no improvement was gained through asset selection. This behavior can be attributed to the fact that the baseline MV portfolio is already a somewhat diversified portfolio with respect to covariances between assets, so that adding the asset selection phase in portfolio construction does not help in terms of the performance measures. In fact, for the S&P 500 data, asset selection only contributed to adverse effect in terms of portfolio size, maximum drawdown, and one-way turnover as indicated by the table. Therefore, to improve the MV strategy in a similar order of magnitude as the EW and RP strategies, it seems that different asset selection

	A	B	C	D	E	F	G	H	I	J
M	6.282	0.085	0.159	0.073	0.180	0.243	1.000	—	2.200	—
EW	73.249	0.190	0.171	0.178	0.183	0.816	0.977	1.000	12.885	0.019
EW_Max	95.972	0.203	0.180	0.190	0.183	0.881	0.965	0.252	16.067	0.370
EW_Min	94.036	0.202	0.179	0.189	0.182	0.881	0.957	0.252	15.730	0.368
EW_MC	93.786	0.202	0.179	0.189	0.183	0.875	0.965	0.252	15.786	0.366
MV	39.299	0.156	0.136	0.150	0.106	1.141	0.453	0.106	4.064	0.379
MV_Max	40.084	0.158	0.130	0.151	0.117	1.049	0.489	0.068	5.505	0.573
MV_Min	39.043	0.156	0.130	0.150	0.116	1.049	0.483	0.070	5.575	0.562
MV_MC	40.218	0.158	0.130	0.151	0.117	1.050	0.489	0.068	5.531	0.574
RP	57.126	0.176	0.154	0.167	0.159	0.868	0.840	1.000	9.487	0.072
RP_Max	68.401	0.185	0.159	0.175	0.160	0.912	0.836	0.252	11.168	0.398
RP_Min	66.100	0.183	0.157	0.173	0.158	0.913	0.823	0.252	10.798	0.392
RP_MC	67.305	0.184	0.158	0.174	0.160	0.908	0.836	0.252	11.040	0.394

Table 3. Performance summary for S&P 500 data.

methods need to be explored. We note that in all of the asset-selected risk-based strategies, as $k - 1$ assets have already been selected before selecting the last asset using the asset selection method μ, the choice of the asset selection method seemed not to matter too significantly in terms of the performances as shown in the table.

Next, in **Figure 4**, we show the curves representing the cumulative returns over the entire investing time horizon for the S&P 500 data. For the asset-selected strategies, only the $\mu = $ Max
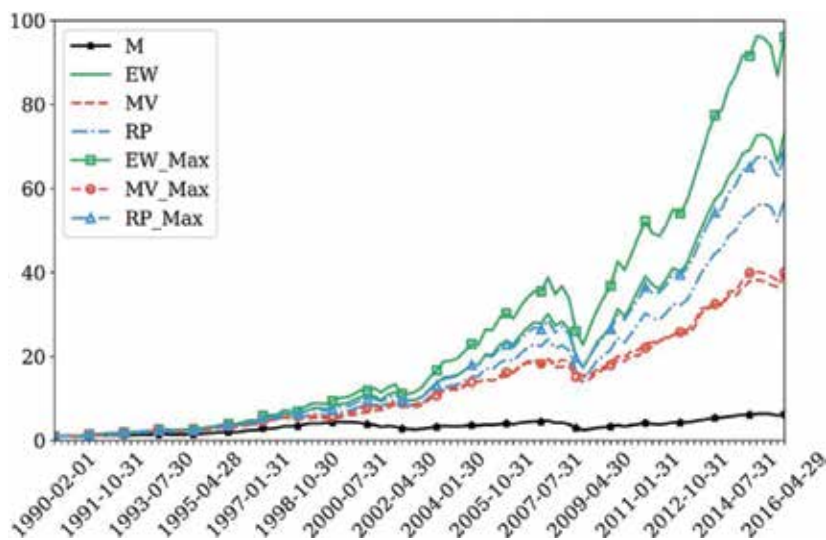
Figure 4. Cumulative return curves for S&P 500 data.

method is shown in the figure as this method was generally the best, albeit marginally, among the three methods. All curves start at 1, and we used the same color to represent the same type of risk-based portfolio strategy. **Figure 4** shows that EW_Max performs the best, and in particular, all asset-selected risk-based strategies outperform their respective baseline strategies, sometimes significantly.
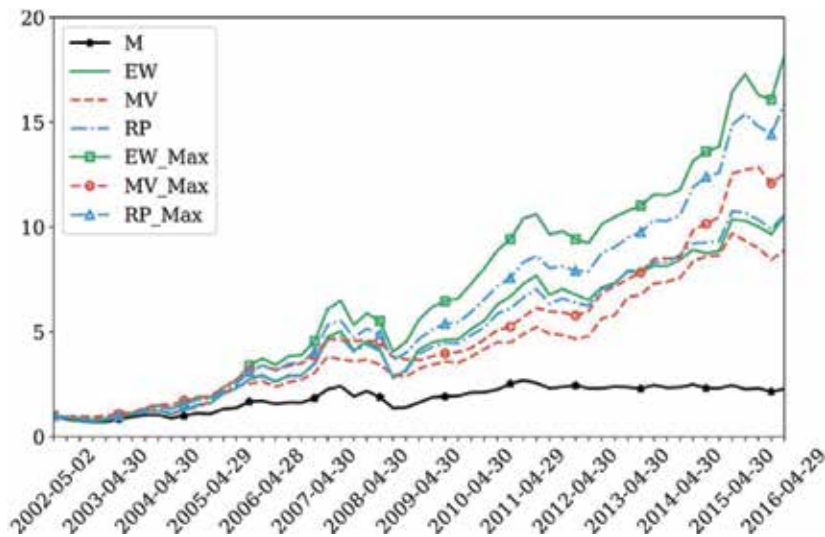
5.2. KOSPI 200 data

Table 4 shows the results of M, EW-based, MV-based, and RP-based strategies for the KOSPI 200 data. Similar to **Table 3**, for the EW-based and the RP-based strategies, our asset-selected strategies showed a clear superior performance in terms of all types of returns considered, in this case including a maximum drawdown as well, and an inferior one-way turnover performance to the respective performances of the baseline strategies. As in **Table 3**, the portfolio sizes are approximately 25% of N which facilitates easy portfolio management in contrast to the baseline strategies. Even for the MV strategy, our asset-selected strategies produced performance improvement across all measures but the one-way turnover. The magnitude of the improvement was not as large as the asset-selected strategies of the EW and RP cases; however, even in this MV case, adding the asset selection phase in portfolio construction facilitated a more comprehensive assets diversification than that obtainable only through variance minimization. As before, the choice of the asset selection method seemed not to matter too significantly while $\mu = $ Max method generally outperformed the others. So for the KOSPI 200 data, our proposed asset-selected strategy produced a clear superior performance to the baseline strategy for all strategy types considered in this chapter.

	A	B	C	D	E	F	G	H	I	J
M	2.267	0.084	0.230	0.061	0.229	0.113	1.000	—	1.044	—
EW	10.495	0.216	0.299	0.186	0.210	0.718	0.840	1.000	2.232	0.021
EW_Max	18.156	0.256	0.283	0.234	0.215	0.924	0.840	0.254	2.439	0.396
EW_Min	18.216	0.257	0.283	0.235	0.211	0.946	0.815	0.254	2.463	0.400
EW_MC	18.003	0.254	0.278	0.234	0.212	0.933	0.834	0.254	2.358	0.390
MV	8.897	0.191	0.237	0.172	0.141	0.971	0.399	0.201	1.257	0.354
MV_Max	12.553	0.217	0.211	0.202	0.169	0.984	0.488	0.120	1.039	0.503
MV_Min	11.964	0.216	0.224	0.198	0.162	0.999	0.439	0.122	1.437	0.495
MV_MC	12.437	0.215	0.208	0.201	0.169	0.981	0.490	0.121	1.022	0.502
RP	10.585	0.214	0.280	0.187	0.185	0.818	0.729	1.000	1.891	0.079
RP_Max	15.868	0.241	0.255	0.222	0.192	0.972	0.734	0.254	1.863	0.417
RP_Min	15.466	0.239	0.256	0.220	0.185	0.997	0.694	0.254	1.767	0.419
RP_MC	15.723	0.239	0.251	0.222	0.191	0.976	0.734	0.254	1.790	0.409

Table 4. Performance summary for KOSPI 200 data.

Figure 5. Cumulative return curves for KOSPI 200 data.

Next, in **Figure 5**, we show the curves representing the cumulative returns over the entire investing time horizon for the KOSPI 200 data. As in **Figure 4**, only the $\mu = $ Max method is shown for the asset-selected strategies as this method was the best among the three methods. As before, all curves start at 1, and we used the same color to represent the same type of risk-based portfolio strategy. **Figure 5** shows that EW_Max performs the best followed by RP_Max and then MV_Max. Consequently, all asset-selected risk-based strategies outperform their respective baseline strategies as in **Figure 4**. Summing up, **Figures 4** and **5** describe that our proposed strategy's performance improvement is robust across both data sets which serves as evidence that diversified assets selection contributes to superior portfolio returns.

6. Conclusions

In this chapter, we considered the three types of risk-based portfolio strategies that have played an important role recently in the area of smart beta strategies. They are the equal-weighted, the minimum variance, and the risk parity portfolios. By establishing an efficient and effective asset selection from assets in the investable universe before the risk-based portfolio strategies are applied, improvements in the characteristic and in the performance of the risk-based portfolio strategies were obtained. The improvement in the characteristic part allows the investor to pick the exact size of the portfolio for the equal-weighted and the risk parity portfolios. The improvement in the performance part is related to the performance improvement in all three risk-based portfolio strategies for various performance measures such as the returns, the Sharpe ratio, and the diversification measures. Empirical results on the data sets of the S&P 500 and the KOSPI 200 have indicated that our asset-selected risk-based

portfolio strategies show, sometimes significant, advantages across various performance measures compared to the baseline risk-based strategies.

Acknowledgements

Conflict of interest

The authors declare no conflict of interest.

Author details

Saejoon Kim* and Soong Kim

*Address all correspondence to: saejoon@sogang.ac.kr

Department of Computer Science and Engineering, Sogang University, Seoul, South Korea

References

[1] Markowitz H. Portfolio selection. The Journal of Finance. 1952;**7**:77-91

[2] Sharpe W. Capital asset prices: A theory of market equilibrium under conditions of risk. The Journal of Finance. 1964;**19**(3):425-442

[3] Carvalho R, Xiao L, Moulin P. Demystifying equity risk-based strategies: A simple alpha plus beta description. The Journal of Portfolio Management. 2012:56-70

[4] Clarke R, de SH, Thorley S. Risk parity, maximum diversification, and minimum variance: An analytic perspective. The Journal of Portfolio Management. 2013;**39**(3):39-53

[5] Jurczenko E, Michel T, Teiletche J. Generalized Risk-Based Investing. SSRN; 2013

[6] Richard J-C, Roncalli T. Smart Beta: Managing Diversification of Minimum Variance Portfolios. SSRN; 2015

[7] Maillard S, Roncalli T, Teiletche J. The properties of equally weighted risk contribution portfolios. The Journal of Portfolio Management. 2010;**36**:60-70

[8] Qian E. Risk parity and diversification. The Journal of Investing. 2011;**20**(1):119-127

[9] Bai SX, Tutuncu R. Least-squares approach to risk parity in portfolio selection. Quantitative Finance. 2016;**16**(3):357-376

[10] Goldschmidt O, Hochbaum D. Polynomial algorithm for the *k*-cut problem. IEEE Annual Symposium on Foundations of Computer Science. 1988:444-451

[11] Saran H, Vazirani V. Finding *k* cuts within twice the optimal. SIAM Journal of Computing. 1995;**24**(1):101-108

[12] Vazirani V. Approximation Algorithms. Springer; 2003

Virtual Reality for Urban Sound Design: A Tool for Architects and Urban Planners

Josep Llorca

Abstract

Urban sound is one of the main concerns of architects and urban planners in contemporary cities: how to control it, what to do about noise pollution, where silent areas should be situated, or which urban decisions must be made. These questions, among others, are based on spatial sound. Virtual reality is a powerful technology that can serve as a design tool to find some answers to these questions. Due to its power to generate realistic images of the environments that are studied, it is easy to see that virtual reality could contribute to the visualization and auralization of spaces before their construction. This task is one of architects' responsibilities, and such a tool could be very useful to them. This chapter highlights the principles and some applications of virtual reality in urban sound design.

Keywords: virtual reality, architect, urban planner, sound design, sound object

1. Introduction

Sound continuously surrounds and envelops us, whether we are indoors or out, at work or play, in cities or the country. We hear birds, voices, machines, wind, water, thunder, whispers, steps, calls, whimpers, doors, windows, floors, chairs, etc. Some of these sounds are heard clearly, others overlap. Sometimes they come together in a yell, at other times they succeed one another. Then, you can hear noises, sounds, music, background, rhythms, accelerando, harmonies, dissonances, cacophonies, echoes, vibratos, repetition, melodies, etc. At times sounds remain masked by distance; at other times, some frequencies are highlighted. They are modified by echo, reverberation, coupling, absorption, brightness or localization. They have the capacity to make us rejoice, to sadden, pacify, sweeten, amaze, frighten, annoy, alert,

stress, upset, attack or madden. At the same time, they can make you dance, swing, feel dizzy, sing, whistle, imitate, keep quiet, laugh or cry. When you try to emulate them, you can beat, rub, blow, strum, nip, strike, haul, move or pat something. Some sounds can be anticipated, others come as a surprise. Sometimes, you know how they are going to sound, but you do not know when. At other times you know when you will hear them, but not how they are going to sound. In some cases, the sound has been heard, but you do not know where it has come from. At certain times, it is known that a sound is going to happen, but you do not know why. Thus, anticipation, prevision, effect, surprise, timbre or production techniques contribute to hearing a sound as a source of pleasure, such as when it turns into music.

This chapter describes some attributes of sound object design in architecture and urbanism, and explains a useful tool for designing and experiencing a sound object. The first part of the chapter defines the sound object in the built environment through some considerations of its temporal composition, and the treatment of the city as a sonic instrument. The second part of the chapter translates these concepts into a powerful representation tool—virtual reality—as a useful instrument for architects and urbanists of today, and then reviews the software that meets these needs and describes some case studies. Finally, the last part of the chapter provides an overview of how acoustic virtual reality can take advantage of some current developments in artificial intelligence.

2. Sound objects

2.1. Some features of the architectural and urban sound object

To define what an architectural and urban sound object is, we must refer to the creator of the term "sound object". In 1966, in his famous *Traité des objets musicaux*, Pierre Schaeffer broke with the academic classifications of noise, sound and music, and created a new musicology. His work presented a phenomenology of the audible. The key concept was not defined as a musical object, but as a sound object that could represent any environmental sound. The notion is quite complex and its richness cannot be demonstrated in a few words [1]. Nevertheless, Pierre Schaeffer himself defined what a sound object is, pointing out that:

> *It is obvious that in saying "that's a violin" or "that's a creaking door" we are alluding to the sound produced by the violin or the creak of the door. But the distinction that we want to establish between the instrument and the sound object is even more radical: when we are presented with a magnetic strip in which an unknown sound is recorded, what are we listening to? It is precisely what we call a sound object, regardless of every causal reference designated by the terms sound body, sound source or instrument* [2].

In this way, the term "sound object" is grounded in our subjectivity, despite the fact that it is not modified by individual variations in hearing, or continuous variations in our attention and sensitivity. Far from a subjective issue—in the individualist, incommunicable and practically elusive sense—sound objects can be well described and analyzed [2]. In other words, the sound object is the sound that reaches the listener's ear and is analyzed just before entering it. The sound object is never well revealed, as in the effect and content of blind hearing.

Pierre Schaeffer defines sound objects by comparing illuminating and sonic phenomena. This comparison is reproduced here, as it provides a clear explanation:

Two big differences separate the experience of illuminating and sonic phenomena. The first consists of the fact that most visual objects are not sources of light, but simply objects, in the usual sense of the word, with light shining on them. Physicists are therefore quite accustomed to distinguishing light from the objects that reflect it. If the object itself gives out light, then we say it is a light "source".

With sound there is nothing like this. In the overwhelming majority of sonic phenomena, sound as originating from "sources" is emphasized. However, the classic distinction in optics between sources and objects has not been imposed in acoustics. Attention has been given to the sound (as we say the light) considered as an emanation from a source, its paths and deformations, without the appreciation of the shapes and contours of this sound apart from the reference to its source [2].

It can be easily assumed that the distinction between sound source and deformations of the sound source can refer to a probable distinction between sound source and modifications to the sound added by the architecture. In other words, we can refer to the distinction between music and architecture in the process of hearing. However, before a conclusion is drawn, another quote by Schaeffer could provide clarification:

This attitude has been reinforced by the fact that sound (prior to the discovery of recording) has always been linked in time with the energy phenomenon that was its origin, to the extent that it has been practically confused with it. However, a fleeting sound is only accessible to one sense and remains under single control: the sense of hearing. In contrast, a visual object has something stable, and this is the second of the aforementioned differences. It is not confused with the light that illuminates it, it appears with permanent contours under different lights, and it is accessible to other senses: it can be felt, weighed and smelled; there is a form that our hands feel, a surface that touch explores, a weight and an odour.

It is understood that the notion of object barely had the strength to impose itself on the physicist's attention. As the natural tendency of physics is to lead facts to their causes, great satisfaction is found in the energetic evidence of the sound source. There is no reason why the ear, at the end of the propagation of the mechanical relations in an elastic medium (the air), should perceive anything other than the sound source itself.

In fact, there is nothing false in the reasoning. Let us just say that, while it is valid for a physicist or an electroacoustic device builder, it is not, however, suitable for a musician or even for an acoustic ear [or an architect]. In fact, the latter do not concern themselves with the way a sound is born and propagated, but only with the way it is heard. Now, what the ear hears is neither the source nor the "sound", but the true sound objects in the same way that the eye does not see directly the source, or even its "light", but the luminous objects.

What Schaeffer wants us to realize is that what the ear hears is not the source, the "sound", or even the pure music, but the real sound objects, which are the-music-with-the-architecture, in the same way that the eye does not see the source directly, or even its "light", but the illuminated architecture. What the ear hears is, in fact, musicalized architecture.

Notably, when virtual reality tries to represent the sound world of an environment, it always does so for a subjective position of the listener, with an individual and unique point of view. The listener as a receiver perceives the effects and the sound content there, and interactively in virtual reality the sound is heard once it has been mixed with the architecture. Therefore, the listening point of the listener includes not only the sound source, but also the modifications caused by the place in which the sound moves, that is, the sound object. In this context, the notion of sound object provides an appropriate theoretical framework for this type of representation, since the sound object is revealed in the blind listening of the effects and the sound content, as explained previously [2].

2.2. The city as a sound instrument and the architect as a luthier of the city

Another issue in considerations of the sound object is the notion of identification of the sound source. While it is true that a sound sounds different in each architectural or urban space, we can also say that we continue to recognize the original sound and distinguish it from other sounds, despite the effects with which the architectural space modifies the sound source. To understand this issue, it can be explained as follows.

We can say that an acoustic musical instrument has three elements, of which the first two are essential. These are: the *vibrator* that starts to vibrate, and the *exciter* that causes the initial vibration or prolongs it in the case of maintained sounds. The third element, which is accessory, but always present, is the *resonator*, that is, a device designed to add its own effects to those of the body in vibration in order to amplify them, prolong them or modify them in some way [2].

If we transfer this concept to the city, we can also distinguish three elements: the *vibrator* or the elements of the construction: stony floors, street furniture, walls, etc.; the *exciter* or the man with his footsteps, movements and actions, the wind, the water; and the *resonator* or architecture. From this point of view, we consider architecture as a large-scale instrument. This classification into three elements presents the city as a great sound instrument and the architect as a manufacturer of musical instruments without knowing it.

Thus, we can easily compare a church, a street and a square. All of them have elements of vibration: the stone walls of the church, the asphalt of the street, and the sheet of water on the fountain of the square. The exciter for the church are the people who pray, while in the street it is the friction of the car wheels, and in the square the jet of water that falls from the top of the fountain. Finally, the first two spaces have an open-air resonator, while the church is a closed space of resonance.

This classification introduces great clarity in the approach, so that we can move on to another more difficult classification: that of the sound objects themselves, obtained from sources or sound bodies. The murmur of the people praying in the church is infinitely closer to the sound of the fountain in the square than to the shrill sound of a shout in the church, which in turn can approach the braking of a car on the street.

Once a sound source has been discovered, two possibilities are offered to the instrument manufacturer: repeat the same source and multiply it in different measures, or keep the same source and try to vary it. Schaeffer argues that the second procedure is not the simplest, because it inextricably links the three elements: vibrator, exciter and resonator. It is likely that contingencies force the instrumentalist not to use these variations in mutual independence, but to associate them immediately with the level of esthetics of the object [2].

However, if we refer to multiple instruments composed of a collection of vibrating bodies, like the city, we see at once that each of them repeats the triple combination of the elements. Architecture proposes a change in the collection of vibrators (each piece of architecture has different construction elements), a change in the collection of exciters (people change), and a change in the resonator (although the architecture does not move, there are as many pieces of

architecture as places; due to the movement of the spectator, the architectural scene changes). Therefore, it is an instrument that varies greatly.

If we listen carefully to the same sound—an oboe—from different points of the same architecture, there is hardly anything in common between the various results produced by the same reproduction-resonance device. But this does not prevent the musician from speaking about the "timbre" of the oboe as an identity. Certainly, the oboe's timbre is recognizable, and the most disfigured of the halls allows the oboe to be identified by an uneducated listener. So, we can state a priori that, although the entire room has a timbre, each of its spatial positions also has its own timbre: the same word with two different meanings. Therefore, we define architecture as an acoustic instrument:

> *Any device that allows us to obtain a varied collection of sound objects or varied sound objects, maintaining in spirit the presence of a cause, is an instrument of music in the traditional sense of experience common to all civilizations* [2].

## 3. Virtual reality as a tool for acoustic design

Virtual reality is the simultaneous representation and perception of reality and its physical attributes in an interactive computer-generated environment [3]. One of these physical attributes is sound. The following diagram shows the operation of acoustic virtual reality applied to a closed architectural environment. Three elements can be distinguished that act in the process: the emission of a sound source, the addition of sound effects, and the perception of the final sound object (**Figure 1**).

The process described above is based on auralization. Following the concepts of simulation in acoustics and vibrations, Vorländer, in [3], describes auralization as (a) the separation of the process of sound generation, propagation and reproduction into three separate blocks, and (b) the representation of these blocks with systems theory tools (**Figure 2**).

The primitive signal, $s(n)$, is called "dry recording". It contains the mono sound signal without any reverberation. Normally it is a sound source recorded at a distance and in a specific direction in an anechoic chamber. The resulting sound signal after the sound propagation in (or between) rooms, $g(n)$, contains characteristics of both the sound source and the transmission system. Here, the propagation of sound in a room usually adds the phenomenon of reverberation to the source signal, while a sound event transmitted through the walls is characterized by a lower sound pressure and a dark sound (with the characteristics of a low pass). The operation of a sound transmission system in physics is represented by the impulse response of the system, $h(n)$. The sound signal at the position of the receiver is achieved by convolution operation of the original "dry recording" with the impulse response (usually represented by a digital filter) [3]. This method can be easily understood as an acoustic filter that contains the impulse response as a function of the position in the hall [4].

In the framework of a tool for the architect and urban planner, we need to summarize which requirements the acoustic virtual reality must meet to satisfy the design needs of architects and urban planners:

1. The tool must be able to correctly represent the location of the sound sources.

2. It must allow the attenuation of sound with distance.

3. It must be able to include the effects added to the source by the geometry of the built environment. To do this, you must calculate the bounces with the adjacent geometry.

4. It must change the resulting sound depending on the materials of the building elements.

5. It must be interactive, that is, allow movement through the environment, and even modification and testing in real time of the elements in the scenario.
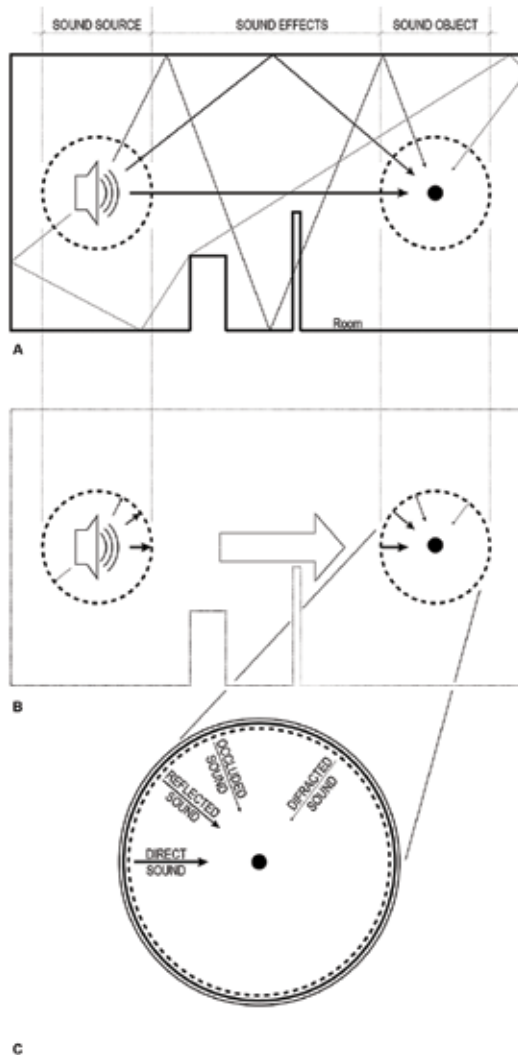
Figure 1. Representation of the operation of acoustic virtual reality applied to a closed architectural environment. A: real environment. B: virtual environment. C: zoom into the sound object.
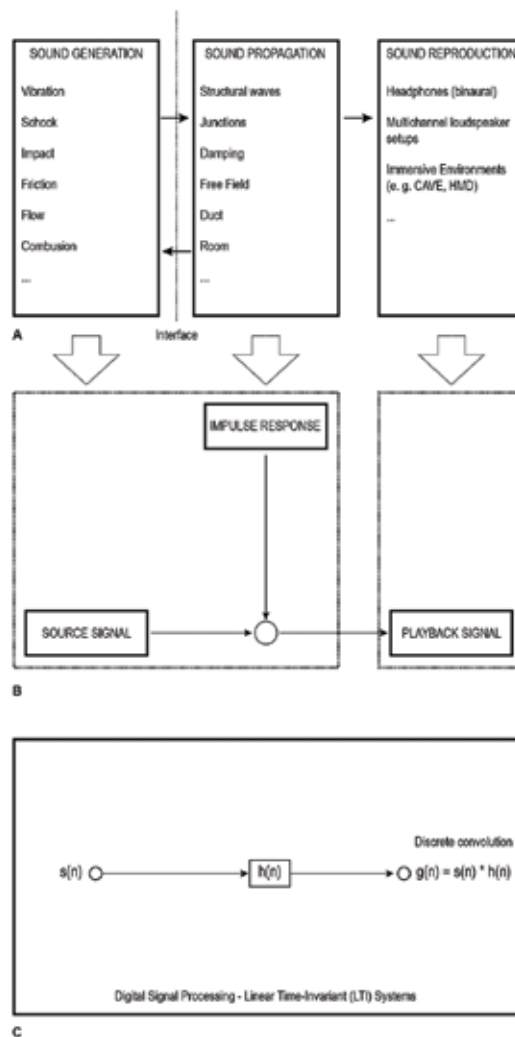
Figure 2. Representation of the operation of auralization. Redrawn from [3]. Generation and propagation of sound and its representation in the physical domain (A and B), and in the domain of acoustic signal processing (C). In the physical domain, sound source characterization and wave propagation can be either modeled or measured. The components will be combined in a synthesis of source signals and impulse responses.

3.1. Virtual acoustics software: a quick review

The market offers a series of virtual reality tools that address virtual acoustics. These can be divided into two types: auralization engines for computer games, and acoustic simulation programs.

Regarding the first type, the sound of a computer game is the result of work done by the sound designer. A sound designer usually creates audio content (sound effects and music)

and then creates sound events to launch the audio content. Sound events are normally monitored using tools such as Wwise or FMOD and are launched by them in the game.

Audiokinetic Wwise (https://www.audiokinetic.com/) is a solution for sound design in computer games that consists of a powerful application to create audio and animation structures; define the propagation; control the sound, music and integration of movement; profile the reproduction; and create banks of sounds. In addition, Wwise is a sophisticated audio engine that handles audio processing, animation and a series of functions optimized for each platform. The program interprets LUA scripts and reproduces exactly how the sounds and the animation behave in the game, allowing the validation of specific behaviors and outlining the Wwise acuity in each platform before the integration of sound into the game. It also contains a series of plug-ins that are divided into those that serve to generate audio and movement, such as the tone generator, and those that create audio effects, such as reverberation. Finally, it is an interface between Wwise and the visualization programs of the three-dimensional world.

FMOD Studio (https://www.fmod.com/), like Wwise, this is an application dedicated to sound in video games. This software, developed by Fireflight Technologies in 2002, is one of the industry's standards and has been the basis of award-winning projects. FMOD Studio is a flexible and intuitive solution for audio in video games. It allows sound in the game to be designed using a DAW interface without knowledge of the required programming. FMOD can be incorporated into almost all platforms. In addition, the program allows real-time mixes and balances. This aspect allows you to make changes and listen to them without having to re-record the scene. FMOD consists of several separate tools in different programs such as "FMOD Designer" that correspond to the main window of the program, where the main work of creating the events and the parameters to be called by the video games is done [5].

Pure Data (https://puredata.info/) is a visual open source environment that works from any personal computer to smartphones and iOS. It is one of the major branches of the family of programming languages known as Max, originally developed by Miller Puckette at IRCAM. Pure Data allows musicians, visual artists, researchers and developers to create programs graphically without writing a line of code. It can be used to process and generate sound, video, 2D/3D graphics, interface and MIDI sensors.

Propagate (https://www.assetstore.unity3d.com/en/#!/content/40200) is a system for Unity that allows you to incorporate immersive audio that propagates realistically through the geometry of the scene quickly and efficiently. It is a simple interface that allows you to propagate the audio in real time, even when the sound sources move. The program is based on three principles of sound reception: the occlusion system simulates the transmission of sound waves by paraments and geometries, taking into account their materials and thus modifying the volume and frequency distribution. The diffraction system simulates the passage of sound waves through holes and corners between geometries. The perception system simulates how the perception of sound changes with the position in geometry.

Acoustic simulation programs, the second group of virtual reality programs with acoustics treatment, are programs that rely on and are completely dedicated to the faithful reproduction of sound in a space. Usually, they do not pay attention to the visual aspect of space, so

their three-dimensional representations are not realistic, but rather schematic. Here we collect three powerful samples: RAVEN, EVERTims and CATT-Walker.

RAVEN. Developed at ITA in RWTH Aachen university, it is based on knowledge about today's acoustic simulation techniques and allows fairly faithful physical auralization of the propagation of sound in complex environments, including important effects such as sound diffusion, room isolation and sound diffraction. Instead of this rendered realistic sound field, the sound sources are distributed and move freely, and the receivers listen to them in real time. In addition, manipulations and modifications of the environment itself are supported. The acoustic simulation of RAVEN combines the method of deterministic image sources (IS) [6] with a stochastic ray-tracing algorithm [7]. This framework allows physical computations of high-quality impulse responses in real time, where, apart from the components of the reflected specular sound field, the phenomenon of sound diffusion, sound transmission, and diffraction are taken into account. The environment is completely written in C ++, supports the operating systems of Windows, Linux and Mac OX X, and allows parallel computing in machines where memory is shared, in machines with memory distributed over the network, or in a combination of both [8].

EVERTims (http://evertims.github.io/) is an open source framework for the auralization of 3D models, which offers real-time feedback on how the acoustics sound in any given room during its creation. The framework is based on three components: a Blender plug-in, a C ++ Ray plotter, and a JUCE auralization motor. While a 3D model in Blender is devised, the plug-in continuously increases the geometry and the details of the materials to the ray tracing. On the basis of this information, the client simulates how the waves propagate there. The result of this simulation is then released to the auralization engine that reconstructs the Ambisonics sound field in any position for binaural listening. The environment takes advantage of the Blender Render Engine to support the auralization of the game for interactive exploration of the designed model [9].

Finally, with the **CATT-Walker** module (https://www.catt.se/walker.htm), CATT-Acoustic software has a powerful tool for real-time auralization in a microcomputer. CATT-Walker manages this dynamic audible restitution by continuously interpolating the impulse responses previously calculated in B format (surround coding). Hearing is performed in binaural mode from appropriate ambisonic decoding. To maintain a sufficiently low latency, CATT uses Lake Technology's split FIR convolutional filtration technique. To reproduce the simulated acoustic environment as closely as possible, the modeled space must be sampled more or less densely by distributing the reception points around the source and in the evolution zone.

4. Five virtual acoustics applications for architects and urban planners

After the description of the properties of Virtual Acoustics and the software on offer that can support virtual acoustic simulation tools, we propose five applications of virtual acoustics in the process of design and architectural and urban analysis.

4.1. Invisible sound objects

The first case is for the phenomenon of sound objects that are heard, but you cannot see them (incongruence), the visible objects that you see, but cannot hear (incongruence), or the sound objects that you see in a different place to which they are heard (delocalization). Do such phenomena generate any design problems?

In contrast to the experimental conditions, the listener immersed in a real environment relies on all the senses to structure a representation of the environment [10]. A sensory modality could also pay attention to a different modality, and even influence the same perception more strongly. This raises questions about whether the resources of attention are controlled by a supramodal system or by various modalities of attention systems. In conditions of focused attention, it is difficult to judge each signal (sound and vision) separately when incongruent signals occur in the same place, at least much more difficult than when the incongruent signals come from different points and attention is divided [11]. The most feasible model for today's knowledge consists of a multilevel attention mechanism with a multimodal component above the sensory component. In the context of perception of the sound environment, this could be interpreted as a stronger emphasis on visible sources, but at the same time, a lower probability of identification of the deviant sounds if these sounds come from the same place as the visual stimulus.

The mechanisms of multisensory attention also have a very strong temporal component. The sound stimuli presented in temporal congruence with the appearance of the visual objective make the visual objective stand out in the scene [12]. Based on this knowledge of multisensory perception, one can solve a nineteenth-century concern of sound landscape designers, at least partially: is it good to hide the sources of unwanted sounds from view? From the perspective of attention, we can conclude that when the sound is not very prominent and therefore does not attract much attention, we can avoid noticing the sound by eliminating congruent visual stimuli. Similarly, a desired sound should be accompanied by a visual stimulus to ensure that it receives sufficient attention. In contrast, we must emphasize that in the case of very prominent sounds that will attract attention, the absence of a visual stimulus could appear as a surprise, which would influence the perception [13].

In this context, the virtual reality tool offers a scenario to test the congruence between the sounds that are seen and heard. For this, the possible options to be evaluated must be simulated:

- A not very prominent and seen sound source: there is weak incongruence.

- A not very prominent and not seen sound source: there is weak congruence.

- A very prominent and not seen sound source: there is strong incongruence.

- A very prominent and seen sound source: there is strong congruence, and control of the situation.

- A desired sound source and view: there is strong congruence, and control of the situation.

• A desired and unseen sound source: there is strong inconsistency, discomfort and lack of control of the situation.

• An unwanted and seen sound source: there is strong congruence, annoyance and control of the situation.

• An unwanted and unseen sound source: there is strong inconsistency, discomfort and lack of control of the situation.

4.2. The influence of materials

Architects and urban planners are constantly concerned about the visual impact that the finishing will have on the built work. Although the visual impact affects the perception of space [14] and can affect the mood of the users [14, 15], the sound impact has a no less important effect on the perception of space [16, 17] and the users' mood [18].

We also know that visual perception in virtual reality environments is related to the geometric representation of the space, and to the material representation of this geometry. If we focus on the representation of the materials, their treatment and properties are paramount. Therefore, base color, glossiness, roughness, normal or bump and displacement are some of the techniques that virtual reality software has developed to simulate the materials of the represented reality.

Even though this fine detailing could also seem to be necessary for acoustic virtual reality representations, there is a big difference between acoustic and visual materials. Small details in visual perception are negligible in acoustic perception. For this reason, a plane wall behaves acoustically similarly to a rough wall at low frequencies [19]. However, some phenomena linked to materiality, such as acoustic porosity and absorption, are linked to visual glossiness and roughness and, therefore, could have a big influence both visually and acoustically. In this context, the fine representation of both visual and acoustical properties of materials in virtual reality would turn this tool into a convincing way of representing the environments designed by architects and urban planners.

4.3. Concavity and convexity in architectural forms

There has always been a debate among architects about the suitability of curved shapes versus straight forms, and vice-versa. At both extremes, these tendencies can be classified as the purest rationalism, that of rectangular forms, straight lines, repeated and constant rhythms [20]; and the organisms that its formal referents have in nature [21–23]. Experimentally, curvilinear, sinuous, parabolic or circular shapes have been shown to affect neural activity more strongly than rectangular or quadrangular ones [24]. The visual influence of this type of geometry both in architectural interiors and in urban exteriors is also very important in the acoustics of these spaces. The graphical acoustics of ray tracing [25, 26] shows us how concave shapes reinforce sound at a point or in a concentrated area, while convex shapes scatter sound in multiple directions.

This affects the way sound is perceived in interiors and can produce undesired effects in these places. One clear example of this phenomena can be tested when seated at specific points in

Coderch's new building for the School of Architecture of Barcelona. The curved walls concentrate student's whispers in some areas that reinforce the noise perception at these points. The same effect can happen when, in an urban environment, a curved design of walls concentrates the sound rays in one area.

Acoustic virtual reality should consider this effect as a product of design. For this purpose, therefore, a rough approximation of the architectural geometry is not enough. More detail in the representation of these nuances would lead to a more realistic and credible representation of reality with this tool.

4.4. The sound object as a reason for the architectural project

This case study deals with the sound object not as a pretext for comfort in architecture, through congruence with visual, acoustic comfort or a sound impact. The sound object is now a topic of design. The design of the sound object must meet all the requirements that the solution to a specific problem requires, as in any architectural design process. Architectural design is a creative process, but its additional emphasis on the definition phase of the problem [27] places it in a special category. When the term "creative" is used in its most general sense, to describe a process in which an agent (a personal product and in the environment) interacts with the material to form a new synthesis of essential novelty, it embraces the entire design process, but is too general to be particularly useful. However, when used in a more specific sense, relating pure arts and sciences, it usually allows for more self-original and self-motivated input, deriving from sensitive perception in art, and critical observation in science of the selected phenomenon. It is, then, a creative process of *synthesis*, preceded by *analysis* and followed by validation, especially in science, which ends up consisting of a piece of art or a hypothesis or validated theory [28]. This process of synthesis-analysis-synthesis, which summarizes all attitudes of observation of reality, [29] is the same as occurs in the process of auralization: *synthesis* of the sound emitted by the sound source, whose sound waves come from a single point; *analysis* or separation by parts of the different types of waves that bounce, pass or diffract in the elements of the built environment; and *synthesis* that is collected in a single point and that we have qualified as the "sound object". Below, we present a series of examples of acoustic targets, extracted from [13]:

- Moving water or sounds of nature should be the dominant sound heard.

- Only the sound of nature should be heard.

- A specific sound should be clearly audible in some areas.

- Mostly (nonmechanical, nonamplified) sounds made by people should be heard.

- The sounds of people cannot be heard.

- Suitable for hearing unamplified/amplified speech (or music).

- Acoustic sculpture/installation sounds should be clearly audible.

- Sounds conveying a city's vitality should be the dominant sounds heard.

- Sounds that convey the identity of a place should be the dominant sounds heard.

However, the process of designing the sound object, like any architectural design process, requires that the solution to the problem addressed by the sound object is not just the solution that meets the functional requirements. In addition to fulfilling the requirements satisfactorily, it offers the user an integrated solution with the pre-existing environment and has its own entity as a newly created element. Here are some possible guidelines for achieving such goals:

- *A solution integrated with the pre-existing environment*: it may contain vernacular sounds of the area or neighboring territory, replicating or imitating them. It can present contrasting sounds, highlighting those of the pre-existing environment. It can modulate pre-existing sounds, reinforcing some frequencies or switching off others.

- *A solution that has its own entity as a newly created element*: it can present a characteristic rhythm, whether regular or irregular. It can present a characteristic tonic, either monophonic or polyphonic. It can present characteristic timbres or textures that are produced by concrete materials.

5. Artificial intelligence and acoustic virtual reality

Even though the framework depicted here seems easy to implement, the reality is rather different. In 2003, Kang et al. highlighted the introduction of new EU noise policies [30] and noted that noise-mapping software/techniques are being widely used in European cities [31]. Nevertheless, they stated that these techniques provide an overall picture for macro-scale urban areas. The micro-scale, for example an urban street or a square, could be more effectively studied using detailed acoustic simulation techniques. In addition, applications that predict and measure micro-scale environments, such as auralization techniques for indoor spaces [32], are still not sufficiently user-friendly, and the computation time is rather long. Kang et al. presented two computer models based on the radiosity and image source methods in an attempt to present to urban designers an interface that could be useful in the design stage, using simple formulae that can estimate sound propagation in micro-scale urban areas. However, the answer still has not been found.

Artificial intelligence can be applied to virtual reality in many ways. In the user interface, we need systems that behave rationally, e.g. reflect user movements as accurately as possible. Content production needs tools for optimizing the layout of virtual worlds, and virtual world simulation needs methods for approximating the behavior of the environment [33]. This last challenge relates strongly acoustic virtual reality and artificial intelligence and might be the key factor in progress in the simulation of both closed spaces and open environments. Some studies have already included artificial intelligence in the field of acoustics. In particular, new methods identify room acoustic properties based on evolutionary algorithms (EA) [34]. Focused on the problem of learning from real acoustics, Cox et al. [35] developed a new method employing machine learning techniques and a modified low frequency envelop spectrum estimator, to estimate important room acoustic parameters including Reverberation Time (RT) and Early Decay Time (EDT) from received music signals. What is known as the machine audition field [36] therefore presents a promising method that can establish and enhance classical methods of acoustics. More specifically, architects and urban

planners always need comprehensive visualizations of the reality that they are designing. Approximation of the behavior of the environment on which they are working is therefore a major concern in their representations. For this reason, artificial intelligence could be a good way to solve some problems that still remain today. These questions draw possible future paths for investigation that we can sum up in the two following points:

- The wide variety of case studies that an architect deals with in everyday practice require an easy method for acoustic representation. Otherwise, decisions are taken by approximation only, or the effort of representation is too great for one studio. For this reason, a database of urban spaces with their defined acoustic features would be useful. The acoustic features could be used as the hidden layer in a neural network framework for rapid prediction of the behavior of the environment in future modifications.

- Measurements of acoustic properties of an architectural space need highly specialized equipment and special conditions. Not every architectural studio has the opportunity or means for making such measurements. Investigations to extract acoustic information from everyday recordings could help not only to analyze the current environment, but also to predict new architectural designs.

6. Conclusions

Acoustic architectural and urban design is an area that still needs to be addressed. Despite the fact that a considerable amount of research has been done on architectural acoustics, urban soundscape, and noise treatment, few real applications of these theories can be found in built projects. This chapter links Schaeffer's theory on sound objects to acoustic virtual reality, to describe a potential way of understanding the role of acoustic virtual reality in the field of architecture and urbanism. The main tools that could help practitioners have been presented here. Moreover, five specific applications have been proposed. However, much more work is needed to apply the theory to daily practice, and all efforts at bringing scientific research into everyday activity are welcome. The direction of artificial intelligence seems plausible in the future. Sound design is an old concern, and still remains as elusive and unpredictable today as it can be when it turns into music.

Acknowledgements

Author details

Josep Llorca

Address all correspondence to: josep.llorca@upc.edu

AR and M, Barcelona School of Architecture, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

References

[1] Augoyard JF, Torgue H, Sonic Experience: A Guide to Everyday Sounds. Montreal: McGill-Queen's University Press; 2010

[2] Schaeffer P. Treatise on Musical Objects: An Essay Across Disciplines. 1966th ed. Oakland: Univertisy of California Press; 2017

[3] Vorländer M, Schröder D, Pelzer S, Wefers F. Virtual reality for architectural acoustics. Journal of Building Performance Simulation. 2014;**8**(1):15-25

[4] Llorca J, Redondo E, Valls F, Fonseca D, Villagrasa S. Acoustic Filter. Cham: Springer; 2017. pp. 22-33

[5] Rehren C, Cárdenas J. Motores de Audio para Video Juegos. Síntesis Tecnológica. 2011; **4**:81-99

[6] Schröder D, Lentz T. Real-time processing of image sources using binary space partitioning. Journal of the Audio Engineering Society. 2006;**54**(7/8):604-619

[7] Schröder D, Dross P, Vorländer M. A fast reverberation estimator for virtual environments—RWTH Aachen University Institute of Technical Acoustics—English. In: Proceedings of the AES 30th International Conference on Intelligent Audio Environments. Saariselkä, Finland; New York, NY: Audio Engineering Society; March 15-17, 2007

[8] Dirk Schröder, Michael Vorländer. RAVEN: A real-time framework for the auralization of interactive virtual environments. In: Conference: Proceedings of the Forum Acusticum, Aalborg, Denmark Volume. January 2011. ISBN: 978-84-694-1520-7

[9] Poirier-Quinot D, Noisternig M, Katz BFG. EVERTims: Open source framework for realtime auralization in architectural acoustics and virtual reality. In: 20th International Conference on Digital Audio Effects (DAFx-17), Edinburgh, United Kingdom; Sep 2017

[10] Driver J, Spence C. Attention and the crossmodal construction of space. Trends in Cognitive Sciences. 1998;**2**(7):254-262

[11] Santangelo V, Fagioli S, Macaluso E. The costs of monitoring simultaneously two sensory modalities decrease when dividing attention in space. NeuroImage. 2010;**49**(3): 2717-2727

[12] Talsma D, Senkowski D, Soto-Faraco S, Woldorff MG. The multifaceted interplay between attention and multisensory integration. Trends in Cognitive Sciences. 2010;**14**(9):400-410

[13] Kang J, Schulte-Fortkamp B. Soundscape and the Built Environment. Boca Raton: CRC Press. Taylor & Francis Group; 2016

[14] Filbrich L, Alamia A, Blandiaux S, Burns S, Legrain V. Shaping visual space perception through bodily sensations: Testing the impact of nociceptive stimuli on visual perception in peripersonal space with temporal order judgments. PLoS One. 2017;**12**(8):e0182634

[15] Vartanian O et al. Impact of contour on aesthetic judgments and approach-avoidance decisions in architecture. Proceedings of the National Academy of Sciences. 2013;**110**(Supplement_2):10446-10453

[16] Al-barrak L, Kanjo E, Younis EMG. NeuroPlace: Categorizing urban places according to mental states. PLoS One. 2017;**12**(9):e0183890

[17] Plack CJ. The Sense of Hearing. 2nd ed. London: Taylor & Francis Group; 2014

[18] Zhang Y, Kang J, Kang J. Effects of soundscape on the environmental restoration in urban natural environments. Noise & Health. 2017;**19**(87):65-72

[19] Kuttruff H. Acoustics. New York: Talyor & Francis; 2004

[20] Frampton K. Modern Architecture: A Critical History. Oxford: Oxford Univeristy Press; 1980

[21] Mumford M. Form follows nature: The origins of American organic architecture. Journal of Architectural Education. 1989;**42**(3):26-37

[22] Krause LR. Frank Lloyd Wright: Organic architecture for the 21st century. Journal of Architectural Education. 2011;**65**(1):82-84

[23] Dennis JM, Wenneker LB. Ornamentation and the organic architecture of Frank Lloyd Wright. Art Journal. 1965;**25**(1):2-14

[24] Banaei M, Hatami J, Yazdanfar A, Gramann K. Walking through architectural spaces: The impact of interior forms on human brain dynamics. Frontiers in Human Neuroscience. 2017;**11**:477

[25] Officer CB. Introduction to the Theory of Sound Transmission: With Application to the Ocean/C.B. Officer.—Version Details—Trove. New York: McGraw-Hill; 1958

[26] Kinsler LE. Fundamentals of acoustics. New York: Wiley; 2000

[27] Kneller GF. The Art and Science of Creativity. Holt: Rinehart and Winston; 1965

[28] Herbert G. The architectural design process. British Journal of Aesthetics. 1966;**6**(2):152

[29] Condillac. La lógica o los primeros elementos del arte de pensar. Imprenta d. Barcelona; 1827

[30] EUR-Lex—32002L0049—GA—EUR-Lex. [Online]. Available from: http://eur-lex.europa. eu/legal-content/GA/TXT/?qid=1399875039336&uri=CELEX%3A32002L0049. [Accessed: November 09, 2017]

[31] Welcome to Schal. [Online]. Available from: http://www.tpsconsult.co.uk/schal.aspx. [Accessed: November 09, 2017]

[32] Jing Y, Xiang N. A modified diffusion equation for room-acoustic predication. The Journal of the Acoustical Society of America. 2007;**121**(6):3284-3287

[33] Laukkanen S, Karanta I, Kotovirta V, Markkanen J, Rönkkö J. Adding intelligence to virtual reality. In: ECAI 2004: Proceedings of the 16th European Conference on Artificial Intelligence. Santa Clara, CA, USA: IEEE; 2004. p. 1136

[34] Poteralski A, Szczepanik M, Ptaszny J, Kuś W, Burczyński T. Hybrid artificial immune system in identification of room acoustic properties. Inverse Problem in Science Engineering. 2013;**21**(6):957-967

[35] Kendrick P, Cox TJ, Zhang Y, Chambers JA, Li FF. Room acoustic parameter extraction from music signals. In: 2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings. Toulouse, France: IEEE; Vol. 5. pp. V-801-V-804

[36] Wang W. Machine audition: Principles, Algorithms, and Systems. Hershey: IGI Global; 2011

Blockchain: The Next Breakthrough in the Rapid Progress of AI

Spyros Makridakis, Antonis Polemitis,
George Giaglis and Soula Louca

Additional information is available at the end of the chapter

"Distributed ledgers, also known as blockchains, could be the most consequential development in information technology since the internet. Created to support the Bitcoin digital currency, the blockchain is actually something deeper: A novel solution to the age-old human problem of trust."

Prof. Kevin Werbach, University of Pennsylvania, Wharton School.

"By far the greatest danger of Artificial Intelligence is that people conclude too early that they understand it."

Eliezer Yudkowsky, an AI theorist.

Abstract

Blockchain technologies, once used exclusively for buying and selling bitcoins, have entered the mainstream of computer applications, fundamentally changing the way Internet transactions can be implemented by ascertaining *trust* between unknown parties. In addition, they ensure *immutability* (once information is entered it cannot be modified) and enable *disintermediation* (as trust is assured, no third party is required to verify transactions). These advantages can produce disruptive changes when properly exploited, inspiring a large number of applications. These applications are forming the backbone of what can be called the Internet of Value, bound to bring as significant changes as those brought over the last 20 years by the traditional Internet. This chapter investigates blockchain and the technologies behind it and explains their technological might and outstanding potential, not only for transactions but also as distributed databases. It also discusses its future prospects and the disruptive changes it promises to bring, while also considering the challenges that would need to be overcome for its widespread adoption. Finally, the chapter considers combining blockchain with Artificial Intelligence (AI) and discusses the revolutionary changes that would result by rapidly advancing the AI field.

Keywords: blockchain applications, AI applications, combining blockchain and AI, disruptive technologies, smart contracts, DAO, decentralized storage, IoT, internet of value, decentralized cloud storage, supply chain operations, blockchain/AI startups

1. Introduction

In a large IBM survey recently conducted by top executives on blockchain [1] it was found that one-third of the almost 3000 who participated responded that they are using, or considering adopting blockchain in their business. According to the survey, 8 in 10 of those exploring blockchain are investing either in response to financial shifts in their industry, or for the opportunity to develop entirely new business models. The results of the survey echo a recent article in **Forbes** [2] entitled *"Blockchain As Blockbuster: Still Too Soon To Tell, But Get Ready"*. The proponents of blockchain talk about its great potential capable of creating the same type of fundamental changes as those brought over the last two decades by the traditional Internet. Yet for the majority of people, including the two-thirds of executives in IBM's survey, blockchain remains an elusive concept, with its advantages not well understood by business people, government officials and the general public (the same thing was true with the Internet in the early 1990s). It is important, therefore, to explain blockchain and its unique advantages as well as its possible drawbacks and in particular the revolutionary changes that would result by integrating it with AI.

The purpose of this chapter is to investigate blockchain and the technologies behind it and explain its might and outstanding potential. It consists of three parts. The first part describes blockchain's achievements and expands on its ability to transform peer-to-peer collaboration by, among its other benefits, removing the need for trusted intermediaries. The second part looks at its future prospects, including its utilization as a distributed ledger and the disruptive changes it will bring while also considering the challenges that would need to be overcome, including the fear of hacking and the possible fraud associated with the utilization of the technology. The final part considers combining blockchain and AI and the breakthrough applications that could result from such a marriage. There is also a concluding section summarizing the chapter and suggesting some directions for future work.

2. The uniqueness of blockchain: decentralized, authenticated and immutable information at lower costs

Blockchain is simply a decentralized, or distributed ledger (versus the centralized ones kept by, say, banks to record transactions and keep customer balances) of trustworthy digital records shared by a network of participants. As such, it expands the traditional Internet of information and communications (emails, sending/receiving/searching for information, exchanging files, participating in social media, etc.) to a new category that can be called the "Internet of Value". Such Internet includes sending/receiving money between two parties without the need for financial intermediaries, buying and selling stocks, keeping/issuing certificates, including real estate titles, creating/executing smart contacts, improving supply chains, etc. Blockchain's uniqueness comes from the following capabilities:

- *Trust:* new information can be added only when the majority of computers in the network give their approval after satisfactory proof is provided that the information, which is transmitted cryptographically, is truthful. The authentication of information is done in

short intervals of time and the updated information is stored (appended) to all participating network computers.

- **Immutability and transparency**: information can be appended only to previous data and, once entered, cannot be changed, modified or lost, providing a permanent, incorruptible historical record that stays in the system permanently. Moreover, changes to public blockchains can be seen by all parties in the network thus ensuring transparency.

- **Disintermediation**: the ledger (database) is not maintained by any single person, company or government but by all participating computers located around the world. This means that two parties are able to generate an exchange without the need for a trusted intermediary to authenticate the transactions or verify the records.

- *Lower costs and greater speeds:* lower transaction costs and greater speed are also characteristics of blockchain in a good number of applications by removing the monopolistic power of powerful intermediaries (e.g. banks) or large, centralized industry leaders (e.g. Airbnb).

3. Why blockchain is a disruptive technology

Blockchain provides a fundamental shift from the Internet of information/communications to the Internet of Value. The difference between the two is fundamental. The first disrupted business models in the 2000s and created the likes of Amazon, Google, Facebook, Alibaba as well as Uber and Airbnb. Its disadvantage is that the information transmitted can be copied, thus making it impossible to guarantee its trustworthiness without the approval of an intermediary, for example, a bank verifying that the money being transmitted is available. The biggest advantage of the Internet of Value is the establishment of trust, through the application of blockchain technology, between strangers who can now trust each other. This means assets can be exchanged in an instant and efficient manner **without** intermediaries who are no longer needed as trust is built into the system. Such an advantage of the Internet of Value is bound to cause even more profound changes than those brought by the Internet of information. Trusted peer-to-peer transactions will encourage the formation of decentralized structures, diminishing the monopolistic power of intermediaries such as banks or firms like Uber and Airbnb [3]. This will be done through the creation of new players that would exploit the blockchain-based platforms of decentralized networks with the potential to dramatically narrow the monopolistic power of today's dominant actors, democratizing the global economy and creating a more efficient and sustainable economic system [3].

Blockchain applications started slowly introducing bitcoins after Nakamoto's 2008 paper and were restricted to cryptocurrencies until July 2015 when the Ethereum platform was released, allowing the issuing of smart contracts. At around the same time Estonia started implementing blockchain technologies in its governmental operations, including an ehealth record system that covered any one of its citizens who had ever visited a doctor. Further applications were introduced in 2016 with smart contracts and decentralized autonomous organizations (DAOs) with huge potential thus fundamentally affecting the legal profession and the management of organizations (see below). However, the most significant applications are taking place since

2016 with a large number of startups working on innovative solutions that are going to change the economic landscape [4] and turn blockchain into a momentous technological force.

3.1. Achievements

Apple, Google, Amazon, Facebook, Tencent, Alibaba, Samsung, Netflix, Baidu and Uber (with a combined market surpassing $4.3 trillion at the beginning of 2018) were created by exploiting the advantages provided by the evolving Internet of the late 1990s and the 2000s until now. These eight firms disrupted the economy and business sector by revolutionizing shopping and viewing habits, the search for information and advertising spending, among others, in ways no one could have predicted in the early 1990s when the Internet was introduced. As blockchain holds the potential for equal or even greater disruptions, particularly when combined with AI (see Section 3), revolutionary changes of considerable magnitude covering a wide range of industries and products/services will emerge over the next 20 years and new firms, corresponding to the eight ones mentioned will probably emerge. The great challenge for entrepreneurs is to direct their startups to exploit the emerging blockchain technologies and develop new applications and innovative products/services at affordable prices to better satisfy existing and emerging needs.

Below is a presentation of what we believe are the 10 most important *existing, or soon to be introduced* blockchain applications, highlighting their usage and advantages and mentioning the startups that have been formed to develop and implement them. These applications have been classified in terms of the industries that are being affected and the various applications being pursued. There is no doubt that many more applications will be introduced in the future, some of them becoming successful breakthroughs, in particular when combined with AI algorithms.

4. Industries

4.1. Banking

Blockchain banking applications can reduce costs by as much as $20 billion by eliminating intermediaries and increasing the safety and efficiency of banking transactions [5]. A leading startup in the field is ThoughtMachine that has developed Vault OS, which is run in the cloud, providing a secure, fast and reliable end to end banking system capable of managing users, accounts, savings, loans, mortgages and more sophisticated financial products (see https://www.thoughtmachine.net/). An alternative blockchain banking application is Corda, a distributed ledger platform that is the outcome of over 2 years of intense research and development by the R3 startup and 80 of the world's largest financial institutions. It meets the highest standards of the banking industry, yet it is applicable to any commercial scenario. Using Corda, participants can transact without the need for central authorities creating a world of frictionless commerce (see https://www.corda.net/). According to *Business Insider* [6], practically all major global banks are experimenting with blockchain technology trying to reduce cost and improve safety and operational efficiencies while, at the same time, making sure that they will not be left behind startups utilizing blockchain technologies to dominate the market.

4.2. Payments and money transfers

By avoiding a central authority to verify payments and money transfers, costs can be substantially reduced. At present, there are a good number of services using the technology aimed primarily at those without bank accounts or those looking for important cost savings. Below is a brief description of six blockchain services located in various parts of the world

- **Abra** (USA) is a mobile application allowing person-to-person money transfers. The app can be downloaded from Apple or Google stores.

- **Align Commerce** (USA) is a payment service provider (PSP) allowing businesses to send and receive payments in local currencies.

- **Bitspark** (Hong Kong) is an end to end remittance platform to any of their 100,000 plus locations worldwide.

- **Rebit** (Philippines) is a money transfer service offering significantly lower rates to the many Philippine immigrants working abroad.

- **CoinRip** (Singapore) is a service offering safe and quick money transfer charging a flat rate of 2%.

- **BitPesa** (Africa) is a cheap and safe money transferring service operating in Africa.

4.3. Securities trading

Blockchain technologies aim to reduce costs and speed up trading while also simplifying the settlement process. For these reasons, many stock exchanges are considering introducing blockchain to their operation. The London Stock Exchange, the Australian Securities Exchange and the Tokyo Stock Exchange are already experimenting with blockchain technologies which are expected to be operational in the near future. Banks and financial companies are also exploring blockchain applications for security trading. T-zero (see https://tzero.com/), a US startup, claims on its website to be the first blockchain-based trading platform that integrate cryptographically secure distributed ledgers with existing market processes to reduce settlement time and costs, increase transparency, efficiency and auditability.

4.4. Health care

Health care costs are skyrocketing, estimated to be around 10% of GDP in developed countries and exceeding 17% (close to $3 trillion) in the USA. This means that any effort to improve health care services can result in substantial savings and blockchain technologies are prime candidates to achieve such savings while improving efficiency and probably saving lives at the same time. There are short-term blockchain applications ready to apply and ambitious, long-term ones aimed at revolutionizing the health industry.

- *Security and trust:* collect complete health data (medical reports for each patient, history of illnesses, lab results and X-rays) in a secure manner, using a unique identifier for every person and only allow the sharing of such data with the express permission of the

individual involved. Blockchain technology will eliminate the more than 450 health data breaches, affecting over 27 million patients, reported in 2016.

- *Exchangeability of information:* health information between the various actors is not communicated freely creating silos that hinder its effective utilization to improve health care. Blockchain technology can improve both the exchangeability of information and its quality leading to significant benefits.

- *Claim settlement and bill management:* facilitate claim settlement by reducing bureaucracy and introduce bill management to reduce fraud and speed up payment. This can be achieved more efficiently by creating consortia of health providers and insurers.

- *Authentication of medical drugs:* ensure the integrity of medical drugs based on current industry estimates pharmaceutical companies incur an estimated annual loss of $200 billion due to counterfeit drugs globally while about 30% of drugs sold in developing countries are considered imitations.

- *Clinical trials and medical research:* it is estimated that as much as 50% of clinical trials go unreported, and that investigators often fail to share their study results. Blockchain technologies can address the issues through the time-stamped, immutable records of clinical trials. Most importantly, the technology could facilitate collaboration between participants and researchers and could contribute in improving the quality of medical research.

Estonia has implemented a blockchain application, eHealth, covering all its citizens. In addition, there are a number of startups like GEM claiming to have developed the first application for health claims based on blockchain technology. This is done by introducing real time transparency and substantially reducing the time for bills to be paid by the sharing of the same platform among those involved. There are several other startups, some of which are already operating, and some on the way to becoming functional, like Guardtime, operating in Estonia and being used by patients, providers, private and public health companies and the government to store and access information from their eHealth system in a safe and efficient way. Similar functions are provided by Brontech, an Australian startup, offering reliable health data to improve the diagnostic process among others; Health Co aims at revolutionizing the relationship between medical researchers and users; Factom, Stratumn and Tierion are mostly concerned with improving the quality of health data while the purpose of Blockpharma is to fight drug counterfeiting.

4.5. Retail

The multinational eBay is the leader for online commerce between consumer-to-consumer sales. OpenBazaar, is a new startup challenging eBay by utilizing blockchain technology to decentralize online person-to-person trade. By running a program on their computer, users can connect to other users in the OpenBazaar network and trade directly with them. This network is not controlled or run by an owning organization but is decentralized and free. This means there are no mandatory fees to pay, and that trades are not monitored by a central organization (see https://www.cbinsights.com/company/openbazaar).

5. Applications

5.1. Smart contracts

Smart contracts are probably the blockchain technology with the highest potential to affect, or even revolutionize all sorts of transactions from the execution of wills to the Internet of Things (IoT). The major innovation of smart contracts is the elimination of trusted intermediaries. Consider, for example, the executor of a will who approves the directives of the deceased of how the money will be spent/allocated. Instead of an executor, a programmable, legally binding smart contract can achieve the same purpose, using blockchain technology, avoiding the trusted intermediary, while reducing costs and improving efficiency. An additional, application of smart contracts is with IoT, facilitating the sharing of services and resources leading to the creation of a marketplace of services between devices that would allow to automate in a cryptographically verifiable manner several existing, time-consuming work flows [7]. Most importantly, such technology is the central principle behind **Ethereum** (see below), a new extension of blockchain technologies focusing on running the programming code of decentralized smart contract applications.

5.2. Supply chain

Supply chain operations are dominated by paper-based methods requiring letters of credit (costing 1–3%) and factoring (costing 5–10%), increasing costs by an estimated trillion dollars and also slowing down transactions. Such costs could be reduced substantially, using blockchain technology that will eliminate intermediaries by establishing trust between buyers and sellers. There are several startups, among them, Skuchain, aiming its blockchain technology at the intersection of payments (letter of credit or wire transfer), finance (operating and short-term trade loans) and Provenance focusing on tracking the authenticity and social and environmental credentials of goods from the source all the way to the final consumer. In addition to startups, big companies like Walmart, are also aiming at exploiting the advantages of blockchain technology to improve efficiency and reduce supply chain costs [8].

5.3. IoT

Blockchain could revolutionize the IoT if applied securely to the estimated 8.5–20 billion of connected IoT devices that existed in 2017 and are expected to grow to 1 trillion by 2020. Exploiting the information generated by IoT devices intelligently can transform our homes and cities and have a profound effect on the quality of our lives while saving energy. According to Compton [9], *"Because blockchain is built for decentralized control, a security scheme based on it should be more scalable than a traditional one. And blockchain's strong protections against data tampering would help prevent a rogue device from disrupting a home, factory or transportation system by relaying misleading information"*. Eciotify, a startup, specializing in applying blockchain to the IoT, plans to roll out applications utilizing blockchain technology for IoT devices.

5.4. Decentralized cloud storage

Computer storage was decentralized in individual computers until about a decade ago when Dropbox was founded providing the first, modern, centralized cloud storage unit. Since then cloud computing was introduced revolutionizing applications by encouraging firms to outsource their storage needs to the likes of Amazon, Google or Microsoft Web Services. The advantage of such services was lower costs and greater reliability. Blockchain technology aims to re-decentralize computer storage to individual computers all over the world. According to experts [10], there are three major reasons for such a switch. First, the cost of most cloud services is around $25 per terabyte per month while the corresponding one of blockchain storage is 12.5 times cheaper at $2 per terabyte/month. Second, there is greater security as blockchain data is encrypted, meaning that only users holding the appropriate keys can view it (data stored in commercial cloud services could be viewed by third parties). Finally, blockchain cloud storage is immutable while providing a record of all historical changes done on the data.

5.5. Certification

One of the great promises of blockchain technology is that it can serve as a decentralized, permanently unalterable storage alternative for all types of information, or assets, not just as a currency or payment system. This makes the technology a prime tool for certifying all sorts of information, transactions, documents and records. What has attracted the greatest interest, however, is the certification of data (with the startup Stampery being the leader) and of identities (with the startup ShoCard being the leader). There are many, additional areas where certification using blockchain technology can be applied including the issuing of IDs and even voting.

6. Other blockchain applications

There are many additional applications exploiting blockchain technologies. These include true decentralized ride-sharing services (Uber and Lyft are actually centralized taxi services) like those offered by La'Zooz and Arcade City. Stratumn, a platform aiming to automate auditing, Synereo whose purpose is to aid users to create content, publish and distribute it online, Docusign offering the eSigniture solution and Steem, a social media platform where anyone can earn rewards, with some of these startups already operational while others are still being developed.

7. Specialized blockchain VC firms and geographical distribution of funding

According to FinTech News, in Switzerland eight major Venture Capital Firms have invested more than $1.55 billion in bitcoin and blockchain startups since 2012. Country wise the USA dominates the race with 55% of the total, followed by UK with 6%, Singapore with 3% and Japan, South Korea and China with 2% each. As interest in blockchain technologies increases, it is expected that VC investments will increase too accelerating the number of available applications.

8. Ethereum

Ethereum, like bitcoin, is a distributed public blockchain network (developed by the non-profit Swiss foundation of the same name) upholding its unique capabilities (Trust, immutability/transparency, disintermediation, low costs) but with the additional three:

- Running applications exactly as programmed without any possibility of downtime, censorship, fraud or third-party interference.

- Enabling developers to build and deploy decentralized applications, serving specific purposes that become part of the blockchain network and as such not controlled by any individual or central entity which is the case of Internet applications.

- Exploiting the ethereum virtual machine (EVM) to run any desired program, written in any programming language, by using the EVM developers, without the need to create blockchain applications from scratch but can utilize the thousands of existing ones already available (one type of such applications can be smart contracts).

8.1. Blockchain technologies: future prospects and major challenges

Blockchain is becoming one of the most remarkable technologies since the appearance of the Internet [11]. The large number of innovative applications based on this technology and the great interest shown from business firms, government organizations and individuals is mainly due to its ability to assure trust between parties that do not know each other, guarantee the safety of transactions and attest to the trustworthiness of the information, in addition to its other advantages. The interest in the technology can be seen from the Consensus Blockchain Conference, held in May 2017, which attracted more than 2000 participants and was just one of the more than 200 conferences held during 2017, as well as the more than 110 startups established in recent years and the exponentially increasing number of students attending blockchain programs. For instance, in the University of Nicosia's online blockchain course, there were 164 registrations from all over the world in 2017, versus 23 when this program was offered for the first time in 2013. In addition, there are 5495 registrations, from all five continents, who follow its MOOC class this year, versus 642 when it was first offered in the Spring of 2014. These numbers show the growing interest from the part of students while the university's blockchain placement office receives numerous requests each week from companies asking for graduates from its blockchain programs that could work for them.

The previous section of this chapter covered the blockchain technology and the various applications already, or in the process of, being implemented. This section discusses its future prospects and the challenges until its widespread adoption by business firms, governmental organizations and individuals. Faster and cheaper computers, lower storage costs and a host of specialized applications (some of them already discussed in the previous section) will accelerate its widespread adoption and will produce disruptive changes that will become revolutionary when blockchain is combined with AI algorithms, exploiting the advantages of both technologies. There are always the doubters saying that blockchain is overhyped [12, 13] but the same

was true when the Internet was in its infancy back in 1995. In a *Newsweek* article in February of that year, Clifford Stoll, a computer expert, wrote *"Baloney. Do our computer pundits lack all common sense? The truth is no online database will replace your daily newspaper, no CD-ROM can take the place of a competent teacher and no computer network will change the way government works"* [14].

9. Future prospects

Recently, Christine Lagarde, IMF's Managing Director, gave a talk at the Bank of England entitled "Central Banking and FinTech, A Brave New World?" [15] providing her views of banking and policy making in the year 2040. Her talk concentrated on three themes (virtual/digital currencies, new models of financial intermediation and AI, all three major concerns of this paper too) and how they will affect the future as well as what should be done to deal effectively with the challenges they will pose. Her advice *was "we–as individuals and communities–have the capacity to shape a technological and economic future that works for all"*, adding that we have a responsibility to make it work, assuring that humans will be needed for all important decisions, even though machines will certainly play a greater role as time passes.

9.1. Governments adopt blockchain for their entire operations

Some countries are experimenting with blockchain while a few are ahead in adopting the technology in some functions of their operations. Estonia is a pioneer having already applied blockchain-based services in eHealth, eSecurity and eSafety, eGovernment Services and eGovernance (including iVoting), estimating that such services save 100 years of working time for its 1.3 million citizens. Countries like Sweden follow Estonia's example while Dubai plans to implement blockchain to its entire government by 2020, reducing CO_2 emission by 114 million tons a year from fewer trips and saving 25.1 million hours and $1.5 billion annually from productivity increases in document processing alone [16]. According to an IBM sponsored survey [16], 9 in 10 government executives plan to make blockchain investments in financial transactions, asset and contract management and regulatory compliance by 2018. **Figure 1** shows the expectation of these executives to implement blockchain. According to the *Economist*'s article [17], governments may become big backers of blockchain technology as they come to understand its benefits that according to Brian Forde, of the Massachusetts
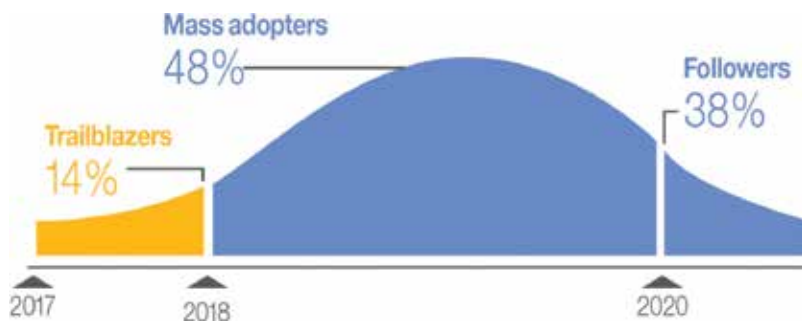
Figure 1. First to finish: Respondents' expectations of when they will have blockchains in productions and at scale.

Institute of Technology, is the driving force behind its widespread adoption. According to **Figure 1** and Forde, the future will probably witness a considerable number of blockchain applications in all areas of governmental operations.

Certificates and IDs are issued exclusively on blockchain: toward the end of 2017, the Dubai Land Department became the world's first government entity to conduct all its transactions through Blockchain technology [18]. Along the same direction, the Swedish National Land Survey and FinTech startup ChromaWay will test launch an initiative to put all land title records on Blockchain, and thereby safeguard the rights and interests of genuine property owners, eliminating or seriously reducing the chance of fraud [19]. Land records is just the tip of the iceberg, with all government (IDs, Passports, Driver Licenses, Birth Certificates, etc.) and educational certificates (Graduation Diplomas, Records of Programs/Courses taken, etc.) of potential candidates to be issued using blockchain technology. It is highly likely, therefore, to see a surge for this to happen with considerable cost savings, reduced bureaucracy and improved level of services.

9.2. Virtual (digital or crypto) currencies are adopted

While governments are buoyant about adopting blockchain for their operations, they are not so sure about virtual currencies, such as bitcoins, afraid of being used for tax evasion and possible criminal activities associated with the dark web. At present, the legal status of virtual currencies varies considerably from one country to another, with no indications of what countries plan to do in the future. China's recent decision to ban Initial Coin Offerings (ICO), calling them 'illegal fundraising' [20] as well as that of Russia to block cryptocurrency exchanges, are an indication of how virtual currencies are being treated by governments. At the same time, some countries (Switzerland, Singapore, South Korea, Japan, Dubai and Bahrain) are more open to adopt virtual currencies alongside their legal money, while others are openly hostile to its adoption. At the same time, international bodies like IFM encourage such an adoption, initially at least from countries with weak institutions and unstable national currencies. As time passes and the problems of volatility and hacking are addressed, virtual currencies are likely to play a complemental role, supplementing national ones, in trade and financial transactions, among others. However, at present, their future prospect is uncertain.

10. eHealth records

For health records to be useful they must be shared among doctors, laboratories, hospitals, pharmacies, government health agencies, insurance companies and researchers while, at the same time, protecting patients' privacy against unauthorized usage and breached health records. Although the challenge for doing so is tremendous, the Estonian eHealth Foundation is operating with considerable benefits, as a secure health record system that can become an example for other countries to follow, although it may be more difficult given the complexities of implementing the system in larger nations. In the USA, there are serious efforts to implement a blockchain health system that among other achievements can reduce fraudulent claims that are estimated at around 5–10% of health care costs at present. The challenge is how to digitize and standardize all health records, some of which are hand written. One system being developed to do so is MedRec [21] that according to its developers *"doesn't store*

health records or require a change in practice. It stores a signature of the record on a blockchain and notifies the patient, who is ultimately in control of where that record can travel. The signature assures that an unaltered copy of the record is obtained. It also shifts the locus of control from the institution to the patient, and in return both burdens and enables the patient to take charge of management" [22]. According to Das [23], blockchain will probably play a significant role in the healthcare industry as it has started "*to inspire both relatively easily achievable and more speculative potential applications*". Healthcare authorities, governments and providers are excited about the available possibilities and are investing to achieve them, although these achievements maybe more evolutionary than abrupt.

10.1. Business firms adopting blockchain for their internal operations and external transactions

Blockchain, as discussed, is a distributed ledger of trustworthy digital records whose safety is assured and its history can be traced as new data is added and chained, at the end of old ones while no information can be erased. Businesses that can leverage these unique advantages can harness significant gains in efficiency, including lower costs, more effective auditing (the data is immutable) and eliminating, or making fraud practically impossible.

10.2. The banking and financial sector and FinTech firms

Blockchain technology can be used for secure and direct alternatives to the complex and expensive banking processes used today, reducing transaction costs from $25 to less than a single dollar and avoiding costly intermediaries [3]. Such a huge saving has obliged practically all major banks to test the technology and many of them have joined R3, a startup developing Corda, a blockchain-based platform geared toward the banking industry. Corda and similar platforms will transform the sector by simplifying operations, eliminating intermediaries, reducing operating costs and offering a wide variety of new, innovative products and services, in addition to opening up banking to billions of people who are excluded at present. Financial firms face similar challenges as banks. In remarks at a FinTech-focused conference at the end of September 2017, Yasuhiro Sato, the president and CEO of the Mizuho Financial Group, said "*the technology could 'change the strategies of international financial institutions,'* adding *'we should have the courage' to make the shift to blockchain **now**"*. The Japanese Bankers Association (JBA) announced earlier in September 2017 that it will partner with IT provider Fujitsu to test the viability of using a blockchain across financial services. Blockchain will transform the banking/financial sectors, as FinTech startups are disrupting incumbents by developing innovative blockchain platforms and offering new products/services at lower prices.

11. Supply chain operations

As mentioned, supply chain transactions are dominated by paper-based, time-consuming and bureaucratic procedures, involving banks, financial firms and custom agencies among others. In the future blockchain can eliminate the paper trail and introduce trust among the

various players while also assuring firms receiving materials/parts and consumers on the authenticity of goods (from the raw materials to the final product). This can be done, for instance, by installing RFID tugs that can immutably record every movement of material/product, guaranteeing its provenance and testifying its physical presence, thus, eliminating the need for letter of credits, factoring and detailed inspections. Moreover, the optimization of supply chain can be achieved at present using AI for its logistics part (scheduling and planning) while it can be extended in the future to automate the majority of supply chain transactions (in conjunction with smart contracts) that could include the majority of AI transactions.

11.1. VA (autonomous vehicles) and IoT (Internet of Things)

The safety provided by blockchain technology is indispensable for the smooth running of self-driving vehicles and the untroubled functioning of IoT devices. By 2020, it is estimated that a sizable number of AVs will be on the road while there will be more than 1 trillion IoT gadgets, providing a unique challenge for blockchain technology to provide interconnectivity for all AVs and the smooth integration of the trillion of IoT devices. The implications are immense. If AVs are interconnected, they could communicate traffic jams, facilitate car sharing, receive and make payments and select the best insurance option among other tasks that can be performed using blockchain. Interconnected IoTs can optimize the functioning of all its devices, say at home, setting optimal temperatures, reducing energy consumption, ordering food and checking and paying utility bills.

12. Smart blockchain contracts instead of lawyers

Despite being in their infancy, smart contracts hold the potential to become a groundbreaking legal innovation, becoming a cornerstone of future commerce. At present, there are several problems limiting its applicability as a legal document [24]. Once these problems can be resolved, they can safely move assets around, interact with IoT devices and automate many business-related processes that demand human resources. How smart contracts will affect lawyers and law practices is debatable, with some predicting a serious decline in the need for lawyers [25] or at least providing an alternative to expensive legal practices.

13. Decentralized autonomous organizations (DAOs)

DAO is another major innovation of blockchain technology. A DAO is a company without a CEO, managers, employees or office buildings. It is created and run based on the computer code included in a smart contract. Although, the first DAO firm was hacked and its assets were stolen [26], the potential for DAOs are significant once the technical security problems are resolved. For instance, there is no reason for portfolio funds solely investing in market indexes to pay expensive executives, employ personnel and occupy offices when they can be

run more effectively as a DAO, open 24/7. There are immense possibilities to be exploited, leading to great cost reductions and more efficient operations as DAOs, once perfected, are not prone to human errors.

14. Other applications

There are numerous, additional applications of the blockchain technology pointing to substantial improvements. Some of them are listed below while there is practically no limit to future ones being developed and implemented

- Blockchain-enabled energy trading saving millions of dollars per year.

- Maritime insurance, reducing costs, decreasing fraud and speeding up the settlement of claims [27].

- Identifying epidemics faster while avoiding to cause panic [28].

- Educational material can be exchanged safely among academic institutions while safeguarding the intellectual rights of the writers [29].

As the adoption of new technologies has accelerated over time [30], the same phenomenon would probably occur with blockchain, resulting in more applications and faster penetration rates allowing us to exploit its considerable benefits in record time and witnessing quickening progress in the field.

15. Challenges

The blockchain challenges can be classified as *general*, referring to the technology itself and *specific* ones concerning virtual currencies.

General: adapting the blockchain technology and integrating it with existing IT systems may require significant changes, or even complete replacement of such systems, considerable initial investments and difficulties in hiring personnel to implement the technology. Although these problems are important, ready-made solutions and open systems may alleviate them, which are no different to when the Internet or other new technologies were first introduced. Another concern is the high electricity consumption required to run all of the computers in the network that some estimate to be equal to that of Ireland [31]. To avoid this problem alternative technologies to pure blockchain have been developed and utilized. DeepMind, for instance, uses a method called Merkle trees to track data changes without requiring verification from all networked machines. Such trees allow the efficient and secure verification of the contents of large data structures when the major objective is the safety and immutability of the data rather than ensuring trust between the parties involved. Similarly, the "algorand" algorithm [32] substantially reduces the amount of computations required and possesses

additional desirable properties. In the future, transaction speeds, verification times and data limits will further improve through innovations in order to deal with the exponentially growing number of transactions.

Specific: virtual currencies are currently too volatile and therefore too risky to be acquired by the public while the fear of hacking and fraud is present. In addition, dealing with technical problems such as programming bugs in the code of smart contracts must be dealt with, as their consequences when the contracts are executed are critical. Finally, the problem of scalability of the blockchain technology must be addressed as some platforms are reaching their capacity and storage limits. The hope is that as prices rise so will the need for innovative solutions that will eventually solve practically all problems.

15.1. Combining blockchain and AI

As we have shown in this chapter, blockchain is a groundbreaking technology permitting the safe and reliable storage and transmission of data, among its other advantages. AI, on the other hand, is a revolutionary technology that can learn on its own by analyzing and discovering patterns in massive amounts of (big) data. There is, therefore, a natural complementarity between the two, as blockchain safely stores/transmits trustworthy data while AI requires huge amounts of reliable data to discover patterns and learn. In this section, we discuss the complementarity between the two technologies and consider the breakthrough innovations that could result by marrying them. The potential benefits are expected to be in the areas of medicine, autonomous vehicles (AV), smart contracts, Internet of Things (IoT), decentralized autonomous organizations (DAOs) and many additional areas of applications, not yet conceived at present. In many cases, AI could not be used without the assurance of the safety and reliability of the data provided by blockchain and vice versa the value of many blockchain applications will be limited without AI.

Two examples can illustrate the complementarity and mutual benefits of joining blockchain and AI. Consider AVs in the simple case, as more carmakers adopt "over the air (OTA)" software updates for their increasingly connected and autonomous cars the risk of a hacker hijacking and stealing the car will also increase. In a worse-case scenario, a car can be forced to cause accidents or create traffic jams while the worst possibility would be to hijack and program the car to accomplish simultaneous terrorist attacks in many cities. Similarly, if IoT devices can be hacked, a home's security will be compromised, or its equipment can malfunction. Therefore, the safety provided by blockchain is indispensable for the smooth utilization of AVs and IoTs. On the other hand, consider a smart contract application that depends on some environmental assumptions for its correct execution. Such a contract would be outdated once some of these assumptions do not hold, making AI monitoring imperative in order to allow learning and determining on its own when the environment has changed. Although at present the blockchain and AI technologies may not be at the point of being successfully combined, the prospects for doing so in the near future are encouraging, motivated by the substantial expected benefits. The remainder of this section describes such advantages, clearly recognized in China where the first alliance for integrating artificial intelligence and blockchain is being established to harness these benefits [33].

16. Government operations

Governments, apart from some pioneering ones already mentioned, are slow in adopting new technologies and blockchain and AI are no exceptions, particularly when AI as a technology is still in a developmental stage, apart from some applications in games and those involving language and image recognition [34]. This does not mean that there will not be significant progress in the future, as the steepest progress in AI only occurred a few years ago. At present, however, the majority of AI applications are centered on digital assistants, answering questions in natural language and in image, including face recognition techniques [35]. The future prospects however are huge, with estimated benefits running into the billions. AI applications could range from fighting tax evasion to establishing monetary and fiscal policies. The catchword of "cognitive AI", if it becomes a reality, can have profound implications in not only saving billions but also providing higher quality services to the public and increasing the level of democratization. Some governments such as those of Dubai are planning to introduce Blockchain into their entire operations reducing bureaucracy, improving their efficiency, reducing waste and pollution and saving billions in the process.

17. Digital currencies

It is not obvious how AI can be combined with the blockchain technology used in bitcoins and other cryptocurrencies, although this could be achieved in the future when DAOs and robots will be introduced, owning property and holding assets. In such a case, they will have to use AI to make the necessary M2M transactions or using bitcoins for making and receiving payments.

17.1. eHealth

While blockchain can assure safety and reliability, adding AI capabilities can greatly benefit the health sector. At present AI is mainly used for detecting abnormalities in X-rays and CT scans, a task performed at least as accurately as humans can, and for assuring a greater level of personalized medicine. According to experts, the future holds significant inventions given the momentous benefits that can be achieved by reducing medical costs and improving the quality of medical care. For this reason, all big players (Google, Microsoft, Apple and Amazon), as well as a host of startups are actively exploring AI for medical applications, aimed at improving the more effective utilization of patients' data, the accuracy of diagnosis, providing better recommendations, based on evidence-based research findings, and several other possibilities. These applications are on top of improvements in robotic surgery and digital advice provided though smartphone applications. According to Accenture [36], key clinical health AI applications can potentially create $150 billion in annual savings for the United States healthcare economy by 2026.

18. The banking and financial sector

The benefits of AI for the back office of banks and financial firms are widespread, as large histories of data are available. For a long time before AI was introduced, risk and fraud

detection was predictive with great success using statistical decision rules. AI has improved such rules to a new level by allowing learning through the analysis of a huge amount (big) of data to identify patterns and improve decision-making. Klarna, a Swedish e-commerce company, provides instant evaluation of customers' credit worthiness for buying goods without a credit card. The same task is done by the Chinese Yongqianbao and several other firms. In addition, *"AI technology is being used to find the speediest way to execute trades, to make bets on market momentum, and to scan press releases and financial reports for keywords that could signal that a stock will rise or fall"* [37]. However, this is not the same with more accurate fore-casting. Unfortunately, stocks and commodities behave like random walks and cannot be predicted any better than using the most recent price for future ones, according to efficient market theory [38]. For instance, in a recent study conducted by one of the authors of this paper [39], comparing statistical and AI (ML or NN) forecasting methods, it was found that the former were more accurate than the AI ones, half of which were less accurate than a random walk benchmark.

Clearly, present AI applications in banking and finance are just the tip of the iceberg and soon the power of AI to deliver better experiences, lower costs, reduce risks and increase revenues will become a reality and they may even progress to more accurate forecasting.

A prime example of successful AI applications is Numerai [40], a San Francisco hedge fund that makes trades using machine-learning models built by thousands of anonymous data sci-entists paid in bitcoin. Another is Polychain, a fund that buys bitcoin and other digital curren-cies and invests in a radically new breed of businesses owned, funded, and operated entirely by decentralized networks of anonymous online investors.

19. Supply chain operations

Blockchain technology is already utilized in supply chains while its integrations with AI is still in its infancy apart from its logistic part (what used to be the old scheduling/planning tasks) which is used extensively by some firms [41]. The challenge is in the future to extend AI to the remaining parts of the supply chain. Amazon, a pioneer in AI, has moved beyond just responding to customer demands by developing a whole profile for each customer and using such data in its AI applications. Manish Chandra and Anand Darvbhe of Accenture [42] point out, *"The use of AI in supply chains will ultimately result in spawning an ecosystem where supply chains link themselves with each other, enabling seamless flow of products and information from one end to the other"*, completely automating the process and achieving significant benefits in the process.

20. AV and IoT

Employing AI to AVs can go beyond just following a set course for taking its passengers from point A to B by continuously analyzing traffic information from connected AVs and learning to determine the route depending on the time, the day, the weather conditions and a host of other factors. Moreover, it can even modify the course of a journey, if necessary, when the AI

determines that traffic patterns are changing. Similarly, IoTs devices can go beyond setting temperatures and ordering food by using AI to predict what the owners want and modify the settings to satisfy their evolving desires.

20.1. Cognitive blockchain smart contracts (IBM) and DAO

IBM is experimenting with turning smart contracts into "cognitive contracts" that can learn and adapt using AI [43]. This can be done by identifying pattern changes in the data, recognizing interesting interactions, detecting suspect activities, etc., in order to make recommendations for updating the smart contracts and taking specific actions based on insights gained from AI. Clearly, such cognitive contracts can be applied to DAOs to improve their effectiveness and value.

20.2. Matrix chain: merging blockchain and AI

Lately, efforts are being made to integrate AI and blockchain technologies into a single application. At the technical level this has been attempted by a new type of blockchain called the "MATRIX Chain" [44] whose aim is to merge blockchain and AI and set the path toward blockchain 3.0. The benefits that such technology will bring to distributed ledger technology comes down to making blockchain smarter and adding its ability to evolve through self-learning without the need to introduce AI as a separate technology.

A summary of the major applications integrating blockchain and AI is presented in **Table 1**, also showing an estimate of the extent of usage of each of the two technologies and the direction of what would need to be done to improve their future integration.

Major applications	Application uses mostly	Future requirements	
		More AI	More BC
Government operations	Neither BC/AI*	Yes	Yes
Digital currencies	BC	Yes	
eHealth	BC	Yes	Yes
Banking	BC	Yes	
FinTech	AI		Yes
Supply chain	Little BC	Yes	Yes
Autonomous vehicles (AV)	All AI		Yes
Internet of Things (IoT)	All BC	Yes	
Smart contracts	All BC	Yes	
DAOs	BC	Yes	

*Apart from exceptions as Estonia and a few other countries.

Table 1. Major applications, their current utilization of BC/AI and their future requirement.

21. Conclusions

Blockchain technology, according to Muneeb Ali, Blockstack Co-Founder, *"can help us advance from a 'don't be evil' world to a 'can't be evil' world"*. Blockchain transactions assure trust and reliability, improve security and remove intermediaries from value chains. In a chapter, Tasca and Ulieru [3] state that, in a not-so-distant future, our economic structure will be organized around person-to-person decentralized platforms that could enable real sharing of marketplaces without intermediaries and central hubs, where all transactions between consumers and service providers will be done through decentralized, person-to-person networks. They discuss Uber and Airbnb as examples. Both companies create extra value exploiting their monopolistic advantage, derived from their centralized, proprietary software platforms, which allow them to dictate their conditions to drivers/owners and customers. LaZooz, using blockchain technology, on the other hand, has developed a decentralized transportation platform owned by the community, utilizing vehicles' unused space to create a variety of smart transportation solutions. LaZooz works with a "Fair Share" rewarding mechanism sharing value creation among developers, users and backers. Similarly, Slock (an Italian startup), uses open source blockchain technology, to develop the Universal Sharing Network (USN) to eliminate Airbnb's monopolistic advantages.

In addition to startups, established companies also seek to exploit the advantages of blockchain technology and diminish the monopolistic advantages of Internet giants. The CEO of TUI, the largest tourist firm in the world, believes that blockchain technology will break the almost "monopolistic" hold that Priceline, Expedia, Booking.com and Airbnb have today in the lodging and distribution ecosystem [45]. He believes that these firms create superior margins because they take advantage of their monopolistic power and that blockchain will destroy that. TUI, he explained, has already moved all of its contracts into its private blockchain. "We are using it today predominantly to have mechanisms to swap bedstock between different PMSs [Property Management Systems]," he said. "The next step is that the whole inventory will be on the blockchain." Then using smart contracts, which are simply code snippets that execute automatically on the blockchain, Joussen argues it can easily manage and automate a large part of bedstock and hotel capacity between all the markets TUI operates.

Clearly, TUI is not the only company developing blockchain applications. So, the critical question is how all these applications will affect the competitive landscape and how innovative startups will utilize blockchain technologies to disrupt established players and create the corresponding success stories of Amazon, Google and Facebook, among others, in the emerging Internet of value. In answering this question, we should have in mind Amara's law that states, *"We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run"*. We strongly believe that in the long term, the Internet of Value will bring changes of equal or greater magnitude to those of the existing Internet of communications. The critical question is how to recognize such changes as soon as possible and how to profit by implementing them to gain competitive advantages. There is little doubt in our minds that in the next couple of decades, innovative, entrepreneurial startups marrying blockchain and AI technologies will disrupt established industry leaders such as Google, Amazon, Facebook,

Uber and Airbnb, although they may not reach their size because of the limitations being imposed by the decentralized attributes of the blockchain technology.

For us, the most interesting question is "who and in what areas are going to emerge the new Googles, Baidus, Facebooks, Amazons and Alibabas and how will they successfully exploit blockchain and AI, although such a marriage may still be several years away?"

Author details

Spyros Makridakis*, Antonis Polemitis, George Giaglis and Soula Louca

*Address all correspondence to: spyrosmakridakis@gmail.com

Faculty University of Nicosia, Members of the Blockchain/AI Team, Institute For the Future (IFF), Nicosia

References

[1] IBM. IBM Study: C-suite executives exploring Blockchain aim to disrupt, Not Defend; 2017

[2] McKendrick J. Blockchain as Blockbuster: Still Too Soon to Tell, But Get Ready. Forbes; 2017

[3] Prisco G. Move Over Uber: Blockchain Technology Can Enable Real, Sustainable Sharing Economy; 2016. Available: https://bitcoinmagazine.com/articles/move-over-uber-blockchain-technology-can-enable-real-sustainable-sharing-economy-1480629178/. Last accessed 30 October, 2017

[4] CBInsights.Com. Banking Is Only The Beginning: 36 Big Industries Blockchain Could Transform. 2018. Available: https://www.cbinsights.com/research/industries-disrupted-blockchain/?utm_source=CB+Insights+Newsletter&utm_campaign=fa48df10a8-ThursNL_02_01_2018&utm_medium=email&utm_term=0_9dc0513989-fa48df10a8-90141994

[5] Tasca P, Aste T, Pelizzon L, Perony N, editors. Banking. Beyond Banks and Money: Springer International Publishing; 2017

[6] Kocianski S. The blockchain in banking report: The future of blockchain solutions and technologies. 2017 Available: http://www.businessinsider.com/blockchain-in-banking-2017-3. Last Accessed 30 October, 2017

[7] Christidis K, Devetsikiotis M. Blockchains and smart contracts for the internet of things. In: IEEE Access. 2016. DOI: 0.1109/ACCESS.2016.2566339

[8] Lohade N. Dubai Aims to Be a City Built on Blockchain, WSJ. 2017

[9] Compton J. How Blockchain Could Revolutionize The Internet Of Things. 2017. Available: https://www.forbes.com/sites/delltechnologies/2017/06/27/how-blockchain-could-revolutionize-the-internet-of-things/#7617423b6eab. Last Accessed 30 October, 2017

[10] Maor R. Cloud Computing: The Future Belongs to Blockchain. . 2017 Available: http://www.forbes.co.il/news/new.aspx?Pn6VQ=L&0r9VQ=EIMKF. Last Accessed 30 October, 2017

[11] Marvin R. Blockchain: The Invisible Technology That's Changing the World. 2017. Available: https://www.pcmag.com/article/351486/blockchain-the-invisible-technology-thats-changing-the-wor. Last Accessed 30 October, 2017

[12] Bloomberg J. Eight Reasons To Be Skeptical About Blockchain. 2017. Available: https://www.forbes.com/sites/jasonbloomberg/2017/05/31/eight-reasons-to-be-skeptical-about-blockchain/1. Last Accessed 30 October, 2017

[13] Flieswasser K. Will blockchain disrupt or go bust?. 2017 Available: https://www.top-bots.com/6-challenges-preventing-widespread-blockchain-technology-adoption/. Last Accessed 30 October, 2017

[14] Stoll C. Why the web won't be nirvana. 1995. Available: http://www.newsweek.com/clifford-stoll-why-web-wont-be-nirvana-185306. Last Accessed 30 October, 2017

[15] Lagarde C. Central Banking and Fintech—A Brave New World?. 2017 Available: https://www.imf.org/en/News/Articles/2017/09/28/sp092917-central-banking-and-fintech-a-brave-new-world. Last Accessed 30 October, 2017

[16] Smart Dubai. Dubai blockchain strategy. 2016. Available: http://www.smartdubai.ae/dubai_blockchain.php. Last Accessed 30 October, 2017

[17] Economist (The). Intelligence Unit Artificial Intelligence in the Real World: The business case takes shape, London. 2016

[18] Gulf News. Dubai has world's first government entity to conduct transactions through Blockchain network. 2017. Available: http://gulfnews.com/business/property/dubai-has-world-s-first-government-entity-to-conduct-transactions-through-blockchain-network-1.2101819. Last Accessed 30 October, 2017

[19] Rajashekara M. 3 Ways In Which Fintech Is Riding The Blockchain Wave. 2017. Available: http://www.huffingtonpost.in/rajashekara-v-maiya/3-ways-in-which-fintech-is-riding-the-blockchain-wave_a_21876915/. Last Accessed 30 October, 2017

[20] BBC. China bans initial coin offerings calling them 'illegal fundraising'. 2017 Available: http://www.bbc.com/news/business-41157249. Last Accessed 30 October, 2017

[21] Ekblaw A et al. A case study for Blockchain in healthcare: "MedRec" prototype for electronic health records and medical research data. IEEE Conference August. 2017;**2016**:22-24

[22] Halamka D, Lippman A, Ekblaw A. The Potential for Blockchain to Transform Electronic Health Records. 2017. Available: https://hbr.org/2017/03/the-potential-for-blockchain-to-transform-electronic-health-records. Last Accessed 30 October, 2017

[23] Das R. Does Blockchain Have A Place In Healthcare?. 2017 Available: https://www.forbes.com/sites/reenitadas/2017/05/08/does-blockchain-have-a-place-in-healthcare/. Last Accessed 30 October, 2017

[24] Agrello. How to Make Smart Contracts Worthy of Their Name Using Artificial Intelligence. 2017. Available: https://blog.agrello.org/how-to-make-smart-contracts-worthy-of-their-name-using-artificial-intelligence-3a90e4dd3c47. Last Accessed 30 October, 2017

[25] Artificiallawyer.com. OpenLaw Brings Legal Norms to Blockchain Token Transactions 2017. Available: https://blog.agrello.org/how-to-make-smart-contracts-worthy-of-their-name-using-artificial-intelligence-3a90e4dd3c47. Last Accessed 30 October, 2017

[26] Levine M. Blockchain Company's Smart Contracts Were Dumb. 2016. Available: https://www.bloomberg.com/view/articles/2016-06-17/blockchain-company-s-smart-contracts-were-dumb. Last Accessed 30 October, 2017

[27] EY.com. EY, Guardtime and industry participants launch the world's first marine insurance blockchain platform 2017. Available: http://www.ey.com/gl/en/newsroom/news-releases/news-ey-guardtime-and-industry-participants-launch-the-worlds-first-marine-insurance-blockchain-platform. Last Accessed 30 October, 2017

[28] Jones B. The CDC Wants to Use Blockchain as a Weapon Against Deadly Epidemics. 2017. Available: https://futurism.com/the-cdc-wants-to-use-blockchain-as-a-weapon-against-deadly-epidemics/. Last Accessed 30 October, 2017

[29] Acheson N. Blockchain and Education: A Big Idea in Need of Bigger Thinking. 2017. Available: https://www.coinndesk.com/blockchain-education-big-idea-need-bigger-thinking/. Last Accessed 30 October, 2017

[30] McGrath RM. The Pace of Technology Adoption Is Speeding up, Harvard Business Review. 2013

[31] O'Dwyer KJ, Malone D. Bitcoin Mining and Its Energy Footprint, ISSC 2014/CIICT 2014, Limerick. 2014

[32] Chen J, Micali S. Algorand. Technical report, 2017. URL http://arxiv.org/abs/1607.01341

[33] AIES. First Alliance for the Development of Artificial Intelligence and Block Chain Technologies will be Established in China for the Development of Integration of the two Technologies. 2017 Available: http://aies.in/first-alliance-development-artificial-intelligence-block-chain-technologies-will-established-china-development-integration-two-technologies/ Last Accessed 30 October, 2017

[34] Pontin J. Greedy, Brittle, Opaque, and Shallow: The Downsides to Deep Learning, Wired.Com. 2018 Available: https://www.wired.com/story/greedy-brittle-opaque-and-shallow-the-downsides-to-deep-learning/

[35] Deloitte. AI-augmented government: Using cognitive technologies to redesign public sector work, The Deloitte Center for Government Insights 2017

[36] Accenture. Why the artificial intelligence is the future of growth. 2017. Available: https://www.accenture.com/us-en/insight-artificial-intelligence-future-growth. Last Accessed 30 October, 2017

[37] Satariano A, Kumar N. The Massive Hedge Fund Betting on AI. 2017 Available: https://www.bloomberg.com/news/features/2017-09-27/the-massive-hedge-fund-betting-on-ai. Last Accessed 30 October, 2017

[38] Malkiel BG. A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing. New York: W.W. Norton; 2017

[39] Makridakis S, Spiliotis E, Assimakopoulos V. The accuracy of machine learning (ML) forecasting methods versus statistical ones: Extending the results of the M3-competition. 2017. In: UNIC Working Paper. 2017

[40] Craib R. A New Cryptocurrency For Coordinating Artificial Intelligence on Numerai. 2017. Available: https://medium.com/numerai/a-new-cryptocurrency-for-coordinating-artificial-intelligence-on-numerai-9251a131419a. Last Accessed 30 October, 2017

[41] CB Insights. Amazon Strategy Teardown: Building New Business Pillars In AI, Next-Gen Logistics, And Enterprise Cloud Apps. 2017. Available: https://www.cbinsights.com/research/report/amazon-strategy-teardown/. Last Accessed 30 October, 2017

[42] Chandra M, Darbhe A. Artificial Intelligence: The next big thing in Supply Chain Management. 2016. Available: http://www.financialexpress.com/industry/artificial-intelligence-the-next-big-thing-in-supply-chain-management/329033/. Last Accessed 30 October, 2017

[43] IBM, Building Trust in Government: Exploring the Potential of Blockchain, IBM Institute of Business Value and the Economist Intelligence Unit, London. 2017 Available: https://www-03.ibm.com/press/us/en/pressrelease/52418.wss. Last Accessed 30 October, 2017

[44] Hebblethwaite C. Merging blockchain and AI with MATRIX Chain. 2017. Available: https://www.blockchaintechnology-news.com/2017/08/08/merging-blockchain-ai-matrix-chain/. Last Accessed 30 October, 2017

[45] Montali D. Blockchain Will Disrupt Expedia and Airbnb, TUI CEO Says. 2017. Available: https://skift.com/2017/07/11/blockchain-will-disrupt-expedia-and-airbnb-tui-ceo-says/. Last Accessed 30 October, 2017

Augmenting Reality with Intelligent Interfaces

Dov Schafer and David Kaufman

Additional information is available at the end of the chapter

Abstract

It is clear that our daily reality will increasingly interface with virtual inputs. We already integrate the virtual into real life through constantly evolving sensor technologies embedded into our smartphones, digital assistants, and connected devices. Simultaneously, we seek more virtual input into our reality through intelligent interfaces for the applications that these devices can run in a context rich, socially connected, and personalized way. As we progress toward a future of ubiquitous Augmented Reality (AR) interfaces, it will be important to consider how this technology can best serve the various populations that can benefit most from the addition of these intelligent interfaces. This paper proposes a new terminological framework to discuss the way AR interacts with users. An intelligent interface that combines digital objects in a real-world context can be referred to as a Pose-Interfaced Presentation (PIP): Pose refers to user location and orientation in space; Interfaced means that the program responds to a user's intention and actions in an intelligent way; and Presentation refers to the virtual object or data being layered onto the perceptive field of the user. Finally, various benefits of AR are described and examples are provided in the areas of education, worker training, and ESL learning.

Keywords: mixed reality, Augmented Reality for education, intelligent interface, Pose-Interfaced Presentation, PIP

1. Introduction

Researchers have long foreseen that Augmented Reality (AR) and Virtual Reality (VR) technologies have the potential to play a large role in our future daily lives, education, and business operations [1, 2]. The ubiquity of sensor-rich technologies in the pockets of the masses, the rapid pace of technological progress for such devices, and widespread interest in mixed-reality interfaces by both researchers and corporations, are harbingers of that foresight. Since it is

clear that we are moving toward a future where the virtual will play a big role in our real lives, how we best interface with the computers in our natural environment will become an increasingly important consideration.

An intelligent interface is one that learns about the user and can adapt to serve them better through a minimally invasive presentation. An interface designed to overcome challenges faced by specific user groups by responding to the context of the user can be called intelligent [3]. Mixed-reality applications should be created to assist people and to be responsive to a diverse range of needs in a variety of contexts. This paper will discuss the rise of Augmented Reality as a means to interface with data in an intelligent way by discussing some examples of AR designed specifically for the needs of different learner groups: workers, students, foreign language learners, and immigrants.

This literature review to parse the current state of Augmented Reality for Education was done between October and November 2017 using ERIC ProQuest, the Social Sciences Citation Index (SSCI) database, and Google Scholar. The search terms 'augmented reality' OR 'mixed reality' OR 'augmenting reality' + 'education' OR 'learning' were used. Search results were limited to articles and conference papers in English. Much of the foundational research and history of AR was discovered through reading the most widely cited works on AR according to Google Scholar's reverse citation search function. The purposes of this review are to place AR in the broader historical context of technologies that inhabit the spectrum of mixed reality within education technology research and to paint a compelling picture of the current state of AR for enhancing learning.

2. The spectrum of mixed reality

In the early 1990s, Tom Caudell and David Mizell were exploring the possibility of a heads-up display to enable workers to more efficiently create wire harness bundles during aircraft manufacturing at Boeing. They developed a see-through visual interface to identify, through the use of virtual cues, which real wires should be bundled together. They published the first academic paper that used the term 'Augmented Reality' [1] and are often credited with coining the term. Since its inception, AR has been widely studied in many contexts, while its underlying technologies have progressed by leaps and bounds [4]. The general principle of a continuum between the real and the virtual still exists [5] and is becoming an increasingly poignant consideration in our daily lives.

Augmented Reality is not a specific device or program; it is a type of human-computer interaction that occurs through a combination of technologies that superimpose computer-generated content over a real-world environment. Encompassing a broad range of technologies and components, AR has been historically defined as any system that: (1) combines the real and the virtual, (2) is interactive in real time, and (3) appears three dimensionally [2]. AR overlays virtual objects into the real world. These virtual objects then appear to coexist in the same space as objects in the real world for the purposes of interacting with the user in some meaningful way [6].

Figure 1. Modified version of Milgram's virtuality continuum [8].

Reality and virtuality act as two opposite ends of a continuum [7]. Milgram and Kishino [7] proposed that there were various ways that reality and virtual content could be mixed and presented. Reality could begin as a location in physical space or could be computer generated. In a VR setting, reality becomes entirely virtual and replaces natural reality to create an immersive experience. Inversely, the goal of an AR system is to "enhance reality with digital content in a non-immersive way." ([4], p. 79). A mixed-reality program can be to varying degrees exocentric or egocentric [7]; the user can feel as if reality is situated within the program, such as when playing an immersive 3D video game with a head-mounted display [HMD], or the program can seem to be layered on top of reality and situated within it, as when viewing context specific information about your environment on a smart phone or interacting with a virtual tour guide in a museum (**Figure 1**).

The real environment acts as a substrate for virtual modifications. In AR, the user can maintain a direct view of the substrate background upon which the virtual is layered, as with HoloLens [9], or alongside the substrate, as exemplified with smart glasses [10]. Alternatively, the substrate can be processed by a camera and presented on a display; this is what smartphones currently do. A seemingly more futuristic option is that the virtual object can be projected directly into reality, such as a spatial AR or hologram [11]. The focus of this paper is the exploration of AR interfaces for the benefit of different learner populations, but the points made herein can be applied to the VR or AV ends of the continuum as well. The exact mix of how much reality is injected into the virtual (or visa-versa) is worthy of consideration when designing and studying mixed-reality applications.

3. Pose-Interfaced Presentation: A classification framework

Klopfer and Squire [12] resist using technological features to define AR; they claim AR occurs in "a situation in which a real-world context is dynamically overlaid with coherent location or context sensitive virtual information" ([12], p. 205). Initially, many researchers have tried to craft an exact definition for AR (e.g., [2, 7]), but in Klopfer and Squire's opinion, any definition

would restrict the exact meaning, so any technology that combines real and digital information could be considered to be augmenting reality. An intelligent AR interface requires the combination of tracking input, to find the pose of the user, with displaying output to combine the real and virtual world together in order to support the user. Therefore, this paper proposes a new terminological framework with which to discuss the way AR interacts with users: an intelligent interface that combines digital objects in a real-world context can be referred to as a Pose-Interfaced Presentation (PIP):

- Pose—user location and orientation in space is accurately tracked and sent to the program;

- Interfaced—the program responds to a user's intention and actions in an intelligent way; and

- Presentation—virtual objects or data are layered on to the perceptive field of the user.

The following three sections discussing the current state of the field based upon the above PIP framework.

3.1. Pose: tracking on smart devices

The keys to a successful device running an AR interface are (1) the accuracy of the tracking inputs to accurately overlay virtual objects in the field of view, (2) speed of the processor to be able to layer more complex virtual objects (such as videos and 3d models) onto the field of view, and (3) ease of portability, so the user is free to move around in the real world unencumbered. An important consideration for successful AR devices is that they have an array of highly accurate sensors in order to establish and maintain detailed data about the environment. The application will make use of the data to situate a virtual output in real space. In a mobile phone, a visual AR interface uses the built-in camera to send image data to the application, which is programmed to find some pre-determined pattern (e.g., a QR code, picture, or building) that exists in the real world to determine user and device pose. The application layers a virtual object over a real-time display of the environment, presented on the screen. A location-based AR interface uses GPS co-ordinates and compass, while inertial tracking uses a gyroscope and accelerometer, rather than the camera, to track pose. Data triangulation through multiple-sensor input is required to eliminate drift in visually-established pose [4]. While AR visual, location-based, and inertial sensors are by far the most common methods of tracking pose, AR could also mean sensing sound, electromagnetic fields, radio waves, infrared light, or temperature. Hybrid pose tracking methods, combining input from multiple sensors at once, are far more reliable [13].

Pose tracking accuracy, and sensors in general, will continue to develop, as smartphone technology investment shows no signs of slowing. As of May 2016, the global average user range error for GPS receivers was ≤2.3 ft. 95% of the time [14]. GPS enhancement technologies such as Real Time Kinematic, which uses the GPS signal's carrier wave for measurement, have the potential to improve accuracy to the centimeter level. This increased level of sensitivity will allow AR-based mobile applications to be able to establish exactly where the device is located in space, and therefore make much better use of location information in context-aware applications.

Another area of rapid progress that will improve AR is visual sensor technology. 3D camera systems are already appearing in flagship smartphones and should continue to proliferate [15]. The use of multiple cameras can provide depth information in order to greatly improve pose

fidelity. One well-known approach for real-time 3D environment modeling is the Microsoft KinectFusion system. Newcombe et al. [16] describe KinectFusion as using visual data obtained from 'structured light depth sensors' to create real-time 3D models of objects and environments, and these models are also used for tracking the pose of the device and the user in the environment. By combining tracking of the environment with tracking of the users' limbs, KinectFusion allows for a highly accurate gesture-based human-computer interface that allows users to interact with a detailed 3D model of the environment. This type of 'natural user interface' [8] has enough fidelity to even track facial expressions [17] and may end up becoming as common in future mobile devices as multi-touch screens are today. We can expect pose tracking to improve in both speed and accuracy over the coming years.

3.2. Interfaced: responding to human intention

Once the user pose is known, the user needs to be able to interact with the program. AR programs, by their physically situated and reality-referenced nature, tap into an intrinsic human affinity for physical rules and spaces. The interface between the virtual and the real world is accomplished through physical gestures, voice, movement through space, and gaze [18]. That interaction should be intelligent; that is, it should allow interaction with computers in a minimally invasive, intuitive, and efficient way that minimizes excessive perceptual and cognitive load [19, 20]. An intelligent interface should allow a program to be able to accurately determine the wishes of the user and respond to their needs with minimal input and data display. The ultimate goal is to create an interface that is functionally invisible to the user and makes interacting with the virtual world as natural as interacting with real-world objects, removing the separation between the digital and physical. Augmented Reality is one of the first technologies to make this type of intelligent, reality-referenced interface possible [4].

Interfaces should support users' plans and goals in order to be effective. An intelligent interface presents information clearly and is easy to understand [19]. There is a lot of overlap between 'intelligent' interfaces and what could be called 'effective user-experience design.' Sullivan et al. [19] state that the requirements of intelligent interfaces are essentially those of human-computer interaction research in general. AR is a form of human-computer interaction that facilitates intelligent interface design because of its ability to make complex interactions feel grounded in reality. Situating the experience of computing in physically referenced space is a good first step toward creating a type of human-computer interface that is responsive and feels meaningful.

3.3. Presentation: Augmented Reality displays and devices

Our reality will be increasingly interfaced with virtual information; the trillion-dollar question is how the digital world should best be interacted with and presented to users. Research firm Global Market Insights [21] predicts that the global market for AR products will surge by 80% to $165 billion by 2024. Giant US blue-chip corporations have recently invested heavily in AR, in products such as Microsoft's AR headset and enterprise feature suite HoloLens [8], Apple's AR platform (ARKit) [22], Google (project X, formerly Glass) [9], Android's AR platform (ARCore) [23], Facebook's VR headset (Oculus Rift) [24], and ambitious startups such as Magic Leap [25] and Meta [26] who seek to create their own entire AR ecosystems. Although

sentiment on VR softened in the first half of 2017, corporate investors are now actively look-ing for mobile AR investment opportunities. Digi-Capital's new AR/VR Investment Report reported that $1.8 billion was invested across 27 AR/VR investment sectors in the first three quarters of 2017 [27].

4. Current and future trends in Augmented Reality

The future of AR lies within the usage and adoption of this new human-information interac-tion (HII) paradigm, not in the AR devices themselves. Technology will evolve over time, but fundamental improvements to HII and intelligent interface design will survive those muta-tions. Although head-mounted AR display technologies such as Meta Vision, Magic Leap, and HoloLens are grabbing all the media attention, they are merely stepping stones toward the widespread proliferation and integration of AR into mainstream culture. These flashy new devices are still too large and expensive for daily consumer use. HMDs have limited field of view and image fidelity as well as practical limitations such as battery life and limited prod-uct lifespan. That being said, HMD AR devices have demonstrated utility within commercial and industrial applications, as is discussed in Section 4.2 of this chapter.

As technology improves, mobile devices are becoming ubiquitous and capable of running AR applications. It used to be the case that immersive reality augmentation would require expensive, stationary computers and bulky HMDs. The first portable AR interfaces involved LCD screens tethered to desktop computers that provided the necessary tracking and pro-cessing power [28]. The sensors required for establishing the position and orientation (pose) of the user and device relative to the environment have also rapidly evolved since mobile phones received their first cameras in 1997 and Mohring and Bimber demonstrated the first mobile phone-based AR application in 2004 [4]. Today, smartphones can easily and smoothly run the software required for AR and are outfitted with a plethora of useful sensors/receivers (e.g., gyroscope, compass, accelerometer, light sensors, cameras, GPS, WIFI radio, near-field communication or NFC) which allow programs on smartphones to accurately track pose and make use of situational data that can be layered on top of the environment to augment the reality that the user experiences. Thanks to the development of more advanced smartphones and widely available software platforms such as Apple's ARKit and Android's ARCore, seam-less dialog between reality and virtuality can now be accomplished much more easily than ever before, and this bodes well for the future of mass AR adoption.

We are progressing toward a technological future where our daily reality is intermixed with digital information that is presented to us in a real-time, user-friendly way. Earlier in this paper we outlined the PIP framework for the discussion of AR; our smartphones are mature, widely used devices capable of the Pose-Interfaced Presentation required for advanced AR interface. Adoption is not going to happen all at once, but many areas of our lives will be noticeably changed over the coming years as AR slowly replaces more traditional means of human-computer interaction. Entertainment, marketing, commerce, online social interaction, and professional practice are all likely to see major technological changes due to AR.

Major corporations are placing a lot at stake in AR. Taking into account these recent trends in the AR development space, it is important to ask how this emergent technology can best be applied, and who might benefit from the paradigm shift of layering data onto reality rather than alongside it. AR development should consider how best to leverage this new form of human-computer interaction to empower people who need support. In order to do this, it is useful to identify what AR is especially good at doing in real-world contexts in order to help focus design efforts and maximize potential benefits.

As early as 2011, Yuen et al. [29] predicted five directions for AR and provided multiple examples. While many current AR educational projects have focused on scientific inquiry (e.g., astronomy, mathematics, architecture, and engineering), educators have been working on AR projects in fields such as such as language, art, political science, sport management, textiles, fashion merchandising, and food and nutrition. AR learning tools allow students access to capabilities and resources that can dramatically increase the effectiveness of individual study and discovery learning [29]. For example, many historic sites supply overlay maps and different points of historic information for their visitors. AR will allow visitors to experience an event such as a treaty signing or a battle while they visit an historic location. Medical students around the world will be able to easily and inexpensively work on digital cadavers or dummies that can be reset after each use. Through the situated learning implicit in AR, learners will be better able to transfer what they learn into authentic situations.

With the continued development of AR-enabled educational books and games, combined with AR models of the real world, AR has the potential to make learning ubiquitous, allowing learners to gain immediate access to a wide range of location-specific information from many sources. In the areas of non-formal and informal education, museums and other public organizations are only beginning to explore how the technology can be used to improve experiences. For example, the Franklin Institute in Philadelphia has created an AR exhibit about the Terracotta Warriors [30].

For many teachers and students, the task of creating 3D models for AR is too difficult, since it requires significant technical knowledge. However, easier-to-use development kits are the goal of many firms investing in AR, so these problems should ease with time. Both Apple and Android are continually improving their native AR platforms to provide greater functionality, easier integration, and better performance. Flagship phones from most manufacturers are more than capable of running these AR platforms. Third party software developers and social media companies, such as Facebook, are working toward incorporating Augmented Reality into their platforms. These developments will accelerate the development of new applications in education and other fields.

4.1. Augmented Reality as a tool for enhancing learning

One consistent promise of AR has been in its application to formal and informal learning settings [31–33]. Indeed, several canonical learning theories describe the potential benefit of an augmented version of reality that is crafted with the intention to give context-specific, just-in-time information to a learner in an unobtrusive way. As a teaching and learning method, AR aligns well with both situated and constructivist learning theory. These theories frame the

learner within a real-world physical and social context. AR affords the type of scaffolding and social context that constructivism values. It allows for participatory and metacognitive learning processes through authentic inquiry, situated observation, peer connections, coaching, and reciprocal peer learning, in line with a situated learning theory-driven view of education [34, 35].

The Cognitive Theory of Multimedia Learning can also be used to discuss the potential benefits of AR in education [36]. This theory states that there are multiple channels (or pathways) for sensory input to be brought into working memory. People learn best by using combined words and patterns, such as images, rather than text alone [37]. In this theory, learning is an active process of filtering, selecting, and using input from different sensory channels. Since each channel can become overloaded, it is important to reduce extraneous load, which is reflected in van Merriënboer and Sweller's notion of cognitive load reduction through mixing of input channels [38]. AR can accomplish this input channel mixing in an elegant way by overlaying intelligently presented interface data onto a view of the actual environment. Germane (as opposed to extraneous) cognitive load can be achieved through a smart mix of real and virtual inputs in order to free up working memory. Limiting extraneous cognitive load paves the way for more effortless learning.

In light of its conceptual alignment with multiple learning theories, AR has been increasingly investigated by educational technology researchers. One of the most widely cited systematic reviews of AR in education is by Wu et al. [39], who reviewed 54 different education-related studies of AR that were indexed in the Social Sciences Citation Index (SSCI) database from January 2000 to October 2012. They found widely-reported positive effects on students' learning, improvements in collaborative learning, and increases in student and teacher reports of motivation. By 2013, a consensus among educational researchers was building that AR was going to play a significant role in the future of instructional design. Researchers also agreed that, at that moment, the research into instructional design using AR was still in its infancy, but AR would become one of the key emerging technologies for education over the next five years [39]. In the past five years, there has been growing scholarly interest in the application of AR in educational settings.

In the SSCI database, Chen et al. [40] found 55 papers published from 2011 to 2016 on the topic of AR in education. Of those 55, only 16 were published between 2015 and 2016, which indicates that the literature search likely took place early in 2016. The authors found widely reported increases in learning performance and motivation; deepened student engagement, improved perceived enjoyment, and positive attitudes were associated with students who learned with AR-based instruction. However, this meta-analysis suffered from some transparency and methodological issues: learning performance was not defined, and data was not presented on how categories were coded or how much support existed between categories. Though they claimed to have searched for papers published from 2011 to 2016, the authors did not report their inclusion criteria or when their search took place. They claimed to have used the search term 'Augmented Reality' and searched through 2016, but in light of other researchers' findings, one would expect many more papers to have been found for 2015 and 2016.

Akçayır and Akçayır [41] performed a more transparent and exhaustive review of AR research related to education written in English from 1980 to January 15, 2016. They located and then analyzed all of the published studies in the SSCI journal database (to the end of 2015) that

addressed educational uses of AR technology. 102 articles were discovered, 68 of which were determined to be relevant after applying the inclusion criteria detailed in the paper. They found that improved learner outcomes were the most common finding, with 'enhanced learning achievement' at the top of the mentions list. Next was motivation, followed by understanding, attitude, and satisfaction. Decreased cognitive load and increased spatial ability were interesting finds. Pedagogical contributions were the second major category of reported findings, with enhanced enjoyment and engagement mentioned most frequently. Pedagogical advantages included both collaboration opportunities and 'promotes self-learning,' which seems a bit contradictory.

The present paper undertook an update to the above literature search to see if the trend has continued since Akçayır and Akçayır's [41] search. Using the same terms and inclusion criteria in December 2017 that they used in early 2016 ('augmented reality' OR 'mixed reality' OR 'augmenting reality') yielded 74 relevant results: 34 from 2016 and 40 from 2017. A search on EBSCO Education Source using the same terms from 2016 to 2017 yielded 145 English scholarly journal results. The trend of articles reporting increases in student learning and motivation continued, and researchers seemed to agree that AR remains one of the most important developments on the educational horizon. AR is expected to achieve widespread adoption in two to three years within higher education, and four to five years in K-12 education [42].

Although all three overviews of the field presented here found similar positive educational outcomes associated with AR, there are some gaps in research that should be addressed. Wu et al. [39] caution that the educational value of AR is not solely based on the technologies themselves but also on how AR is designed, implemented, and integrated into formal and informal learning settings. In light of the fact that most AR research has been short-term, quasi-experimental, mixed-methods, or design-based research using small sample sizes, Wu et al. recommend that more rigorous studies are needed to examine the true learning effects.

Akçayır and Akçayır [41] found that some studies reported that AR decreases cognitive load, while others reported that it can actually lead to cognitive overload. They note that reasons for this difference are unclear, but it is likely to be due to design differences across studies. Similarly, there are conflicting reports of usability issues. "Whether there is a real usability issue […] that stems from inadequate technology experience, interface design errors, technical problems, or the teacher's lack of technology experience (or negative attitude) still needs to be clarified" ([41], p. 8). These discrepancies are emblematic of the broad array of design strategies implemented across different iterations of AR-based learning and of the lack of long-term, rigorous studies that compare larger samples of learner populations. Future research should aim to develop empirically-proven design principles for intelligent AR interface design that eliminate cognitive overload and facilitate ease of use.

Another continuing trend is that educational AR research has largely ignored diverse learner populations. Wu et al. [39] and Akçayır and Akçayır [41] both note that educational AR should be extended through design for different educational purposes, such as assessment and tutoring, as well as for students with special needs. It will be important, moving forward, to design AR with a focus on diverse learner populations in order to maximize its potential benefits, especially when one considers the potential power of AR as an assistive device for those with language barriers and learning disabilities [43, 44].

4.2. Augmented Reality for worker training

Education does not only occur in the classroom. Industrial training is an area that has received considerable attention by educational technologists; to wit, AR itself was first conceived of as worker training system [1]. Tools that provide customized instruction for workers have been repeatedly shown to significantly improve task performance [45]. Although seemingly futuristic, the application of AR interfaces for educating and guiding workers has been developing for quite some time. Nearly 20 years ago, AR displays were being shown to significantly improve performance on manual assembly tasks [46]. For example, the ARVIKA consortium was a German government-supported research initiative with 23 partners from the automotive, aerospace, manufacturing, and industrial research sectors that ran for 4 years from 1999 to 2003, aimed at developing AR technology to improve manufacturing [47]. AR has been shown to aid industrial and military maintenance-related tasks by increasing the speed at which components and schematics can be compared by over 56%, compared to traditional information display methods [48].

Today, the use of AR interfaces to superimpose useful training and technical information onto the real-world view of workers is gaining widespread adoption [49]. AR enables people to visualize concepts that would otherwise be challenging, in order to make use of abstract concepts and interface with complex systems [50]. One salient example of an AR interface used for abstract concept visualization is the CreativiTIC Innova SL [51]. The Innova SL can recognize individual circuit boards and overlay useful information about their components, aiding workers in industrial settings or functioning as a lab guide for engineering students. AR can also overlay instructions over real-world object targets, thereby increasing accuracy and speed [47]. Westerfield et al. [45] explored the use of an intelligent AR tutor to assist with teaching motherboard assembly procedures, rather than simply guiding people through the steps, and was able to show a 40% improvement in learning outcomes in a post-test for students who used the AR intelligent tutor.

Technically complex assembly tasks are well suited to AR interfaces. General Electric (GE) has leveraged the use of consumer-grade AR glasses, powered by custom learning software, for wind turbine assembly [52]. With the glasses, workers can view installation instructions and component diagrams in the same visual frame as the actual task, thereby increasing efficiency and reducing the number of errors. GE reports a 34% increase in productivity from the very first time the technician uses the AR interface, compared to non-augmented controls.

Many HMD AR devices, such as the Microsoft HoloLens, are being used in enterprise education settings. Thyssenkrupp Elevator have developed a custom AR interface for 24,000 service engineers using the Microsoft HoloLens, and the application has reportedly reduced the average length of service calls by a factor of four by allowing technicians to access interactive 3D schematics of the elevator systems they are working on without looking at a separate screen [53]. Technicians can also communicate and share field of view directly with support staff, in order to give situationally specific advice.

It is often the case that industry takes the reins of research for emerging assistive technologies in order to develop them for profit. This is certainly true for Augmented Reality devices; for-profit industry has driven the development of AR technology precisely because of its

demonstrable benefits to their learner populations. The issue with industry-led technology development is that it usually occurs behind closed doors, out of the light of scrutiny of competitors and of researchers who might be able to benefit from their study. Because neither GE nor Thyssenkrupp has made this research public, it is unclear whether there are any potential downsides to AR interfaces in complex assembly tasks, such as the potential for distractions and workplace accidents relating to visual field obstruction. There is clear evidence of the distracting nature of virtual objects [54]; in fact, virtual objects can be so visually captivating that they have even been used to distract patients enough to reduce the perception of chronic pain and phantom limb pain in clinical settings [55, 56]. Poorly designed AR interfaces have also been shown to be distracting to drivers [57]. In light of such findings, it is reasonable to ask if layering virtual interfaces over complex assembly tasks could lead to dangerous distraction. It will be increasingly important for AR interface designers to consider the cognitive load of information that is presented to users in order to make them as minimally invasive as possible as we move toward a future that promises an increasingly augmented view of reality.

4.3. Augmented Reality for language learners

Although the vast majority of Augmented Reality systems have been designed to assist in industrial settings and STEM-related education [40], AR has also been shown to work well in other disciplines, since it is an effective tool for situating knowledge in real contexts [58]. One promising educational application of AR technology, both inside and outside the classroom, is facilitating language acquisition [59, 60].

English is arguably the most popular second language (L2) in the world [43]. For most students who study English in their home countries, English is learned as a foreign language (EFL). Unfortunately, not everyone has the means or opportunity to study abroad in order to reap the benefits of immersion. In countries like Taiwan, Japan, and Korea, English is not widely spoken in everyday life, yet learning English remains an important priority for a large number of people there due to its prominence as the language of international commerce, entertainment, and higher education. In EFL classrooms, English teaching is not connected to daily reality. Traditional L2 instruction in these countries usually relies more on knowledge acquisition than on skill or fluency, although simply repeating words, explaining grammatical rules, and reading irrelevant text leads to low motivation and poor learning outcomes [61]. Compounding these issues, EFL students often have few opportunities to practice English outside the classroom.

Contextualizing learning improves EFL performance [62]. Augmented Reality has been shown to contextualize information [63] while providing increased opportunities for language practice and serendipitous language learning [64]. AR has also been shown to increase long-term retention in vocabulary acquisition tasks [58]. This makes AR a fitting platform for EFL learners, who lack exactly what it may be able to provide: real-world context, motivation, exposure to vocabulary, and opportunities to practice what was learned.

AR for language learning has only recently begun to receive attention in research. Ho et al. [65] studied AR as a tool for ubiquitous learning (U-learning) in EFL at the University of Taiwan. U-learning is a concept closely related to the same proliferation of powerful mobile devices that was discussed earlier in this paper, which is facilitating the rise of mixed-reality applications.

Essentially, U-learning refers to any learning environment, combined with pervasive technology, which is responsive to user needs. These environments support sharing and context-aware content and deliver personalized content to the user [66]. U-learning is a paradigm made possible by technologies like Augmented Reality that operate in a ubiquitous computing environment.

Some research on EFL U-learning using AR has been shown to improve learner outcomes. Santos et al. [58] used multimedia learning theory as a framework to apply AR to EFL settings in order to reduce cognitive load during vocabulary learning tasks. Building on the design work of previous AR language researchers [67–69], they found significant student improvements in Filipino and German vocabulary retention when implementing situated vocabulary learning by overlaying relevant words and animations over real objects found within the learner environment. Students in the study also reported higher satisfaction with their own learning outcomes and had an easier time paying attention to lessons. There appears to be promise in teaching vocabulary through the relationship between virtual data and real objects in AR, through its ability to situate information in context.

Augmented Reality language education extends beyond EFL settings; for the ever-increasing number of voluntary migrant learners and involuntary refugees, rising to the challenges of sociocultural integration and language education are of utmost importance for their success [70, 71]. Recent work by Bradley et al. [72] suggests that language learning on mobile devices can assist in easing the sociocultural transition of migrant learners. This view is supported by previous researchers [73, 74], who affirm that AR platforms allow migrant learners to practice language in context and to interact with socially normative language usage. It is also widely held that AR can be designed to increase social interactions by requiring the learner to use real-world information and locations [41, 64].

Augmented Reality games are emerging as a powerful tool for language learning. Godwin-Jones [59] offers a concise overview of the last decade of AR language learning projects, noting that while commercially available games can be made to fit pedagogical purposes, so-called 'serious games' that are designed for learning are easier to align to learning objectives. The problem is that mechanics and narrative immersion are often compromised in the pursuit of pedagogical targets, resulting in games that feel less enjoyable than commercial titles to play [75]. In addition, the financing available to create serious games is often orders of magnitude less than for big commercial titles, leading to smaller and less experienced development teams.

Commercially available AR games can be used for educational purposes. AR gaming experienced a recent spike in popularity with the viral outbreak of Niantic's Pokémon GO [76]. Niantic's first AR title, Ingress (www.ingress.com), is still popular and is scheduled for a major revamp in 2018 [77], and the company is poised to introduce the smartphone AR game Harry Potter: Wizards Unite in 2018. Recently, researchers have examined commercial mixed-reality games for language learning, pointing to various pedagogical applications and outcomes including opportunities for vocabulary learning [78], digital storytelling as a product of play and engagement [79], increased opportunities for L2 practice [59], improved learner confidence [80], and improved willingness to communicate in L2 [81].

AR games that are explicitly designed for language learning have been the subject of increasing scholarly attention. Games like Mentira [73], designed using the open source platform Augmented Reality Interactive Storytelling (ARIS), have demonstrated that place-based

mobile games can be powerful tools for situating learning by creating authentic opportunities for collaborative engagement with both language and culture. In Mentira, for example, students interact with a historical version of Los Griegos, New Mexico, in order to solve a murder mystery in Spanish. The game blends location cues, group activities, and over 70 pages of scripted dialog spoken by AR characters to create an immersive language learning experience.

Though ARIS games tend to have favorable learning outcomes documented by design-based research, place-based mobile games are limited by a lack of usability away from the specific place for which they are designed. Their limitation of relying on hard-coded content is being addressed by object recognition-based AR platforms such as WordSense [64], which overlays contextual information on everyday objects in order to create serendipitous learning experiences in mixed reality. Current technological limitations in 3D object recognition will need to be overcome in order to facilitate accurate object recognition-based AR language learning tools.

5. Benefits of Augmented Reality in education

The examples described above demonstrate the potential of Augmented Reality to enhance education and training in diverse settings. **Table 1** lists these benefits and provides a brief description and reference for each. It is clear from the table that AR can provide many powerful and unique benefits that support the claims made about its future potential.

AR benefits	Description	Example in research
Visualization	Helping students to visualize and comprehend complex, abstract data	Klopfer and Squire [12]
Contextualization	Connecting new information with the physical world in which it occurs	Ternier et al. [82]
Situation	Blurring the boundaries between inside- and outside-of-classroom activity while encouraging social interaction	Dunleavy et al. [34]
Attention	Drawing and keeping the voluntary focus of students who attend to digital information	Ab Aziz et al. [83]
Integration	Increasing opportunities for ubiquitous interactions with knowledge and incidental learning in everyday life	Vazquez et al. [64]
Motivation	Facilitating gameful design and increasing student engagement in learning tasks	Godwin-Jones [59]

Table 1. Benefits of Augmented Reality.

6. Discussion

It is worth noting that studies reporting on the use of serious games in education, anywhere on the virtuality spectrum, tend to focus solely on the positive effects rather than examining possible negative outcomes [84]. While indications of improved learning outcomes and motivation are good to see, it is important to consider that the use of technology in the classroom tends to have the typical cycle of hype, research, and disillusionment [85]. It is possible that

we may be on the cusp of disillusionment, as AR interfaces become ubiquitous in everyday life and begin to lose some of their novelty for students. It is also important to hedge discussion of positive research outcomes by considering that investigations of novel technology in educational settings may often demonstrate best-case-scenario possibilities [86]. Much of the motivational and attentional findings of AR research could be due to an uncontrolled novelty effect.

The Gartner Group, in its 2017 Hype Cycle for Emerging Technologies chart (**Figure 2**) reported that AR is moving into the 'slope of enlightenment' [87]. They offer an optimistic outlook for AR in the next five to 10 years, as much of the hype and novelty has worn off and the real work of designing useful products and integrating this new technology into the productive fabric of culture is underway. Interestingly, they note that VR is well on its way to becoming a productive tool in the next two to five years. We should view these findings cautiously, since they are simply measures of consumer expectation, corporate investment, and technology industry sentiment. However, while it is likely true that AR is five to 10 years from mass-market adoption, the applicability of these tools to create intelligent interfaces that empower learners is real.

For users who can most benefit from this technology, specific learning and accommodation needs should serve to guide researchers and designers who will be responsible for shaping the way that these new technologies will come into being.

It is important that we not only build on the successes of previous design-based educational research but recognize why certain design strategies have failed [88]. It may be beneficial to replicate and extend some of the AR designs in which researchers have invested so much time and effort over the last few years rather than 'throw the baby out with the bathwater.'

Figure 2. Years to mainstream adoption of various technologies [87].

Longer-term studies, and studies that control for time-on-task during learning, are needed in order to determine whether motivational and learning outcome findings persist and can be transferred to real-world contexts. Educational technology researchers tend to design new interventions from the ground up for their unique student-learning contexts. A key limitation of AR studies has been a lack of maintenance probes, as most studies have only measured initial learning [39]. Longer-term effects of AR learning interventions would help support findings of improved learner outcomes.

7. Conclusion

AR should not be looked at as a panacea to fix all of education, but rather a form of human-information interaction that is very effective at serving context-specific and spatially-situated virtual data. The design of each application and its appropriate alignment to learning targets will inevitably determine its efficacy in helping learners to achieve learning targets. AR educational games also rely on thoughtful design and enjoyable mechanics in order to be successful tools for learning and engagement. However, not all AR interfaces are created equal, and therefore it is not always possible to generalize findings. Regardless, there are innate features of AR that allow it to be used as an empowering platform. Taken as a whole, the corpus makes clear that AR has identifiable strengths and will clearly have an impact in the near future.

Author details

Dov Schafer and David Kaufman*

*Address all correspondence to: dkaufman@sfu.ca

Simon Fraser University, Burnaby, BC, Canada

References

[1] Caudell TP, Mizell DW. Augmented reality: An application of heads-up display technology to manual manufacturing processes. In: Proceedings of the Twenty-Fifth Hawaii International Conference on Systems Sciences; 7-10 January 1992; Kauai, HI, USA. New York, NY, USA: IEEE; 1992. pp. 659-669. DOI: 10.1109/HICSS.1992.183317

[2] Azuma R. A survey of augmented reality. Presence: Teleoperators and Virtual Environments. 1997;**6**(4):355-385. DOI: 10.1162/pres.1997.6.4.355

[3] Aztiria A, Augusto JC, Basagoiti R, Izaguirre A, Cook DJ. Discovering frequent user–environment interactions in intelligent environments. Personal and Ubiquitous Computing. 2012;**16**(1):91-103. DOI: 10.1007/s00779-011-0471-4

[4] Billinghurst M, Clark A, Lee G. A survey of augmented reality. Foundations and Trends® in Human–Computer Interaction. 2015;**8**(2-3):73-272. DOI: 10.1561/1100000049

[5] Wilson AD, Benko H. Holograms without headsets: Projected augmented reality with the RoomAlive Toolkit. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17); 6-11 May 2017; Denver, CO, USA. New York, NY, USA: ACM; 2017. pp. 425-428. DOI: 10.1145/3027063.3050433

[6] Azuma R, Baillot Y, Behringer R, Feiner S, Julier S, MacIntyre B. Recent advances in augmented reality. IEEE Computer Graphics and Applications. 2001;**21**(6):34-47. DOI: 10.1109/38.963459

[7] Milgram P, Kishino F. A taxonomy of mixed reality visual displays. IEICE Transactions on Information and Systems. 1994;**E-77D**(12):1321-1329

[8] Hurley D. Mixed Reality, Education, & Collaboration [Internet]. leveragEd Updates. 2016. Available from: https://medium.com/leveraged-technology/mixed-reality-education-collaboration-f0fc28b3e959 [Accessed: 2018-02-07]

[9] Microsoft. Microsoft HoloLens [Internet]. 2017. Available from: www.microsoft.com/microsoft-hololens/en-us [Accessed: 2017-10-16]

[10] Alphabet. Project X [Internet]. 2017. Available from: www.x.company/glass/ [Accessed: 2017-10-15]

[11] Raskar R, Low KL. Interacting with spatially augmented reality. In: Proceedings of the 1st International Conference on Computer Graphics, Virtual Reality and Visualisation (AfriGraph '01); Nov 5-7 2001; Cape Town, South Africa. New York: ACM. pp. 101-108. DOI: 10.1145/513867.513889

[12] Klopfer E, Squire K. Environmental detectives—The development of an augmented reality platform for environmental simulations. Educational Technology Research and Development. 2008;**56**(2):203-228. DOI: 10.1007/s11423-007-9037-6

[13] Li J, Besada JA, Bernardos AM, Tarrío P, Casar JR. A novel system for object pose estimation using fused vision and inertial data. Information Fusion. 2017;**33**:15-28. DOI: 10.1016/j.inffus.2016.04.006

[14] Gps.gov. GPS Accuracy [Internet]. 2016. Available from: https://www.gps.gov/systems/gps/performance/accuracy/ [Accessed: 2017-10-20]

[15] Forbes. Why Apple Will Win the Augmented Reality Race [Internet]. 2017. Available from: https://www.forbes.com/sites/allabouttherupees/2017/09/07/why-apple-will-win-the-augmented-reality-race/ [Accessed: 2017-10-15]

[16] Newcombe RA, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison AJ, Kohi P, Shotton J, Hodges S, Fitzgibbon A. KinectFusion: Real-time dense surface mapping and tracking. In: Proceedings of the 10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 2011; Oct 26-29 2011; Basel, Switzerland. New York: IEEE; 2011. pp. 127-136. DOI: 10.1109/ISMAR.2011.6092378

[17] Zhang Z. Microsoft Kinect sensor and its effect. IEEE Multimedia. 2012;**19**(2):4-10. DOI: 10.1109/MMUL.2012.24

[18] Höllerer T, Feiner S, Hallaway D, Bell B, Lanzagorta M, Brown D, Julier S, Baillot Y, Rosenblum L. User interface management techniques for collaborative mobile augmented reality. Computers & Graphics. 2001;**25**(5):799-810. DOI: 10.1016/S0097-8493(01)00122-4

[19] Sullivan J, Sullivan JW, Tyler SW. Intelligent user interfaces: Introduction. In: Luff P, editor. Intelligent User Interfaces. New York: ACM; 1994

[20] Browne D, Norman M, Riches D. Why build adaptive systems? In: Browne D, Totterdell P, Norman M, editors. Adaptive User Interfaces. London, UK: Elsevier Science; 2016. pp. 15-58

[21] Global Market Insights Inc. Market Research Report [Internet]. 2017. Available from: https://www.gminsights.com/ [Accessed: 2017-10-20]

[22] Apple Developer. ARKit [Internet]. 2017. Available from: https://developer.apple.com/arkit/ [Accessed: 2017-10-20]

[23] Google Developers. ARCore [Internet]. 2017. Available from: https://developers.google.com/ar/ [Accessed: 2017-10-20]

[24] Oculus VR. Oculus Rift [Internet]. 2017. Available from: https://www.oculus.com/rift/ [Accessed: 2017-10-20]

[25] Magic Leap Inc. Magic Leap [Internet]. 2017. Available from: www.magicleap.com/ [Accessed: 2017-10-16]

[26] Meta Company. Meta 2 [Internet]. 2017. Available from: www.metavision.com/ [Accessed: 2017-10-16]

[27] Digi-Capital. $1 billion AR/VR investment in Q4, $2.5 billion this year so far [Internet]. 2017. Available from: https://www.digi-capital.com/news/2017/11/1-billion-ar-vr-investment-in-q4-2-5-billion-this-year-so-far/#.WifPGLQ-fOQ [Accessed: 2017-10-20]

[28] Fitzmaurice GW. Situated information spaces and spatially aware palmtop computers. Communications of the ACM. 1993;**36**(7):39-49. DOI: 10.1145/159544.159566

[29] Yuen SC-Y, Yaoyuneyong G, Johnson E. Augmented reality: An overview and five directions for AR in education. Journal of Educational Technology Development and Exchange. 2011;**4**(1):119-140. DOI: 10.18785/jetde.0401.10

[30] The Franklin Institute. How To Experience Terra Cotta Warriors Augmented Reality. 2018. Available from https://www.fi.edu/mobile [Accessed: 2018-02-20]

[31] El Sayed NA, Zayed HH, Sharawy MI. ARSC: Augmented reality student card. Computers & Education. 2011;**56**(4):1045-1061. DOI: 10.1016/j.compedu.2010.10.019

[32] Ibanez M, Kloos CD, Leony D, Rueda JJ, Maroto D. Learning a foreign language in a mixed-reality environment. IEEE Internet Computing. 2011;**15**(6):44-47. DOI: 10.1109/MIC.2011.78

[33] Wei X, Weng D, Liu Y, Wang Y. Teaching based on augmented reality for a technical creative design course. Computers & Education. 2015;**81**:221-234. DOI: 10.1016/j. compedu.2014.10.017

[34] Dunleavy M, Dede C, Mitchell R. Affordances and limitations of immersive participatory augmented reality simulations for teaching and learning. Journal of Science Education and Technology. 2009;**18**(1):7-22. DOI: 10.1007/s10956-008-9119-1

[35] Squire K. From information to experience: Place-based augmented reality games as a model for learning in a globally networked society. Teachers College Record. 2010; **112**(10):2565-2602. Available from: http://www.litmedmod.ca/sites/default/files/pdf/ squire_2010_technologies_realite_augmentee.pdf [Accessed 2018-01-25]

[36] Joo-Nagata J, Abad FM, Giner JG, García-Peñalvo FJ. Augmented reality and pedestrian navigation through its implementation in m-learning and e-learning: Evaluation of an educational program in Chile. Computers & Education. 2017;111:1-7. DOI: 10.1016/j. compedu.2017.04.003

[37] Mayer RE. Multimedia Learning. Cambridge, UK: Cambridge University Press; 2009

[38] van Merriënboer JJ, Sweller J. Cognitive load theory and complex learning: Recent developments and future directions. Educational Psychology Review. 2005;**17**(2):147-177. DOI: 10.1007/s10648-005-3951-0

[39] Wu HK, Lee SW, Chang HY, Liang JC. Current status, opportunities and challenges of augmented reality in education. Computers & Education. 2013;**62**:41-49. DOI: 10.1016/j. compedu.2012.10.024

[40] Chen P, Liu X, Cheng W, Huang R. A review of using Augmented Reality in Education from 2011 to 2016. In: Popescu E, Kinshuk Khribi MK, Huang R, Jemni M, Chen N-S, Sampson DG, editors. Innovations in Smart Learning. 2017. Singapore: Springer; 2017. pp. 13-18. DOI: 10.1007/978-981-10-2419-1_2

[41] Akçayır M, Akçayır G. Advantages and challenges associated with augmented reality for education: A systematic review of the literature. Educational Research Review. 2017;**20**:1-11. DOI: 10.1016/j.edurev.2016.11.002

[42] Johnson L, Becker SA, Estrada V, Freeman A. NMC Horizon Report: 2015 Museum Edition. Austin, TX: The New Media Consortium; 2015. Available from: https://eric.ed.gov/ ?id=ED559371 Accessed 2018-01-25

[43] Liu TY. A context-aware ubiquitous learning environment for language listening and speaking. Journal of Computer Assisted Learning. 2009;**25**(6):515-527. DOI: 10.1111/ j.1365-2729.2009.00329.x

[44] McMahon DD, Cihak DF, Wright RE, Bell SM. Augmented reality for teaching science vocabulary to postsecondary education students with intellectual disabilities and autism. Journal of Research on Technology in Education. 2016;**48**(1):38-56. DOI: 10.1080/ 15391523.2015.1103149

[45] Westerfield G, Mitrovic A, Billinghurst M. Intelligent augmented reality training for motherboard assembly. International Journal of Artificial Intelligence in Education. 2015; **25**(1):157-172. DOI: 10.1007/s40593-014-0032-x

[46] Baird KM, Barfield W. Evaluating the effectiveness of augmented reality displays for a manual assembly task. Virtual Reality. 1999;**4**(4):250-259. DOI: 10.1007/BF01421808

[47] Weidenhausen J, Knoepfle C, Stricker D. Lessons learned on the way to industrial augmented reality applications, a retrospective on ARVIKA. Computers & Graphics. 2003; **27**(6):887-891. DOI: 10.1016/j.cag.2003.09.001

[48] Henderson SJ, Feiner S. Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret. In: Proceedings of 8th IEEE International Symposium on Mixed and Augmented Reality (ISMAR 2009); October 19-22 2009; Orlando, FL, USA. New York: IEEE; 2009. pp. 135-144. DOI: 10.1109/ ISMAR.2009.5336486

[49] Kloberdanz K. How Augmented Reality Glasses Are Being Used In Industry—GE [Internet]. GE Reports. 2017. Available from: https://www.ge.com/reports/looking-smart-augmented-reality-seeing-real-results-industry-today/ [Accessed: 2017-11-13]

[50] Stirenko S, Gordienko Y, Shemsedinov T, Alienin O, Kochura Y, Gordienko N, Rojbi A, Benito JR, González EA. User-Driven Intelligent Interface on the Basis of Multimodal Augmented Reality and Brain-Computer Interaction for People with Functional Disabilities [Internet]. 2017. Available from: https://arxiv.org/abs/1704.05915 [Accessed 2018-01-25]

[51] Artetxe González E, Souvestre F, López Benito JR. Augmented reality interface for E2LP: Assistance in electronic laboratories through augmented reality. In: Szewczyk R, Kaštelan I, Temerinac M, Barak M, Sruk V, editors. Embedded Engineering Education. Cham, Switzerland: Springer International Publishing; 2016. pp. 93-108. DOI: 10.1007/ 978-3-319-27540-6_6

[52] Upskill. Upskill Industrial Augmented Reality Software Platform [Internet]. 2017. Available from: https://upskill.io/ [Accessed: 2017-10-16]

[53] Erickson, C. Microsoft HoloLens Enables Thyssenkrupp to Transform the Global Elevator Industry [Internet]. 2016. Available from: https://blogs.windows.com/ devices/2016/09/15/microsoft-hololens-enables-thyssenkrupp-to-transform-the-global-elevator-industry/ [Accessed: 2017-10-20]

[54] Joseph B, Armstrong DG. Potential perils of peri-Pokémon perambulation: The dark reality of augmented reality? Oxford Medical Case Reports. 2016;**2016**(10):265-266. DOI: 10.1093/omcr/omw080

[55] Gromala D, Tong X, Shaw C, Jin W. Immersive virtual reality as a non-pharmacological analgesic for pain management: Pain distraction and pain self-modulation. In: Rodrigues J, Cardoso P, Monteiro J, Figueiredo M, editors. Handbook of Research on Human-Computer

Interfaces, Developments, and Applications. Hershey, PA: IGI Global; 2016. pp. 478-500. DOI: 10.4018/978-1-5225-0435-1.ch019

[56] Dunn J, Yeo E, Moghaddampour P, Chau B, Humbert S. Virtual and augmented reality in the treatment of phantom limb pain: A literature review. NeuroRehabilitation. 2017;**40**(4): 595-601. DOI: 10.3233/NRE-171447

[57] Smith M, Gabbard JL, Burnett G, Doutcheva N. The effects of augmented reality head-up displays on drivers' eye scan patterns, performance, and perceptions. International Journal of Mobile Human Computer Interaction. 2017;**9**(2):1-7. DOI: 10.4018/IJMHCI. 2017040101

[58] Santos ME, Taketomi T, Yamamoto G, Rodrigo MM, Sandor C, Kato H. Augmented reality as multimedia: The case for situated vocabulary learning. Research and Practice in Technology Enhanced Learning. 2016;**11**(1):1-23. DOI: 10.1186/s41039-016-0028-2

[59] Godwin-Jones R. Augmented reality and language learning: From annotated vocabulary to place-based mobile games. Language Learning & Technology. 2016;**20**(3):9-19. Available from: http://llt.msu.edu/issues/october2016/emerging.pdf [Accessed 2018-01-25]

[60] Richardson D. Exploring the potential of a location based augmented reality game for language learning. International Journal of Game-Based Learning. 2016;**6**(3):34-49. DOI: 10.4018/IJGBL.2016070103

[61] Liu PH, Tsai MK. Using augmented-reality-based mobile learning material in EFL English composition: An exploratory case study. British Journal of Educational Technology. 2013; **44**(1):E1-E4. DOI: 10.1111/j.1467-8535.2012.01302.x

[62] Sandberg J, Maris M, de Geus K. Mobile English learning: An evidence-based study with fifth graders. Computers & Education. 2011;**57**(1):1334-1347. DOI: 10.1016/j.compedu. 2011.01.015

[63] Dunleavy M, Dede C. Augmented reality teaching and learning. In: Spector M, Merrill MD, Elen J, Bishop MJ, editors. Handbook of Research on Educational Communications and Technology. New York: Springer. pp. 735-745

[64] Vazquez CD, Nyati AA, Luh A, Fu M, Aikawa T, Maes P. Serendipitous language learning in mixed reality. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI '17); 6-11 May 2017; Denver, CO. New York: ACM; 2017. pp. 2172-2179. DOI: 10.1145/3027063.3053098

[65] Ho SC, Hsieh SW, Sun PC, Chen CM. To activate English learning: Listen and speak in real life context with an AR featured u-learning system. Educational Technology & Society. 2017;**20**(2):176-187

[66] Huang YM, Chiu PS, Liu TC, Chen TS. The design and implementation of a meaningful learning-based evaluation method for ubiquitous learning. Computers & Education. 2011;**57**(4):2291-2302. DOI: 10.1016/j.compedu.2011.05.023

[67] Lin C-C, Hsiao H-S. The effects of multimedia annotations via PDA on EFL learners' vocabulary learning. In: Yu F-Y, Hirashima T, Supnithi T, Biswas G, editors. Proceedings of the 19th International Conference on Computers in Education (ICCE 2011); 28 November-2 December 2011; Chiang Mai, Thailand. Tha Khlong, Thailand: NECTEC. pp. 579-586. Available from: http://www.apsce.net/uploaded/filemanager/9f035690-cc65-4d33-aa29-0a4d953357c4.pdf [Accessed: 2018-01-28]

[68] Lin C-C, Yu Y-C. EFL learners' cognitive load of learning vocabulary on mobile phones. In: Biswas G, Wong L-H, Hirashima T, Chen W, editors. Proceedings of the 20th International Conference on Computers in Education (ICCE 2012). 26-30 November 2012; Singapore. Singapore: National Institute of Education; 2012. pp. 545-552. Available from: http://www.apsce.net/uploaded/filemanager/31fec10b-54f9-45c9-992d-b97834deb457.pdf [Accessed 2018-01-25]

[69] Lin C-C, Wu Y-C. The effects of different presentation modes of multimedia annotations on sentential listening comprehension. In: Proceedings of the 21st International Conference on Computers in Education 2013. 18-22 November 2013; Denpasar Bali, Indonesia. Jhongli City, Taiwan: Asia-Pacific Society for Computers in Education. pp. 668-678

[70] Kukulska-Hulme A, Viberg O. Mobile collaborative language learning: State of the art. British Journal of Educational Technology. 2017. DOI: 10.1111/bjet.12580

[71] Demouy V, Jones A, Kan Q, Kukulska-Hulme A, Eardley A. Why and how do distance learners use mobile devices for language learning? The EuroCALL Review. 2016;**24**(1):10-24. DOI: 10.4995/eurocall.2016.5663

[72] Bradley L, Lindström NB, Hashemi SS. Integration and language learning of newly arrived migrants using mobile technology. Journal of Interactive Media in Education. 2017;**2017**(1):1-9. DOI: 10.5334/jime.434

[73] Holden CL, Sykes JM. Leveraging mobile games for place-based language learning. International Journal of Game-Based Learning. 2011;**1**(2):1-18. DOI: 10.4018/ijgbl.2011040101

[74] Thorne, S. Language learning, ecological validity, and innovation under conditions of superdiversity. Bellaterra Journal of Teaching & Learning Language & Literature. 2013;**6**(2):1-27

[75] Bellotti F, Kapralos B, Lee K, Moreno-Ger P, Berta R. Assessment in and of serious games: An overview. Advances in Human-Computer Interaction. 2013;**2013**:1-11. DOI: 10.1155/2013/136864

[76] Niantic. Pokémon GO [Internet]. 2017. Available from: http://pokemongolive.com/en/ [Accessed: 2017-12-10]

[77] Webster A. Niantic's First AR Game Ingress is Getting a Massive Overhaul in 2018 [Internet]. The Verge. 2017. Available from: https://www.theverge.com/2017/12/2/16725884/ingress-prime-update-niantic-pokemon-go [Accessed: 2017-12-10]

[78] King A. Pokémon GO for Listening and Language Development [Internet]. 2016. Available from: https://avteducationtalk.wordpress.com/2016/07/13/pokemon-go-for-listening-and-language-development/ [Accessed: 2017-11-20]

[79] Schrock K. August 2016: Pokémon Go in the Classroom [Internet]. 2016. Available from: http://blog.discoveryeducation.com/blog/2016/07/13/pokemongo/ [Accessed: 2017-11-20]

[80] Di Serio Á, Ibáñez MB, Kloos CD. Impact of an augmented reality system on students' motivation for a visual art course. Computers & Education. 2013;**68**:586-596. DOI: 10.1016/j.compedu.2012.03.002

[81] Shea AM. Student Perceptions of a Mobile Augmented Reality Game and Willingness to Communicate in Japanese [Ed.D. Dissertation]. Malibu, CA: Pepperdine University; 2014

[82] Ternier S, Klemke R, Kalz M, van Ulzen P, Specht M. ARLearn: Augmented reality meets augmented virtuality. Journal of Universal Computer Science. 2012;**18**(15):2143-2164. DOI: 10.3217/jucs-018-15-2143

[83] Ab Aziz N, Ab Aziz K, Paul A, Yusof A, Noor NSM. Providing augmented reality based education for students with attention deficit hyperactive disorder via cloud computing: Its advantages. In: Proceedings of the 14th International Conference on Advanced Communication Technology (ICACT 2012); 19-22 February 2012; PyeonhChang, Republic of Korea. New York: IEEE; 2012. pp. 577-581

[84] Martí-Parreño J, Méndez-Ibáñez E, Alonso-Arroyo A. The use of gamification in education: A bibliometric and text mining analysis. Journal of Computer Assisted Learning. 2016;**32**(6):663-676. DOI: 10.1111/jcal.12161

[85] Cuban L. Teachers and Machines: The Classroom Use of Technology Since 1920. New York: Teachers College Press; 1986

[86] Selwyn N. Looking beyond learning: Notes towards the critical study of educational technology. Journal of Computer Assisted Learning. 2010;**26**(1):65-73. DOI: 10.1111/j.1365-2729.2009.00338.x

[87] Gartner Group. Gartner Hype Cycle for Emerging Technologies [Internet]. 2017. Available from: https://blogs.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/ [Accessed: 2017-11-10]

[88] O'Neill DK. When form follows fantasy: Lessons for learning scientists from modernist architecture and urban planning. Journal of the Learning Sciences. 2016;**25**(1):133-152. DOI: 10.1080/10508406.2015.1094736

The Today Tendency of Sentiment Classification

Vo Ngoc Phu and Vo Thi Ngoc Tran

Abstract

Sentiment classification has already been studied for many years because it has had many crucial contributions to many different fields in everyday life, such as in political activities, commodity production, and commercial activities. There have been many kinds of the sentiment analysis such as machine learning approaches, lexicon-based approaches, etc., for many years. The today tendency of the sentiment classification is as follows: (1) Processing many big data sets with shortening execution times (2) Having a high accuracy (3) Integrating flexibly and easily into many small machines or many different approaches. We will present each category in more details.

Keywords: sentiment classification, machine learning approaches, lexicon-based approaches, today tendency of the sentiment classification, big data set

1. Introduction

Many different approaches have already been developed for sentiment analysis for many years because a lot of researchers have already desired to find many optimal algorithms and optimal approaches for many surveys and commercial applications.

The sentiment classification, called opinion mining, is the computational studies of opinions, sentiments, evaluations, attitudes, appraisal, affects, views, emotions, subjectivity, etc., expressed in texts (reviews, blogs, discussions, news, comments, feedbacks, etc.)

The different approaches have been used to cross-check with each other to reform their accuracies.

One document (one sentence or one phrase) is classified into the positive polarity, the negative polarity or the neutral polarity.

The positive polarity is a polarity of a word or a phrase (a sentence or a document) which performs aspects about good, nice, like, love, delicious, happiness, enthusiasm, kindness, etc. Examples of phrases: very good, very nice, etc. Examples of sentences: "He is very handsome"; "She is very beautiful." Examples of documents: "He is very handsome. He is also good at Sports."

The negative polarity is a polarity of a word or a phrase (a sentence or a document) which expresses aspects about bad, evil, poor, ugly, wrong, inclement, foul, shabby, sinister, rotten, ill, shoddy, etc. Examples of phrases: very bad, very evil, etc. Examples of sentences: "He is very bad"; "She is very wrong." Examples of documents: "She is very bad. She is very stupid."

The neutral polarity is a polarity of a word or a phrase (a sentence or a document) which is not both the positive polarity and the negative polarity. Examples of neutral words: eat, talk, drink, etc. Examples of phrases: a bucket of water, 1 kg, and etc. Examples of documents: "He eats a banana. He drinks a glass of water."

The polarity (positive, negative, or neutral) of a sentence or a document has been identified by using many machine learning algorithms in the surveys of the sentiment classification in [1–83].

The sentiment polarity of a word or phrase (a sentence or a document) is also expressed through a valence (sentiment score or sentiment value) of this word or this phrase (this sentence or this document).

The polarity and valence of a word or a phrase in English have been calculated by using many different approaches such as many sentiment dictionaries. Besides, the polarity and sentiment value of a word or a phrase have been identified by using many similarity measures in English and Vietnamese in [49, 50, 51, 52]. In addition, according to our opinion, the polarity and sentiment score of a word or phrase of all languages (Chinese, French, etc.) can be calculated easily by using the similarity coefficients.

If the valence of a word or phrase (a sentence or a document) is greater than 0, this word or phrase (this sentence or this document) is the positive polarity. A word or phrase (a sentence or a document) is the neutral polarity if the sentiment score of this word or phrase (this sentence or this document) is as equal as 0. If the sentiment value of a word or phrase (a sentence or a document) is less than 0, this word or phrase (this sentence or this document) is the negative polarity.

Many machine learning algorithms have already had two kinds (supervised Learning and unsupervised learning) comprising a lot of algorithm groups such as: deep learning group, ensemble group, neural networks group, regularization group, rule system group, regression group, Bayesian group, decision tree group, dimensionality reduction group, instance based group, and clustering group.

The sentiment analysis has had many machine learning approaches and lexicon-based approaches.

The lexicon-based approaches comprise many dictionary-based approaches and corpus-based approaches. The corpus-based approaches include statistical and semantic.

In this chapter, we display the dictionary-based approaches and the corpus-based approaches of the sentiment classification basically; and we also present the today tendency of the sentiment analysis in more details as follows: (1) Processing many big data sets with shortening execution times (2) Having a high accuracy (3) Integrating flexibly and easily into many small machines or many different approaches, because there have been a lot of documents, reviews, discussions, blogs, news, comments, feedbacks, etc., on many websites, online news sites, and social networks.

There have also been many big corporations in the world. The corporations have had many branches in many different countries in the world. Each branch of a corporation has had thousands of employees. Therefore, the corporations have had a lot of big information and big data sets about their employees, their businesses, etc. Processing the big information and the big data sets is very difficult by using the old algorithms, the old surveys, and old applications; and sometimes the big information and the big data set cannot be processed successfully.

Thus, the researchers now find the approaches for the surveys and the commercial applications to process the big data set for shortening execution times, improve the accuracies of these approaches. In addition, they can flexibly be integrated, and easily into the small machines or the different approaches because these small machines can be used conveniently in anywhere, for any type of users, and for various purposes. In the near future, these small machines can be produced easily, and they can be very cheap and easy to carry in everywhere.

This chapter includes six sections: The first section is the Introduction section. The second section is the Approaches of the Sentiment Classification section. The third section is the Today Tendency of the Sentiment Analysis section. The fourth section is the Conclusion section. The fifth section is the Conflict of Interest section, and the sixth section is the References section.

2. The approaches of sentiment classification

This section comprises two sub-sections as follows: In the first Section 2.1, we present the lexicon-based approaches of the opinion analysis. The machine learning approaches are displayed in the second Section 2.2.

2.1. Lexicon-based approaches

The lexicon-based approaches are comprised of multiple approaches, both dictionary-based and corpus-based.

The dictionary-based approaches involve using a dictionary that contains synonyms and antonyms of a word: for example [1], this study used seed sentiment words from a dictionary.

The approaches based on the corpus find opinion words with context-specific orientations according to a seed list of opinion words, to find other opinion words in a large corpus. There are two approaches within the category of corpus-based approaches:

a. Statistical Approach (example in [2]): If a word appears intermittently amid positive texts, then its polarity is positive. If it appears frequently among negative texts, then its polarity can be considered negative. If it has equal frequencies in positive and negative texts, then it can be considered a neutral word. Seed opinion words can be found using statistical techniques. Most state-of-the-art methods are based on the observation that similar opinion words often appear together in a corpus. Thus, if two words appear together frequently within the same context, then the probability is high that they have same polarity. Therefore, the polarity of an unknown word can be determined by calculating the relative frequency of co-occurrence with another word.

b. Semantic approach (example in [3]): This principle assigns similar sentiment values to semantically-close words. These semantically-close words can be obtained by getting a list of sentiment words, iteratively expanding the initial set with synonyms and antonyms, and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word.

A lexicon-based method was used for the sentiment classification of Twitter data in [4]. The approaches were used to identify and extract sentiments from emotions and hashtags. Also used in [4] was the practice of converting non-grammatical words to grammatical words, and normalizing non-root to root words to extract sentiments.

The survey in [5] used lexicon–based classification and included two techniques: a method-of-moments estimator for word, and a Bayesian adjustment for repeated counts of the same word.

A structured approach was used in [6] for domain-dependent sentiment analysis, using lexicon expansion aided by emoticons.

The survey [7] introduced was a new approach to lexicon extraction, which can be successfully used for sentiment polarity assignment. It has been shown that the accuracy obtained from such lexicons outperforms other lexicon-based approaches.

The lexicon–based approach that [8] used was the Semantic Orientation CALculator (SO-CAL), which includes dictionaries of words annotated with their semantic orientation (polarity and strength), and incorporates intensification and negation.

The survey in [9] proposes a framework for sentiment analysis using dictionary-based approach. An approach to sentiment analysis is proposed that uses dictionary-based approach incorporating fuzzy logic.

In the research in [10], a lexicon-based approach was proposed to calculate reputation scores from Twitter. A Saudi-dialect lexicon was developed from Saudi tweets, to improve addressing the sentiment of the Arabic tweets.

The authors of [11] propose a lexicon-based approach to sentiment classification of Twitter posts. Their approach is based on the exploitation of widespread lexical resources such as SentiWordNet, WordNet-Affect, MPQA, and SenticNet.

The lexical or lexicon-based approach is a method for a teaching dictionary-based approach described by Mechael Lewis in the early 1990s in [12]. The basic concept and methods of this approach represent an idea that signifies how education involves the understanding and production of lexical phrases. This pattern of language has grammar as well as a meaningful collection of words.

Sentiment analysis performs a role in the lexicon-based approach in [13]. It plays a significant role in determining classes such as positive, negative, and neutral.

Lexicon based approach [14] is to extract and handle the sentiment as no-slang words.

The sentiments are as followed in many dictionaries which are named as lexicon based dictionaries which are: (1) Bing Liu's opinion lexicon. (2) MPQA subjectivity lexicon. (3) SentiWordNet lexicon. (4) Semantic Evaluation (SemEval).

The acronym dictionary included in [15, 16] is very helpful in expanding tweets and improve overall sentiments scores.

In [17, 18, 19], the emoticons have a different combination of symbols as different abbreviations.

The lexicon-based antonym dictionary in [20] contains set of well-lexicons, such as WordNet dictionary in English. WordNet dictionary maintains the set of lexical datasets for English words and also keeps record of semantic relationship between works.

The authors in [21–35] use the equations determining Pointwise Mutual Information (PMI) between two words wi and wj as follows:

$$PMI(wi, wj) = log_2\left(\frac{P(wi, wj)}{P(wi)xP(wj)}\right) \tag{1}$$

They use the equations determining SO (sentiment orientation) of word wi as follows:

$$SO\ (wi)\ =\ PMI(wi, positive) - PMI(wi, negative) \tag{2}$$

In [21–28], the positive and the negative of Eq. (2) in English are: positive = {good, nice, excellent, positive, fortunate, correct, superior} and negative = {bad, nasty, poor, negative, unfortunate, wrong, inferior}.

The AltaVista search engine (AVSE) is used in the PMI equations of [22, 23, 25], and the Google search engine (GSE) is used in the PMI equations of [24, 26, 28]. In addition, the authors of [24] also use German, the authors of [25] also use Macedonian, the authors of [26] also use Arabic, the authors of [27] also use Chinese, and the authors of [28] also use Spanish. In addition, the Bing search engine (BSE) is also used in [26].

With [29–32], the PMI equations are used in Chinese, not English, and Tibetan is also added in [29]. In terms of the search engine, AVSE is used in [31], and the authors of [32] use three search engines: GSE, the Yahoo search engine (YSE), and the Baidu search engine (BSE). The PMI equations are also used in Japanese with GSE in [33]. The authors in [34, 35] also use the PMI equations and Jaccard equations with GSE in English.

The Jaccard equations with GSE in English are used in [34, 35, 37]. The authors in [36, 41] use the Jaccard equations in English. The authors in [40, 42] use the Jaccard equations in Chinese. The authors in [38] use the Jaccard equations in Arabic. The Jaccard equations with the Chinese search engine (CSE) in Chinese are used in [39].

The authors in [48] use the Ochiai Measure through GSE with the AND and OR operators, to calculate the sentiment values of the words in Vietnamese. The authors in [49] use the Cosine Measure through GSE with the AND and OR operators, to identify the sentiment scores of the words in English. The authors in [50] use the Sorensen Coefficient through GSE with the AND and OR operators, to calculate the sentiment values of the words in English. The authors in [51] use the Jaccard Measure through GSE with the AND and OR operators, to calculate the sentiment values of the words in Vietnamese. The authors in [52] use the Tanimoto Coefficient through GSE with the AND and OR operators, to identify the sentiment scores of the words in English.

With the above proofs of the surveys in [21–52], according to our evaluation, all the similarity coefficients (or the similarity measures) can be applied with certainty to identify valences (or the sentiment scores) of all the words in many different languages.

2.2. Machine-learning approaches

The supervised learning algorithms and the unsupervised learning algorithms of the machine learning algorithms have been developed for the sentiment classification in **Figure 1**.

For the deep learning group of the sentiment analysis, deep learning (also known as deep structured learning or hierarchical learning) is based on learning data representations. Learning can be supervised, semi-supervised, or unsupervised. Examples of deep learning include deep neural networks, deep belief networks, and recurrent neural networks. They have been applied to many fields, including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, and drug design.

In the survey in [54], the deep learning techniques showed promising accuracy in this domain on English tweet corpus. The authors conducted the first study that applies deep learning
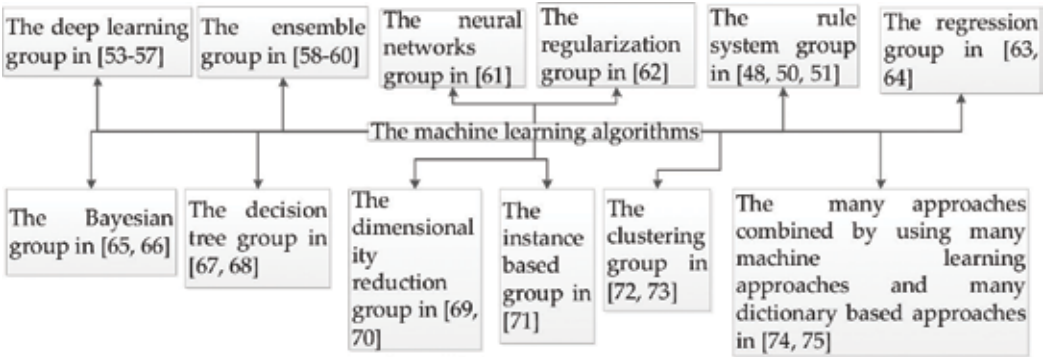
Figure 1. The machine learning algorithms.

techniques to classifying sentiment of Thai Twitter data. Two deep-learning techniques are included in the study: Long Short Term Memory (LSTM) and Dynamic Convolutional Neural Network (DCNN).

The authors of [55] used a new model to initialize the parameter weights of the convolutional neural network. They also used an unsupervised neural language model to train initial words.

Deep learning and micro-blog sentiment analysis were proposed in [56].

The authors in [57] fine-tuned a convolutional neural network (CNN) for image sentiment analysis and train a paragraph vector model for textual sentiment analysis. The authors conducted extensive experiments on both machine weakly-labeled and manually-labeled image tweets.

Ensemble approaches in statistics and machine learning use multiple learning algorithms to get better predictive performance than constituent learning algorithms. A machine learning ensemble, unlike a statistical ensemble in statistical mechanics, comprises only a concrete, finite set of alternative models, but typically allows for much more flexible structures to exist among those alternatives.

A comparative study of the effectiveness of ensemble technique for sentiment classification was proposed in [58]. This survey used the ensemble framework for sentiment classification, with the aim of efficiently integrating different feature sets and classification algorithms in order to synthesize a more accurate classification procedure. The research in [59] presents an ensemble learning method for sentiment classification of reviews. The ensemble learning framework, or stacking generalization, is introduced based on different algorithms with different settings, and compared with the majority voting. An ensemble sentiment classification strategy in [60] was applied based on Majority Vote principle of multiple classification methods, including Naive Bayes, SVM, Bayesian Network, C4.5 Decision Tree, and Random Forest algorithms.

The simplest definition of a neural network—more properly referred to as an "artificial" neural network (ANN)—is provided by the inventor of one of the first neurocomputers, Dr. Robert Hecht-Nielsen. The neural networks (NN)-based method in [61] combines the BPN and SO indexes to classify bloggers' sentiment. The NN-based method can reduce training time when classifying textual data. The NN-based method outperforms the traditional sentiment classification methods (BPN and SO index) in experimental results.

In mathematics, statistics, and computer science—particularly in the fields of machine learning and inverse problems—regularization is the process of introducing additional information in order to solve an ill-posed problem or to prevent over-fitting. The authors in [62] discussed a relation between Learning Theory and Regularization of linear ill-posed inverse problems. The authors showed that a notion of regularization (defined according to what is usually done for ill-posed inverse problems) allows derivation of learning algorithms that are consistent and that provide a fast convergence rate.

The authors in [48, 50, 51] used the rules of rule systems for the sentiment classification in Vietnamese and English.

Regression analysis in statistical modeling is a set of statistical processes for estimating the relationships among variables, and it comprises many techniques for modeling and analyzing

several variables. In regression analysis, we can see how the typical value of the dependent variable (or "criterion variable") changes when any one of the independent variables is varied while the other independent variables are held fixed. Regression analysis is a form of predictive modeling technique, which investigates the relationship between a dependent (target) and an independent variable (s) (predictor). The study in [63] analyzed the effect of using regression on sentiment classification of Twitter data.

Sentiment analysis was used in [64] to predict the Indonesian stock market. This study used the Naive Bayes and Random Forest algorithms to calculate sentiment regarding a company. The results of sentiment analysis were used to predict the company stock price. A linear regression method was used to build the prediction model.

Naïve Bayes classifiers in machine learning are a family of simple probabilistic classifiers according to Bayes' theorem, with strong (naive) independence assumptions between the features. Naïve Bayes was developed in 1950, and it was introduced under a different name to the text retrieval community in the early 1960s. It remains a popular (baseline) method for text categorization, considering the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.), with word frequencies as the features. It is competitive in this domain, with more advanced methods including support vector machines, and it also finds application in automatic medical diagnosis.

The authors in [65] explored different methods of improving the accuracy of a Naive Bayes classifier for sentiment analysis. The supervised learning algorithm was used to classify a review document as either positive or negative in [66]. The authors also improved the Naïve Bayes algorithm.

A decision tree is a tool supporting a decision, and it uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Operation research commonly uses decision trees, specifically in decision analysis, to help identify a strategy that is most likely to reach a goal; it is a popular tool in machine learning.

The authors in [67] proposed a new model using C4.5 Algorithm of a decision tree to classify semantics (positive, negative, neutral) for the English documents. A novel model using an ID3 algorithm of a decision tree was used to classify sentiments for the documents in English in [68]. This survey was based on many rules which are generated by applying the ID3 algorithm to 115,000 English sentences of our English training data set.

Dimensionality reduction, or dimension reduction in machine learning and statistics, is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. Dimensionality reduction comprises feature selection and feature extraction.

Naive Bayes and Support Vector Machine were used in [69] to analyze the sentiments of huge amount of tweets generated from Twitter users (they are stored in Twitter database). Unigram and bigram as feature extractors along with Chi2 and Singular Value Decomposition were also used for dimensionality reduction.

A novel, semi-supervised Laplacian eigenmap (SS-LE) was proposed in [70]. Redundant features were removed by decreasing its detection errors of sentiments. It enabled visualization of documents in perceptible, low-dimensional embedded space, to provide a useful tool for

text analytics. The authors evaluated the novel approach by comparing it to other dimensionality reduction methods.

Instance-based learning (memory-based learning) in machine learning is a family of learning algorithms that, instead of performing explicit generalization, compare new problem instances with instances seen in training, which have been stored in memory.

Naive Bayes, Instance Based Learning, Decision Tree, SVM, and IB1 (Instance Based Learning 1) were implemented for sentiment classification of the class of reviews from Rotten Tomatoes in [71].

Clustering data concerns a set of objects processed into classes of similar objects. One cluster is a set of data objects that are similar to each other and are not similar to objects in other clusters. A number of data clusters can be clustered, which can be identified by following experience or can be automatically identified as part of the clustering method. The authors of [72] proposed a new model for big-data sentiment classification in the parallel network environment. The authors' proposed model used the Fuzzy C-Means (FCM) method for English sentiment classification, with Hadoop MAP (M) /REDUCE (R) in Cloudera. The authors [73] proposed a new model for Big Data sentiment classification in the parallel network environment. Our new model uses the STING Algorithm (SA) (in the data mining field) for English document-level sentiment classification with Hadoop Map (M)/Reduce (R), based on the 90,000 English sentences of the training data set in a Cloudera parallel network environment—a distributed system.

Furthermore, many approaches have combined several machine-learning and dictionary-based approaches. The authors in [74] proposed a system for sentiment analysis and classification using NLP, machine-learning technique, and dictionary-based approach; our proposed methodology classifies peoples' sentiments into different polarity classes (positive, negative, and neutral).The main objective of the proposed system is to address and solve the polarity shift problem and to provide feasible solutions to the BOW model in sentiment classification; we achieved that objective by Detecting, Eliminating, and Modifying negation polarity shifter from a given text.

Two main approaches (lexical approach and machine learning) were applied to sentiment analysis in [75]. The lexicon-based method was used to create emotional dictionaries for each domain, as well as the algorithm that calculates the weight of texts. The Maximum Entropy method and the Support Vectors Machines were used in the machine learning approach to create a dictionary and an algorithm for the construction of the feature vector for the Maximum Entropy method.

3. The today tendency of the sentiment analysis

According to a testing data set and a training data set, the opinion classification has been classified into different categories in **Figure 2**.

With the category (1), the authors [49] used two testing data sets in English and they did not use any training data set. Each testing data set has the 25,000 English documents. The authors
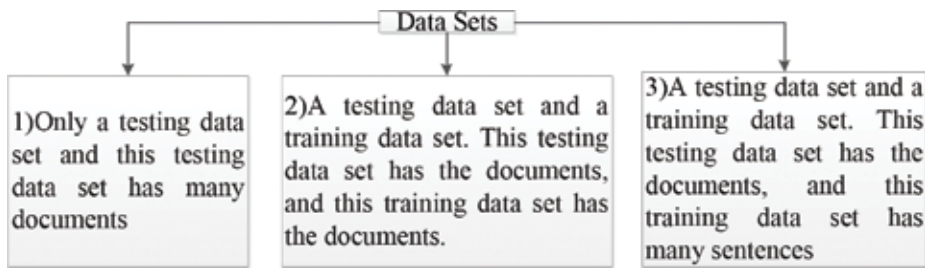
Figure 2. Categories of the sentiment classification based on the data sets.

[51] used one testing data set in Vietnamese and they did not use any training data set. The testing data set has the 30,000 Vietnamese documents. The survey [83] used one testing data set in English and it did not use any training data set. The testing data set has the 5,000,000 English documents.

The category (1) uses the lexicon-based approaches in [1–52, 77]. In addition, category (1) uses a Self-Organizing Map Algorithm—The Self-Organizing Map is based on unsupervised learning.

a. With one document of the testing data set, the SOM is used to cluster all the sentences of this document into either the positive or the negative sections on a map. The sentiment classification of this document is identified completely based on this map. There is no training data set in this category.

b. With many documents of the testing data set, the SOM is used to cluster all the documents into either the positive or the negative sections on a map. The sentiment classification of all the documents is identified completely based on this map. There is no training data set in this category.

Category (1) uses many similarity coefficients (or similarity measures) to classify one document of the testing data set into either the positive polarity or the negative polarity. According to our opinion, all the similarity measures can be used for the sentiment analysis of category (1).

In addition, category (1) also uses many rules for the sentiment classification in [48- 52], in many different languages.

The category (2) has used a testing data set and a training data set. This testing data set has the documents, and this training data set has the documents. The authors [82] used one testing data set including 1,000,000 documents and one training data set comprising 2,000,000 documents in English. This category has used many machine learning algorithms (supervised learning, unsupervised learning, semi-supervised learning, etc.). The authors in [78] use a Machine Learning algorithm, Support Vector Machines, for their sentiment classification. Latent semantic analysis (LSA) has proven to be extremely useful in information retrieval in [79]. A novel approach based on LSA and support vector machine (SVM) aims to improve the sentiment classification performance. Three machine learning approaches (Naive Bayes, maximum entropy classification, and support vector machines) were used for sentiment
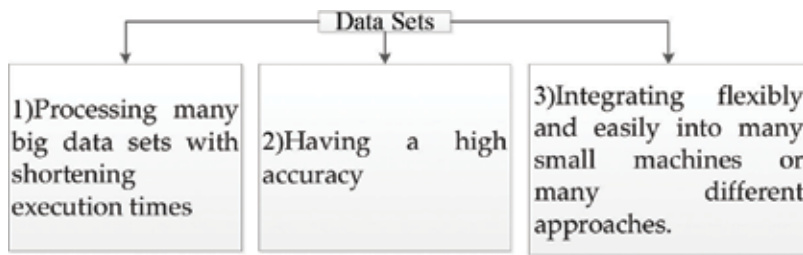
Figure 3. The today tendency of the sentiment analysis.

classification with movie reviews in [80]. The vote algorithm in [81] was used in conjunction with three classifiers, namely Naive Bayes, Support Vector Machine (SVM), and Bagging.

The category (3) uses a testing data set and a training data set. This testing data set has the documents, and this training data set has many sentences. The authors in [67] used one training data set that included 140,000 sentences and two testing data sets in English. Each testing data set has 25,000 documents. The research in [68] used one training data set that included 115,000 sentences and two testing data sets in English. Each testing data set has 25,000 documents. The authors in [72] used one training data set that included 60,000 sentences and two testing data sets in English. Each testing data set had 25,000 documents. The survey in [73] used one training data set that included 90,000 sentences and two testing data sets in English. Each testing data set had 25,000 documents.

This category also uses many machine-learning algorithms (supervised learning, unsupervised learning, semi-supervised learning, etc.). The authors in [67] used a decision tree—a C4.5 algorithm to generate many association rules for English sentiment classification. The authors in [68] also used a decision tree—an ID3 algorithm to generate many association rules for English sentiment classification. The authors in [72, 73] used the clustering algorithms of machine learning to cluster the documents of the testing data set into either the positive polarity or the negative polarity, based on the training data set. The authors in [76] used a SVM algorithm of machine learning to classify the documents of the testing data set into either the positive polarity or the negative polarity, according to the sentences of the training data set.

Paying attention to the current statuses of the economies of the world (we have presented information about big corporations, many documents, etc., in the Introduction section), we show the today tendency of the opinion analysis in **Figure 3**.

1. Processing many big data sets with shortened execution times: As we have presented the information about big corporations, many documents, etc., in the Introduction section, many old approaches (methods or models) cannot process the big data sets with certainty, or they can process the big data sets but only with long times and high costs. The processing of big data sets can be implemented in many parallel network systems. The authors' proposed model in [72] used the Fuzzy C-Means (FCM) method for English sentiment classification, with Hadoop MAP (M) /REDUCE (R) in Cloudera, a parallel network environment. The authors in [73] used a STING Algorithm for English Sentiment Classification

in A Parallel Environment. The authors of [76] used a SVM algorithm for English Semantic Classification in Parallel Environment. Furthermore, lexicon-based approaches can be performed in the distributed network systems with certainty. In the near future, there will be many small machines that can implement the parallel systems. The execution time of the proposed model is dependent on many factors: (1) the parallel network environment, such as the Cloudera system; (2) the distributed functions, such as Hadoop Map (M) and Hadoop Reduce (R); (3) the algorithms in the approach; (4) the performance of the distributed network system; (5) the number of nodes of the parallel network environment; (6) the performance of each node (each server) of the distributed environment; and (7) the sizes of the training data set and the testing data set.

2. Having high accuracy: A high accuracy is crucial for surveys and commercial applications. We can use the works of sentiment classification to cross-check in order to improve their accuracies. The accuracy of the proposed model is dependent on several factors: (1) the algorithms in the approach; (2) the testing data set and the training data set; (3) whether the documents of the testing data set are standardized carefully; and (4) whether the documents (or the sentences) of the training data set are standardized carefully.

3. Integrating flexibly and easily into many small machines or many different approaches: This category is very important for surveys, researchers, and commercial applications. The small machines used in many different fields can be conveniently used anywhere, for any type of users, and for various purposes. These small machines can be produced easily, and can be very cheap and easy to carry. The easy and flexible integration of sentiment classification into the small machines helps save a lot of time and cost. The lexicon-based approaches and the rules-based approaches can be integrated into the small machines, because the small machines have the space to store their data. In addition, the lexicons and the rules can be implemented easily in the small machines. We will not spend much time studying and implementing the surveys that currently exist.

4. Conclusion

In summary, we have presented the dictionary-based approaches and the corpus-based approaches of the sentiment classification basically; and we have also shown the today tendency of the sentiment analysis in more details.

We have displayed the information about the surveys in each section of this chapter. We have also displayed the advantages of the studies in more details.

According to the above proofs and our opinion, three tendencies of the sentiment classification will strongly have developed more and more in the near future because they have the advantages in the different fields and commercial applications.

There will be the surveys developed for the sentiment analysis.

Conflict of interest

We declare that we have no conflict of interest in this chapter.

Notes/Thanks/Other declarations

Thank Dr. Marco Antonio Aceves-Fernandez so much for inviting us to contribute this chapter to the book "Artificial Intelligence."

Author details

Vo Ngoc Phu[1]* and Vo Thi Ngoc Tran[2]

*Address all correspondence to: vongocphu03hca@gmail.com

1 Institute of Research and Development, Duy Tan University – DTU, Da Nang, Vietnam

2 School of Industrial Management (SIM), Ho Chi Minh City University of Technology - HCMUT, Vietnam National University, Ho Chi Minh City, Vietnam

References

[1] Goyal A, Daume III H. Generating semantic orientation lexicon using large data and thesaurus. WASSA '11 Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Portland, Oregon; 2011. pp. 37-43

[2] Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia; July 2002. p. 417-424

[3] Alena N, Helmut P, Mitsuru I. Recognition of affect, judgment, and appreciation in Text. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing; 2010. pp. 806-14

[4] Palanisamy P, Yadav V, Serendio HE. Simple and practical lexicon based approach to sentiment analysis. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, June 14-15; 2013. pp. 543-548

[5] Eisenstein J. Unsupervised learning for lexicon-based classification. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), the Hilton San Francisco, San Francisco, California, USA; 2017. pp. 3188-3194

[6]   Zhou Z, Zhang X, Sanderson M. (2014) sentiment analysis on twitter through topic-based lexicon expansion. In: Wang H, Sharaf MA, editors. Databases Theory and Applications. ADC 2014. Lecture Notes in Computer Science. Vol. 8506. Cham: Springer; 2014

[7]   Augustyniak L, Kajdanowicz T, Szymanski P, Tuligłowicz W, Kazienko P, Alhajj R, Szymanski B. Simpler is better? lexicon-based ensemble sentiment classification beats supervised methods. International Workshop on Curbing Collusive Cyber-gossips in Social Networks (C3-2014). Proc. IEEE/ACM Int. Conf. Advances in Social Network Analysis and Mining, ASONAM, Beijing, China; August 17, 2014

[8]   Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-based methods for sentiment analysis. Computational Linguistics. 2010;**37**(2):267-307

[9]   Hardeniya T, Borikar DA. An approach to sentiment analysis using lexicons with comparative analysis of different techniques. IOSR Journal of Computer Engineering (IOSR-JCE). 2016;**18**(3):53-57. Ver. I; e-ISSN: 2278-0661,p-ISSN: 2278-8727

[10]   Al-Hussaini H, Al-Dossari H. A lexicon-based approach to build service provider reputation from Arabic tweets in twitter. (IJACSA) International Journal of Advanced Computer Science and Applications. 2017;**8**(4)

[11]   Musto C, Semeraro G, Polignano M. A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts. Proceedings of the 8th International Workshop on Information Filtering and Retrieval, Pisa, Italy; December 10th 2014

[12]   Hamdan H, Bellot P, Bechet F. lsislif: Feature extraction and label weighting for sentiment in twitter. Proceedings of the 9th International Workshop on Semantic Evaluation, At Denver, Colorado, USA; 2015. p. 568-573

[13]   Pan Y, Li X, Shi H, Liu H. Research of methods in sentiment orientation analysis of text based on domain sentiment lexicon. Information Technology Journal. 2014;**13**(9): 1612-1621. DOI: 10.3923/itj.2014.1612.1621

[14]   Park S, Kim Y. Building thesaurus lexicon using dictionary-based approach for sentiment classification. 14th IEEE International Conference on Software Engineering Research, Management and Applications (SERA), Towson, MD, USA; 2016. pp. 39-44

[15]   Ren F, Matsumoto K. Semi-automatic creation of youth slang corpus and its application to affective computing. IEEE Transactions on Affective Computing. 2016;**7**(2):176-189

[16]   Xing L, Yuan L, Qinglin W, Yu L. An approach to sentiment analysis of short Chinese text based on SVMs. 34th IEEE Chinese Control Conference (CCC), China; 2015. pp. 9115-9120

[17]   Kundi FM, Ahmed S, Khan A, Asghar MZ. Detection and scoring of internet slangs for sentiment analysis using SentiWordNet. Life Science Journal. 2014;**11**:66-72

[18]   Huang S, Han W, Que X, Wang W. Polarity identification of sentiment words based on emoticons. 9th Conference on Computational Intelligence and Security, Emei Mountain, Sichuan Province, China; 2013. pp. 134-138

[19] Dayalani GG. Emoticon based unsupervised sentiment classifier for polarity analysis in tweets. International Journal of Engineering Research and General Science. 2014;**2**:438-445

[20] Xia R, Xu F, Zong C, Li Q, Qi Y, Li T. Dual sentiment analysis: Considering two sides of one review. IEEE Transactions on Knowledge and Data Engineering. 2015;**27**(8):2120-2133

[21] Bai A, Hammer H. Constructing sentiment lexicons in Norwegian from a large text corpus. 2014 IEEE 17th International Conference on Computational Science and Engineering, Chengdu, China; 2014

[22] Turney PD, Littman ML. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. arXiv:cs/0212012, Learning (cs.LG); Information Retrieval (cs.IR); 2002

[23] Malouf R, Mullen T. Graph-based user classification for informal online political discourse. In: Proceedings of the 1st Workshop on Information Credibility on the Web; 2017

[24] Scheible C. Sentiment translation through lexicon induction. Proceedings of the ACL 2010 Student Research Workshop, Sweden; 2010. pp. 25-30

[25] Jovanoski D, Pachovski V, Nakov P. Sentiment analysis in twitter for Macedonian. Proceedings of Recent Advances in Natural Language Processing, Bulgaria; 2015. pp. 249-257

[26] Htait A, Fournier S, Bellot P. LSIS at SemEval-2016 Task 7: Using Web Search Engines for English and Arabic Unsupervised Sentiment Intensity Prediction. Proceedings of SemEval-2016, 2016, California, p. 481-485

[27] Wan X. Co-training for cross-lingual sentiment classification. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Singapore; 2009. p. 235-243

[28] Brooke J, Tofiloski M, Taboada M. Cross-linguistic sentiment analysis: From English to Spanish. International Conference RANLP 2009, Borovets, Bulgaria; 2009. pp. 50-54

[29] Jiang T, Jiang J, Dai Y, Li A. Micro–blog emotion orientation analysis algorithm based on Tibetan and Chinese mixed text. International Symposium on Social Science (ISSS 2015); 2015

[30] Tan S, Zhang J. An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications. 2007;**34**(4):2622-2629. DOI: 10.1016/j.eswa.2007.05.028

[31] Du W, Tan S, Cheng X, Yun X. Adapting Information Bottleneck Method for Automatic Construction of Domain-oriented Sentiment Lexicon. WSDM'10, New York, USA; 2010

[32] Zhang Z, Ye Q, Zheng W, Li Y. Sentiment classification for consumer word-of-mouth in Chinese: Comparison between supervised and unsupervised approaches. The 2010 International Conference on E-Business Intelligence; 2010

[33] Wang G, Araki K. Modifying SO-PMI for Japanese weblog opinion mining by using a balancing factor and detecting neutral expressions. Proceedings of NAACL HLT 2007, Companion Volume, NY; 2007. pp. 189-192

[34] Feng S, Zhang L, Li B, Wang D, Yu v, Wong K-F. Is twitter a better corpus for measuring sentiment similarity? Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, USA; 2013. pp. 897-902

[35] An NTT, Hagiwara M. Adjective-based estimation of short sentence's impression. (KEER2014) Proceedings of the 5th Kanesi Engineering and Emotion Research; International Conference, Sweden; 2014

[36] Shikalgar NR, Dixit AM. JIBCA: Jaccard index based clustering algorithm for mining online review. International Journal of Computer Applications (0975-8887). 2014;**105**(15):1-6

[37] Ji X, Chun SA, Wei Z, Geller J. Twitter sentiment classification for measuring public health concerns. Social Network Analysis and Mining. 2015;**5**:13. DOI: 10.1007/s13278-015-0253-5

[38] Omar N, Albared M, Al-Shabi AQ, Al-Moslmi T. Ensemble of Classification algorithms for subjectivity and sentiment analysis of Arabic Customers' reviews. International Journal of Advancements in Computing Technology (IJACT). 2013;**5**

[39] Mao H, Gao P, Wang Y, Bollen J. Automatic construction of financial semantic orientation lexicon from large-scale Chinese news corpus. 7th financial risks international forum, Institut Louis Bachelier; 2014

[40] Ren Y, Kaji N, Yoshinaga N, Kitsuregaw M. Sentiment classification in under-resourced languages using graph-based semi-supervised learning methods. IEICE Transactions on Information and Systems. 2014;**E97-D**(4):790-797. DOI: 10.1587/Transinf.E97.D.1

[41] Netzer O, Feldman R, Goldenberg J, Fresko M. Mine your own business: Market-structure surveillance through text mining. Marketing Science. 2012;**31**(3):521-543

[42] Ren Y, Kaji N, Yoshinaga N, Toyoda M, Kitsuregawa M. Sentiment classification in resource-scarce languages by using label propagation. Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation, Institute of Digital Enhancement of Cognitive Processing, Waseda University; 2011. pp. 420-429

[43] Alfredo Hernández-Ugalde J, Mora-Urpí J, Rocha OJ. Genetic relationships among wild and cultivated populations of peach palm (Bactris gasipaes Kunth, Palmae): Evidence for multiple independent domestication events. Genetic Resources and Crop Evolution. 2011;**58**(4):571-583

[44] Ponomarenko JV, Bourne PE, Shindyalov IN. Building an automated classification of DNA-binding protein domains. Bioinformatics. 2002;**18**:S192-S201

[45] da Silva Meyer A, Garcia AAF, de Souza AP, de Souza CL Jr. Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (*Zea mays* L). Genetics and Molecular Biology. 2004;**27**(1):83-91

[46] Drinić SM, Nikolić A, Perić V. Cluster analysis of soybean genotypes based on RAPD markers. Proceedings. 43rd Croatian And 3rd International Symposium On Agriculture, Opatija, Croatia; 2008. pp. 367-370

[47] Tamás J, Podani J, Csontos P. An extension of presence/absence coefficients to abundance data: A new look at absence. Journal of Vegetation Science. 2001;**12**:401-410

[48] Phu VN, Chau VTN, Tran VTN, Dat ND. A Vietnamese adjective emotion dictionary based on exploitation of Vietnamese language characteristics. International Journal of Artificial Intelligence Review (AIR). 2017;**47**:67. DOI: 10.1007/CZEKANOWSKI462-017-9538-6

[49] Phu VN, Chau VTN, Dat ND, Tran VTN, Nguyen TA. A valences-totaling model for English sentiment classification. International Journal of Knowledge and Information Systems. 2017;**53**(3):579-636. DOI: 10.1007/CZEKANOWSKI115-017-1054-0

[50] Phu VN, Chau VTN, Tran VTN. Shifting semantic values of English phrases for classification. International Journal of Speech Technology (IJST). 2017;**20**(3):509-533. DOI: 10.1007/CZEKANOWSKI772-017-9420-6

[51] Phu VN, Chau VTN, Tran VTN, Dat ND, Duy KLD. A valence-totaling model for Vietnamese sentiment classification. International Journal of Evolving Systems (EVOS). 2017;**8**:47. https://doi.org/10.1007/s12530-017-9187-7

[52] Vo NP, Vo TNC, Tran VTN, Dat ND, Duy KLD. Semantic lexicons of English nouns for classification. International Journal of Evolving Systems. 2017;**8**:69. DOI: 10.1007/s12530-017-9188-6

[53] Shirani-Mehr H. Applications of Deep Learning to Sentiment Analysis of Movie Reviews. Technical Report. Stanford University; 2014

[54] Vateekul P, Koomsubha T. A study of sentiment analysis using deep learning techniques on Thai twitter data. 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen, Thailand; 2016. DOI: 10.1109/JCSSE.2016.7748849

[55] Severyn A, Moschitti A. Twitter sentiment analysis with deep convolutional neural networks. SIGIR '15 Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile; 2015. pp. 959-962

[56] Yanmei L, Yuda C. Research on Chinese micro-blog sentiment analysis based on deep learning. 8th Int. Symp. Comput. Intell. Des., Hangzhou, China; 2015. pp. 358-361

[57] You Q, Luo J, Jin H, Yang J. Joint visual-textual sentiment analysis with deep neural networks. MM '15 Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia; 2015. pp. 1071-1074

[58] Xia R, Zong C, Li S. Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences. 2015;**181**(6):1138-1152. DOI: 10.1016/j.ins.2010.11.023

[59] Su Y, Zhang Y, Ji D, Wang Y, Wu H. Ensemble learning for sentiment classification. In: Ji D, Xiao G, editors. Chinese Lexical Semantics. CLSW 2012. Lecture Notes in Computer Science. Vol. 7717. Berlin, Heidelberg: Springer; 2013

[60] Wan Y, Gao Q. An ensemble sentiment classification system of twitter data for airline services analysis. IEEE International Conference on Data Mining Workshop (ICDMW). Atlantic City, NJ, USA; 2015. DOI: 10.1109/ICDMW.2015.7

[61] Chen L-S, Liu C-H, Chiu H-J. A neural network based approach for sentiment classification in the blogosphere. Journal of Informetrics. 2011;**5**(2):313-322. DOI: 10.1016/j. joi.2011.01.003

[62] Bauer F, Pereverzev S, Rosasco L. On regularization algorithms in learning theory. Journal of Complexity. 2007;**23**(1):52-57. DOI: 10.1016/j.jco.2006.07.001

[63] Onal I, Ertugrul AM. Effect of using regression in sentiment analysis. Signal Processing and Communications Applications Conference (SIU), 2014 22nd, Trabzon, Turkey; 2014. DOI: 10.1109/SIU.2014.6830606

[64] Cakra YE, Trisedya BD. Stock price prediction using linear regression based on sentiment analysis. International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia; 2015. DOI: 10.1109/ICACSIS.2015.7415179

[65] Narayanan V, Arora I, Bhatia A. Fast and accurate sentiment classification using an enhanced naive bayes model. In: Yin H et al., editors. Intelligent Data Engineering and Automated Learning – IDEAL 2013. IDEAL 2013. Lecture Notes in Computer Science. Vol. 8206. Berlin, Heidelberg: Springer

[66] Kang H, Yoo SJ, Han D. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications. 2012;**39**(5):6000-6010. DOI: 10.1016/j.eswa.2011.11.107

[67] Phu VN, Ngoc CVT, Ngoc TVT, Duy DN. A C4.5 algorithm for english emotional classification. International Journal of Evolving Systems. 2017;**8**:1-27. DOI: 10.1007/s12530-017-9180-1

[68] Vo NP, Vo TNT, Vo TNC, Dat ND, Duy KLD. A decision tree using ID3 algorithm for English semantic analysis. International Journal of Speech Technology (IJST). 2017;**20**(3):593-613. DOI: 10.1007/s10772-017-9429-x

[69] Shyamasundar LB, Jhansi Rani P. Twitter sentiment analysis with different feature extractors and dimensionality reduction using supervised learning algorithms. India Conference (INDICON), 2016 IEEE Annual, Bangalore, India; 2016. DOI: 10.1109/INDICON.2016.7839075

[70] Kim K, Lee J. Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction. Pattern Recognition. 2014;**47**(2):758-768. DOI: 10.1016/j.patcog.2013.07.022

[71] Oswin Rahadiyan H, Virginia G, Antonius Rachmat C. Sentiment Classification of Film Reviews Using IB1. 7th International Conference on Intelligent Systems, Modeling and Simulation (ISMS). Bangkok, Thailand; 2016. DOI: 10.1109/ISMS.2016.38

[72] Phu VN, Dat ND, Tran VTN, Chau VTN, Nguyen TA. Fuzzy C-means for English sentiment classification in a distributed system. International Journal of Applied Intelligence (APIN). 2017;**46**(3):717-738. DOI: 10.1007/s10489-016-0858-z

[73] Dat ND, Phu VN, Chau VTN, Tran VTN, Nguyen TA. STING algorithm used English sentiment classification in a parallel environment. International Journal of Pattern Recognition and Artificial Intelligence. 2017;**31**(7):30. DOI: 10.1142/S0218001417500215

[74] Kolekar NV, Rao G, Dey S, Mane M, Jadhav V, Patil S. Sentiment analysis and classification using lexicon-based approach and addressing polarity shift problem. Journal of Theoretical and Applied Information Technology. 2016;**90**(1):1-8

[75] Blinov PD, Klekovkina MV, Kotelnikov EV, Pestov OA. Research of lexical approach and machine learning methods for sentiment analysis. Proceedings of. Dialogos. 2013;**2**:51-61

[76] Vo NP, Vo TNC, Vo TNT. SVM for English semantic classification in parallel environment. International Journal of Speech Technology (IJST). 2017;**20**(3):487-508. DOI: 10.1007/s10772-017-9421-5

[77] Vo NP, Phan TT. Sentiment classification using Enhanced Contextual Valence Shifters. International Conference on Asian Language Processing (IALP), Kuching, Malaysia; 2014. DOI: 10.1109/IALP.2014.6973485

[78] Kennedy A, Inkpen D. Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence. 2006;**22**(2):110-125. DOI: 10.1111/J.1467-8640.2006.00277.X

[79] Wang L, Wan Y. Sentiment classification of documents based on latent semantic analysis. In: Lin S, Huang X, editors. Advanced Research on Computer Education, Simulation and Modeling. Communications in Computer and Information Science. Vol. 176. Berlin, Heidelberg: Springer; 2011

[80] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of EMNLP. 2002:79-86

[81] Catal C, Nangir M. A sentiment classification model based on multiple classifiers. Applied Soft Computing. 2017;**50**:135-141. DOI: 10.1016/j.asoc.2016.11.022

[82] Vo NP, Vo TNT. A STING algorithm and multi-dimensional vectors used for English sentiment classification in a distributed system. American Journal of Engineering and Applied Sciences. 2017;**12**:1-19. DOI: 10.3844/ajeassp.2017

[83] Vo NP, Vo TNT. English sentiment classification using only the sentiment lexicons with a JOHNSON coefficient in a parallel network environment. American Journal of Engineering and Applied Sciences. 2017;**12**:1-28. DOI: 10.3844/ajeassp.201

A Multilevel Genetic Algorithm for the Maximum Satisfaction Problem

Noureddine Bouhmala

Additional information is available at the end of the chapter

Abstract

Genetic algorithms (GA) which belongs to the class of evolutionary algorithms are regarded as highly successful algorithms when applied to a broad range of discrete as well continuous optimization problems. This chapter introduces a hybrid approach combining genetic algorithm with the multilevel paradigm for solving the maximum constraint satisfaction problem (Max-CSP). The multilevel paradigm refers to the process of dividing large and complex problems into smaller ones, which are hopefully much easier to solve, and then work backward toward the solution of the original problem, using the solution reached from a child level as a starting solution for the parent level. The promising performances achieved by the proposed approach are demonstrated by comparisons made to solve conventional random benchmark problems.

Keywords: maximum constraint satisfaction problem, genetic algorithms, multilevel paradigm

1. Introduction

Many problems in the field of artificial intelligence can be modeled as constraint satisfaction problems (CSP). A CSP is a tuple $\langle X, D, C \rangle$ where, $X = \{x_1, x_2, ..., x_n\}$ is a finite set of variables, $D = \{D_{x_1}, D_{x_2}, ..., D_{x_n}\}$ is a finite set of domains. Thus each variable $x \in X$ has a corresponding discrete domain D_x from which it can be instantiated, and $C = \{C_1, C_2, ..., C_k\}$ is a finite set of constraints. Each k-ary constraint restricts a k-tuple of variables, $(x_1, x_2, ..., x_k)$ and specifies a subset of $D_1 \times ... \times D_k$, each element of which are values that the variables cannot take simultaneously. A solution to a CSP requires the assignment of values to each of the variables from their domains such that all the constraints on the variables are satisfied. The maximum constraint satisfaction problem (Max-CSP) aims at finding an assignment so as to maximize the number of

satisfied constraints. Max-CSP can be regarded as the generalization of CSP; the solution maximizes the number of satisfied constraints. In this chapter, attention is focused on binary CSPs, where all constraints are binary, that is, they are based on the cartesian product of the domains of two variables. However, any non-binary CSP can theoretically be converted to a binary CSP [1]. Algorithms for solving CSPs apply the so-called 1-exchange neighborhood under, which two solutions are direct neighbors if, and only if, they differ at most in the value assigned to one variable. Examples include the minimum conflict heuristic MCH [2], the break method for escaping from local minima [3], and various enhanced MCH (e.g., randomized iterative improvement of MCH called WMCH [4], MCH with tabu search [5], and evolutionary algorithms [6]). Algorithms based on assigning weights on constraints are techniques that work by introducing weights on variables or constraints in order to avoid local minima. Methods belonging to this category include genet [7], guided local search [8], the exponentiated subgradient [9], discrete Lagrangian search [10], the scaling and probabilistic smoothing [11], evolutionary algorithms combined with stepwise adaptation of weights [12], methods based on dynamically adapting weights on variables [13], or both (i.e., variables and constraints) [14]. Methods based on large neighborhood search have recently attracted several researchers for solving the CSP [15]. The central idea is to reduce the size of local search space relying on a continual relaxation (removing elements from the solution) and re-optimization (re-inserting the removed elements). Finally, the work introduced in [16] introduces a variable depth metaheuristic combing a greedy local search with a self-adaptive weighting strategy on the constraints weights.

2. Algorithm

2.1. Multilevel context

The multilevel paradigm is a simple technique, which at its core involves recursive coarsening to produce smaller and smaller problems that are easier to solve than the original one. Multilevel techniques have been developed in the period after 1960 and are among the most efficient techniques used for solving large algebraic systems arising from the discretization of partial differential equations. In recent years, it has been recognized that an effective way of enhancing metaheuristics is to use them in the multilevel context. The pseudo-code of the multilevel genetic algorithm is shown in **Algorithm 1**. **Figure 1** illustrates the multilevel paradigm used for six variables and two coarsening levels. The multilevel paradigm consists of four phases: coarsening, initial solution, uncoarsening, and refinement. The coarsening phase aims at merging the variables associated with the problem to form clusters. The clusters are used in a recursive manner to construct a hierarchy of problems each representing the original problem but with fewer degrees of freedom. The coarsest level can then be used to compute an initial solution. The solution found at the coarsest level is uncoarsened (extended to give an initial solution for the parent level) and then improved using a chosen optimization algorithm. A common feature that characterizes multilevel algorithms, is that any solution in any of the coarsened problems is a legitimate solution to the original one. Optimization algorithms using the multilevel paradigm draw their strength from coupling the refinement process across different levels.

```
input  : Problem P₀
output: Solution S_final(P₀)
begin
    level := 0;
    while Not reached the desired number of levels do
        P_level+1 := Coarsen (P_level);
        level := level + 1;
    end
    /* Initial Solution is computed at the lowest level */;
    S(P_level) = Initial Solution (P_level) ;
    while (level > 0) do
        S_start(P_level−1): = Uncoarsen (S_final(P_level));
        S_final(P_level−1) := Refine (S_start(P_level−1));
        level := level - 1;
    end
end
```

Algorithm 1. The multilevel genetic algorithm.
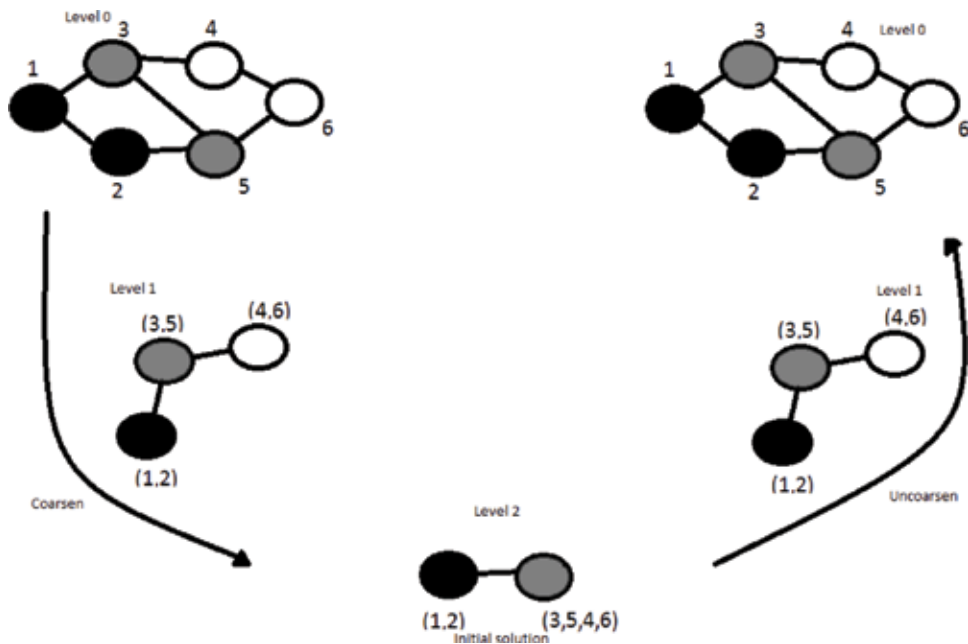
Figure 1. The different steps of the multilevel paradigm.

2.2. Multilevel genetic algorithm (GA)

GAs [17] are stochastic methods for global search and optimization and belong to the group of nature-inspired metaheuristics leading to the so-called natural computing. It is a fast-growing interdisciplinary field in which a range of techniques and methods are studied for dealing with large, complex, and dynamic problems with various sources of potential uncertainties. GAs simultaneously examine and manipulate a set of possible solutions. A gene is a part of a chromosome (solution), which is the smallest unit of genetic information. Every gene is able to assume different values called allele. All genes of an organism form a genome, which affects the appearance of an organism called phenotype. The chromosomes are encoded using a chosen representation and each can be thought of as a point in the search space of candidate solutions. Each individual is assigned a score (fitness) value that allows assessing its quality. The members of the initial population may be randomly generated or by using sophisticated mechanisms by means of which an initial population of high-quality chromosomes is produced. The reproduction operator selects (randomly or based on the individual's fitness) chromosomes from the population to be parents and enter them in a mating pool. Parent individuals are drawn from the mating pool and combined so that information is exchanged and passed to off-springs depending on the probability of the crossover operator. The new population is then subjected to mutation and enters into an intermediate population. The mutation operator acts as an element of diversity into the population and is generally applied with a low-probability to avoid disrupting crossover results. Finally, a selection scheme is used to update the population giving rise to a new generation. The individuals from the set of solutions, which is called population will evolve from generation to generation by repeated applications of an evaluation procedure that is based on genetic operators. Over many generations, the population becomes increasingly uniform until it ultimately converges to optimal or near-optimal solutions. The different steps of the multilevel weighted genetic algorithm are described as follows:

- *construction of levels*: let $G_0 = (V_0, E_0)$ be an undirected graph of vertices V and edges E. The set V denotes variables and each edge $(x_i, x_j) \in E$ implies a constraint joining the variables x_i and x_j. Given the initial graph G_0, the graph is repeatedly transformed into smaller and smaller graphs $G_1, G_2, …, G_m$ such that $|V_0| >, |V_1| >, … > |V_m|$. To coarsen a graph from G_j to G_{j+1}, a number of different techniques may be used. In this chapter, when combining a set of variables into clusters, the variables are visited in a random order. If a variable x_i has not been matched yet, then the algorithms randomly select one of its neighboring unmatched variable x_j, and a new cluster consisting of these two variables is created. Its neighbors are the combined neighbors of the merged variables x_i and x_j. Unmatched variables are simply left unmatched and copied to the next level.

- *initial assignment*: the process of constructing a hierarchy of graphs ceases as soon as the size of the coarsest graphs reaches some desired threshold. A random initial population is generated at the lowest level $G_k = (V_k, E_k)$. The chromosomes, which are assignments of values to the variables are encoded as strings of bits, the length of which is the number of variables. At the lowest level, the length of the chromosome is equal to the number of clusters. The initial solution is simply constructed by assigning to all variable in a cluster, a random value v_i. In this work, it is assumed that all variables have the same domain

(i.e., same set of values), otherwise different random values should be assigned to each variable in the cluster. All the individuals of the initial population are evaluated and assigned a fitness expressed in Eq. (1), which counts the number of constraint violations where $< (x_i, s_i), (x_j, s_j) >$ denotes the constraint between the variables x_i and x_j where x_i is assigned the value s_i from D_{x_i} and x_j is assigned the value s_j from D_{x_j}.

$$Fitness = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Violation\left(W_{i,j} < (x_i, s_i), (x_j, s_j) >\right) \tag{1}$$

- *initial weights*: the next step of the algorithm assigns a fixed amount of weight equal to 1 across all the constraints. The distribution of weights to constraints aims at forcing hard constraints with large weights to be satisfied thereby preventing the algorithm at a later stage from getting stuck at a local optimum.

- *optimization*: having computed an initial solution at the coarsest graph, GA starts the search process from the coarsest level $G_k = (V_k, E_k)$ and continues to move toward smaller levels. The motivation behind this strategy is that the order in which the levels are traversed offers a better mechanism for performing diversification and intensification. The coarsest level allows GA to view any cluster of variables as a single entity leading the search to become guided in faraway regions of the solution space and restricted to only those configurations in the solution space in which the variables grouped within a cluster are assigned the same value. As the switch from one level to another implies a decrease in the size of the neighborhood, the search is intensified around solutions from previous levels in order to reach better ones.

- *parent selection*: during the optimization, new solutions are created by combining pairs of individuals in the population and then applying a crossover operator to each chosen pair. Combining pairs of individuals can be viewed as a matching process. In the version of GA used in this work, the individuals are visited in random order. An unmatched individual i_k is matched randomly with an unmatched individual i_l.

- *genetic operators*: the task of the crossover operator is to reach regions of the search space with higher average quality. The two-point crossover operator is applied to each matched pair of individuals. The two-point crossover selects two randomly points within a chromosome and then interchanges the two parent chromosomes between these points to generate two new offspring.

- *survivor selection*: the selection acts on individuals in the current population. Based on each individual quality (fitness), it determines the next population. In the roulette method, the selection is stochastic and biased toward the best individuals. The first step is to calculate the cumulative fitness of the whole population through the sum of the fitness of all individuals. After that, the probability of selection is calculated for each individual as being $P_{Selection_i} = f_i / \sum_1^N f_i$, where f_i is the fitness of individual i.

- *updating weights*: the weights of each current violated constraint is then increased by one, whereas the newly satisfied constraints will have their weights decreased by one before the start of new generation.

- *termination condition*: the convergence of GA is supposed to be reached if the best individual remains unchanged during five consecutive generations.

- *projection*: once GA has reached the convergence criterion with respect to a child level graph $G_k = (V_k, E_k)$, the assignment reached on that level must be projected on its parent graph $G_{k-1} = (V_{k-1}, E_{k-1})$. The projection algorithm is simple; if a cluster belongs to $G_k = (V_k, E_k)$ is assigned the value vl_i, the merged pair of clusters that it represents belonging to $G_{k-1} = (V_{k-1}, E_{k-1})$ are also assigned the value vl_i,

3. Experimental results

3.1. Experimental setup

The benchmark instances were generated using model A [18] as follows: each instance is defined by the 4-tuple n, m, p_d, p_t, where n is the number of variables; m is the size of each variable's domain; p_d, the constraint density, is the proportion of pairs of variables, which have a constraint between them; and p_t, the constraint tightness, is the probability that a pair of values is inconsistent. From the $(n \times (n-1)/2)$ possible constraints, each one is independently chosen to be added in the constraint graph with the probability p_d. Given a constraint, we select with the probability p_t which value pairs become no-goods. The model A will on average have $p_d \times (n-1)/2$ constraints, each of which has on average $p_t \times m^2$ inconsistent pairs of values. For each pair of density tightness, we generate one soluble instance (i.e., at least one solution exists). Because of the stochastic nature of GA, we let each algorithm do 100 independent runs, each run with a different random seed. Many NP-complete or NP-hard problems show a phase transition point that marks the spot where we go from problems that are under-constrained and so relatively easy to solve, to problems that are over-constrained and so relatively easy to prove insoluble. Problems that are on average harder to solve occur between these two types of relatively easy problem. The values of p_d and p_t are chosen in such a way that the instances generated are within the phase transition. In order to predict the phase transition region, a formula for the constrainedness [19] of binary CSPs was defined by:

$$\kappa = \frac{n-1}{2} p_d \, log_m \left(\frac{1}{1-pt} \right). \tag{2}$$

The tests were carried out on a DELL machine with 800 MHz CPU and 2 GB of memory. The code was written in C and compiled with the GNU C compiler version 4.6. The following parameters have been fixed experimentally and are listed below:

- Population size = 50

- Stopping criteria for the coarsening phase: the reduction process stops as soon as the number of levels reaches 3. At this level, MLV-WGA generates an initial population.

- Convergence during the optimization phase: if there is no observable improvement of the fitness function of the best individual during five consecutive generations, MLV-WGA is assumed to have reached convergence and moves to a higher level.

3.2. Results

The plots in **Figures 2** and **3** compare the WGA with its multilevel variant MLV-WGA. The improvement in quality imparted by the multilevel context is immediately clear. Both WGA and MLV-WGA exhibit what is called a plateau region. A plateau region spans a region in the search space where crossover and mutation operators leave the best solution or the mean solution unchanged. However, the length of this region is shorter with MLV-WGA compared to that of WGA. The multilevel context uses the projected solution obtained at $G_{m+1}(V_{m+1}, E_{m+1})$ as the initial solution for $G_m(V_m, E_m)$ for further refinement. Even though the solution at $G_{m+1}(V_{m+1}, E_{m+1})$ is at a local minimum, the projected solution may not be at a local optimum with respect to $G_m(V_m, E_m)$. The projected assignment is already a good solution leading WGA to converge quicker within few generations to a better solution. **Tables 1**–**3** show a comparison of
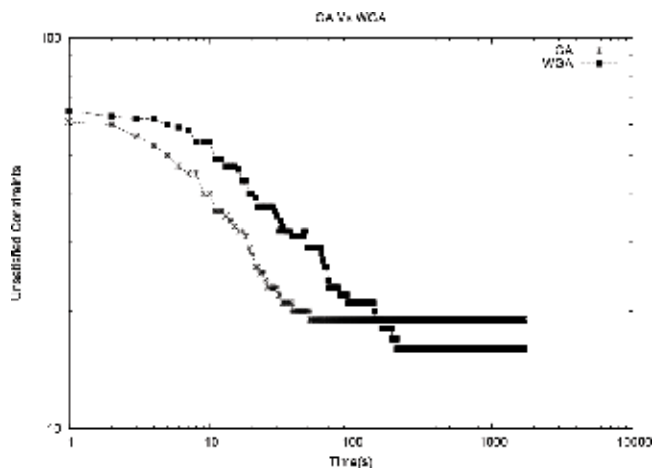
Figure 2. MLV-GA vs. GA: evolution of the mean unsatisfied constraints as a function of time. Csp-N30-DS40-C125-cd026ct063.
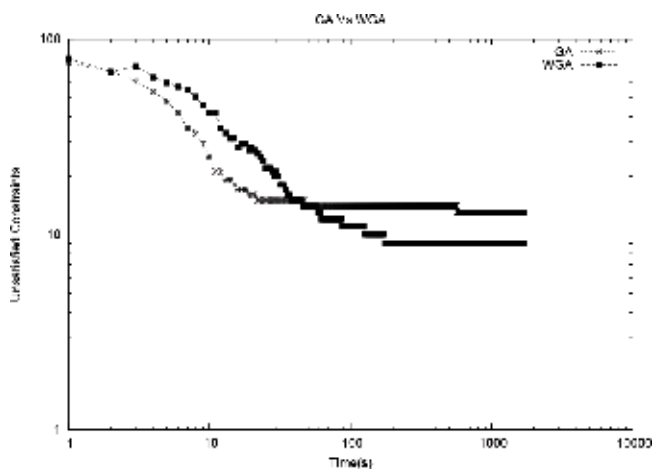
Figure 3. MLV-GA vs. GA: evolution of the mean unsatisfied constraints as a function of time. Csp-N35-DS20-C562-cd094-ct017.

Instance	MLV-WGA				WGA			
	Min	Max	Mean	RE_{av}	Min	Max	Mean	RE_{av}
N25-DS20-C36-cd-014-ct083	3	7	4.58	**0.128**	3	8	5.41	0.151
N25-DS20-C44-cd012-ct087	6	10	8.04	**0.183**	8	14	9.91	0.226
N25-DS20-C54-cd018-ct075	3	7	5.37	**0.100**	4	9	6.91	0.128
N25-DS20-C78-cd026-ct061	2	8	4.33	**0.056**	2	10	5.79	0.073
N25-DS20-C225-cd078-ct027	3	8	4.16	**0.019**	3	9	5.66	0.026
N25-DS20-C229-cd072-ct029	4	9	6.04	**0.014**	4	11	8.16	0.036
N25-DS20-C242-cd086-ct025	1	6	3.5	**0.015**	3	10	5.70	0.024
N25-DS20-C269-cd086-ct025	4	10	5.66	**0.022**	4	10	7.54	0.029
N25-DS20-C279-cd094-ct023	2	7	4.75	**0.018**	4	9	6.75	0.025
N25-DS40-C53-cd016-ct085	6	11	8.91	**0.169**	8	13	10.70	0.202
N25-DS40-C70-cd026-ct069	2	6	4.25	**0.061**	3	8	5.75	0.083
N25-DS40-C72-cd022-ct075	6	12	9	**0.125**	6	15	10.45	0.146
N25-DS40-C102-cd032-ct061	5	12	8.12	**0.080**	7	14	10.33	0.102
N25-DS40-C103-cd034-ct059	5	9	6.83	**0.067**	4	12	8.79	0.086
N25-DS40-C237-cd082-ct031	3	8	5.66	**0.024**	5	10	7.87	0.034
N25-DS40-C253-cd088-ct029	3	7	4.95	**0.020**	5	12	8.04	0.032
N25-DS40-C264-cd088-ct029	5	10	6.91	**0.027**	6	16	8.91	0.034
N25-DS40-C281-cd096-ct027	3	9	5.62	**0.020**	4	12	8.54	0.031
N25-DS40-C290-cd096-ct027	4	10	7.08	**0.025**	6	14	9	0.032

REav denotes the relative error in percent. The value in bold shows the algorithm with the lowest RE.

Table 1. MLV-WGA vs. WGA: number of variables: 25.

Instance	MLV-WGA				WGA			
	Min	Max	Mean	RE_{av}	Min	Max	Mean	RE_{av}
N30-DS20-C41-cd012-ct083	2	6	3.70	**0.026**	3	7	5.08	0.124
N30-DS20-C71-cd018-ct069	1	7	3.37	**0.048**	3	10	5.66	0.080
N30-DS20-C85-cd020-ct065	3	9	6	**0.071**	5	12	8.37	0.099
N30-DS20-C119-cd028-ct053	3	10	5.70	**0.048**	6	12	8.83	0.075
N30-DS20-C334-cd074-ct025	6	13	8.16	**0.025**	6	14	9.87	0.030
N30-DS20-C387-cd090-ct021	3	9	6.66	**0.018**	5	13	8.70	0.033
N30-DS20-C389-cd090-ct021	2	9	6.08	**0.016**	4	14	8.95	0.024
N30-DS20-C392-cd090-ct021	3	10	7.08	**0.019**	5	15	9.16	0.024
N30-DS20-C399-cd090-ct021	5	13	7.70	**0.020**	6	14	9.79	0.025
N30-DS40-C85-cd020-ct073	5	11	7.75	**0.092**	7	14	10.87	0.152
N30-DS40-C96-cd020-ct073	8	12	16	0.167	11	19	14.58	0.015

	MLV-WGA				WGA			
Instance	Min	Max	Mean	RE_{av}	Min	Max	Mean	RE_{av}
N30-DS40-C121-cd026-ct063	8	14	10.5	**0.087**	9	19	14.33	0.152
N30-DS40-C125-cd026-ct063	8	18	12.20	**0.098**	10	19	15.58	0.125
N30-DS40-C173-cd044-ct045	4	10	6.41	**0.038**	6	14	9.20	0.054
N30-DS40-C312-cd070-ct031	7	14	10.5	0.033	7	19	13.33	0.025
N30-DS40-C328-cd076-ct029	6	13	10.37	**0.032**	10	18	13.45	0.042
N30-DS40-C333-cd076-ct029	7	13	10.25	**0.031**	9	18	12.62	0.038
N30-DS40-C389-cd090-ct025	6	13	9.33	**0.024**	9	17	12.20	0.032
N30-DS40-C390-cd090-ct025	6	14	9.29	**0.024**	10	17	13	0.031

REav denotes the relative error in percent. The value in bold shows the algorithm with the lowest RE.

Table 2. MLV-WGA vs. WGA: number of variables: 30.

	MLV-WGA				WGA			
Instance	Min	Max	Mean	RE_{av}	Min	Max	Mean	RE_{av}
N40-DS20-C78-cd010-ct079	6	12	8.91	**0.115**	5	13	9.04	0.116
N40-DS20-C80-cd010-ct079	7	13	9.62	**0.121**	7	13	10.04	0.153
N40-DS20-C82-cd012-ct073	4	9	6.25	**0.073**	4	11	6.95	0.085
N40-DS20-C95-cd014-ct067	2	8	4.45	0.047	2	7	4.12	**0.044**
N40-DS20-C653-cd084-ct017	2	14	9.37	**0.015**	6	16	10.62	0.018
N40-DS20-C660-cd084-ct017	6	14	9.12	**0.014**	7	6	9.75	0.015
N40-DS20-C751-cd096-ct015	6	13	9.91	0.014	5	13	9.83	0.014
N40-DS20-C752-cd096-ct015	5	17	9.29	0.013	3	13	9.20	0.013
N40-DS20-C756-cd096-ct015	6	15	9.95	0.014	5	16	8.75	**0.012**
N40-DS40-C106-cd014-ct075	7	14	11.08	**0.105**	7	16	11.5	0.109
N40-DS40-C115-cd014-ct075	12	20	15.5	0.135	11	20	15.5	0.135
N40-DS40-C181-cd024-ct055	6	17	12.04	0.067	7	17	11.75	**0.065**
N40-DS40-C196-cd024-ct055	11	12	16.58	0.085	12	20	15.54	**0.080**
N40-DS40-C226-cd030-ct047	7	14	10.91	0.051	7	16	11.16	**0.050**
N40-DS40-C647-cd082-ct021	11	23	15.66	0.025	11	20	15.20	**0.024**
N40-DS40-C658-cd082-ct021	11	22	16.33	**0.025**	13	21	16.70	0.026
N40-DS40-C703-cd092-ct019	9	21	13.41	0.020	8	20	13.58	0.020
N40-DS40-C711-cd092-ct019	12	23	15.75	0.023	8	20	14.87	0.021
N40-DS40-C719-cd092-ct019	8	21	16.54	0.024	10	20	15.16	**0.022**

REav denotes the relative error in percent. The value in bold shows the algorithm with the lowest RE.

Table 3. MLV-WGA vs. WGA: number of variables 40.

the two algorithms. For each algorithm, the best (Min) and the worst (Max) results are given, while mean represents the average solution. MLV-WGA outperforms WGA in 53 cases out of 96, gives similar results in 20 cases, and was beaten in 23 cases. The performance of both algorithms differs significantly. The difference for the total performance is between 25 and 70% in the advantage of MLV-GA. Comparing the worst performances of both algorithms, MLV-WGA gave bad results in 15 cases, both algorithms give similar results in 8 cases, and MLV-WGA was able to perform better than WGA in 73 cases. Looking at the average results, MLV-WGA does between 16 and 41% better than WGA in 84 cases, while the differences are very marginal in the remaining cases where WGA beats MLV-WGA.

4. Conclusion

In this work, a multilevel weighted based-genetic algorithm is introduced for MAX-CSP. The results have shown that the multilevel genetic algorithm returns a better solution for the equivalent run-time for most cases compared to the standard genetic algorithm. The multilevel paradigm offers a better strategy for performing diversification and intensification. This is achieved by allowing GA to view a cluster of variables as a single entity thereby leading the search becoming guided and restricted to only those assignments in the solution space in which the variables grouped within a cluster are assigned the same value. As the size of the clusters gets smaller from one level to another, the size of the neighborhood becomes adaptive, and allows the possibility of exploring different regions in the search space while intensifying the search by exploiting the solutions from previous levels in order to reach better solutions.

Author details

Noureddine Bouhmala

Address all correspondence to: noureddine.bouhmala@usn.no

Department of Maritime Technology and Innovation, University SouthEast, Raveien, Borre, Norway

References

[1] Dechter R, Pearl J. Tree clustering for constraint networks. Artificial Intelligence. 1989;**38**: 353366

[2] Minton S, Johnson M, Philips A, Laird P. Minimizing conflicts: A heuristic repair method for constraint satisfaction and scheduling scheduling problems. Artificial Intelligence. 1992;**58**:161-205

[3] Morris P. The breakout method for escaping from local minima. In: Proceeding AAAI'93 Proceedings of the Eleventh National Conference on Artificial Intelligence. 1993. pp. 40-45

[4] Wallace R, Freuder E. Heuristic methods for over-constrained constraint satisfaction problems. In: Over-Constrained Systems. LNCS. Vol. 1106. Berlin, Germany: Springer Verlag; 1995. pp. 207-216

[5] Galinier P, Hao, J. Tabu search for maximal constraint satisfaction problems. In: Principles and Practice of Constraint Programming CP 1997. LNCS. Vol. 1330. Berlin, Germany: Springer Verlag; 1997. pp. 196-208

[6] Zhou Y, Zhou G, Zhang J. A hybrid glowworm swarm optimization algorithm for constrained engineering design problems. Applied Mathematics and Information Sciences. 2013;**7**(1):379-388

[7] Davenport A, Tsang E, Wang C, Zhu K. Genet: A connectionist architecture for solving constraint satisfaction problems by iterative improvement. In: Proceedings of the Twelth National Conference on Artificial Intelligence. 1994

[8] Voudouris C, Tsang E. Guided local search: Handbook of metaheuristics. International Series in Operation Research and Management Science. 2003;**57**:185-218

[9] Schuurmans D, Southey F, Holte E. The exponentiated subgradient algorithm for heuristic Boolean programming. In: 17th International Joint Conference on Artificial Intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers; 2001. pp. 334-341

[10] Shang E, Wah B. A discrete Lagrangian-based global-search method for solving satisfiability problems. Journal of Global Optimization. 1998;**12**(1):6199

[11] Hutter F, Tompkins D, Hoos H. Scaling and probabilistic smoothing: Efficient dynamic local search for SAT. In: Principles and Practice of Constraint Programming CP 2002. LNCS. Vol. 2470. Berlin, Germany: Springer Verlag; 2002. pp. 233-248

[12] Amante D, Marin A. Adaptive penalty weights when solving congress timetabling. Advances in Artificial Intelligence, Lectures Notes in Computer Science. 2004;**3315**:144-153

[13] Pullan W, Mascia F, Brunato M. Cooperating local search for the maximum clique problems. Journal of Heuristics. 2011;**17**:181-199

[14] Fang S, Chu Y, Qiao K, Feng X, Xu K. Combining edge weight and vertex weight for minimum vertex cover problem. In: FAW 2014. 2014. pp. 71-81

[15] Lee H, Cha S, Yu Y, Jo G. Large neighborhood search using constraint satisfaction techniques in vehicle routing problem. In: Gao Y, Japkowicz N, editors. Advances in Artificial Intelligence. Lecture Notes in Computer Science. Vol. 5549. Heidelberg: Springer Berlin; 2010. pp. 229-232

[16] Bouhmala N. A variable depth search algorithm for binary constraint satisfaction problems. Mathematical Problems in Engineering. 2015;**2015**:Article ID 637809, 10 pages. DOI: 10.1155/2015/637809

[17] Holland J. Adaptation in Natural and Artificial Systems. Ann Arbor: The University of Michigan Press; 1975

[18] Xu W. Satisfiability transition and experiments on a random constraint satisfaction problem model. International Journal of Hybrid Information Technology. 2014;**7**(2):191-202

[19] Gent IP, MacIntyre E, Prosser P, Walsh T. The constrainedness of search. In: Proceedings of the AAAI-96. 1996. pp. 246-252

Artificial Intelligence Application in Machine Condition Monitoring and Fault Diagnosis

Yasir Hassan Ali

Additional information is available at the end of the chapter

Abstract

The subject of machine condition monitoring and fault diagnosis as a part of system maintenance has gained a lot of interest due to the potential benefits to be learned from reduced maintenance budgets, enhanced productivity and improved machine availability. Artificial intelligence (AI) is a successful method of machine condition monitoring and fault diagnosis since these techniques are used as tools for routine maintenance. This chapter attempts to summarize and review the recent research and developments in the field of signal analysis through artificial intelligence in machine condition monitoring and fault diagnosis. Intelligent systems such as artificial neural network (ANN), fuzzy logic system (FLS), genetic algorithms (GA) and support vector machine (SVM) have previously developed many different methods. However, the use of acoustic emission (AE) signal analysis and AI techniques for machine condition monitoring and fault diagnosis is still rare. In the future, the applications of AI in machine condition monitoring and fault diagnosis still need more encouragement and attention due to the gap in the literature.

Keywords: artificial intelligence, machine condition monitoring, fault diagnosis

1. Introduction

In the current commercial production industries, there is an increasing trend towards the need for higher availability equipment that can work nonstop 24/7. Thus, any type of failure, even minor, cannot be accepted as it can significantly affect the cost and the production. Hence, a very accurate monitoring of the machine condition and a proper fault diagnosis of the machine failure is necessary. The machine fault diagnosis had seen a vast improvement since the maintenance was provided after the machine had developed a fault and affected the

production. After that, it developed into preventive maintenance in the past few years before all the industries started using the condition-based maintenance. Preventive maintenance can be defined as providing maintenance before the machinery faces any fault.

On the other hand, condition-based maintenance can be defined as providing maintenance depending on the data obtained from target measurements. The efficiency of this technique is measured depending on the accurate diagnostic tactics, which are fulfilled. For surviving in the current competitive market, the industries need to improve their product reliability and also reduce their production costs. The product reliability is more important for specific productions like for the aviation, nuclear and the petrochemical industries where any failure can lead to severe environmental disasters.

Currently, industries have shifted from using the condition-based (predictive) approach to the maintenance-based approach depending on the trending and the data analysis from one or more parameters that indicate the development or the presence of known failures or faults. The effective machine condition monitoring technique must be able to determine the onset of any fault in its early stages and also provide an accurate diagnosis regarding the type of the fault and its location. Ideally, the condition monitoring technique must give an overall and a detailed accurate health assessment of the equipment.

However, conventionally, it would include the aural and the visual inspection (applying all the human senses), temperature monitoring, oil analysis (known as the wear debris analysis), measurement of the vibrations and its analysis, motor current signature analysis, airborne sounds and the acoustic emission (AE) analysis. In acoustic emission analysis, the waves are sent from an emission source and transferred to the surface by the transmission medium. The low-displacement or high-frequency mechanical waves can be picked up as electronic signals. The signal strength can be increased by using a preamplifier before the data are interpreted by the AE equipment [1, 2].

Furthermore, there is a growing interest in developing new technologies to overcome the problems in condition monitoring and diagnostics of complex industrial machinery applications, which were not resolved till now. This provides excellent opportunities for the AI technology to grow continuously, with the rapid increase in the growth of intelligent information, sensor and data acquisition capabilities, combined with the rapid advances in intelligent signal processing techniques [3]. AI techniques that have been extensively used in the field of engineering include genetic algorithms (GA), support vector machine (SVM), fuzzy logic system (FLS) and artificial neural network (ANN). As compared to the common fault diagnostic approaches, the AI techniques are instrumental if they can be improved [4]. Apart from improving performance, these techniques can be easily extended and modified. These can be made adaptive by the integrating new data or information [5].

In this chapter, an attempt has been made to review the recent developments in the field of acoustic emission signal analysis for fault diagnostics of the machine based on the aforementioned AI techniques. These systems can be mutually integrated with each other and also with other traditional techniques.

2. Artificial intelligence

The AI is the system that thinks and acts like a human being. It can also imitate human behaviour. It is majorly concerned with the development of a computer's ability to engage in human-like thought processes like learning, reasoning and self-correction [6]. In the last decade, there has been a growing need in AI to solve the problems of engineering. Earlier, these problems were considered hard to be solved analytically or by using mathematical modelling and needed human intelligence [7].

Nowadays, there is an increased demand for advanced AE analysis tools. This chapter shows that many scholars have studied the detection and diagnostic of several faults by using the AE methods in AET and signal analysis. The AI techniques as mentioned earlier have also been extensively used in the field of engineering.

2.1. Artificial neural network-based fault diagnosis

Artificial neural network (ANN) is an information-processing approach. It works like the biological nervous systems like how the brain processes the information in the human body. The discussion was limited to an introduction of many components, which were involved in the ANN implementation. The network architecture or topology (including number of nodes in hidden layers, network connections, initial weight assignments and activation functions) played a key role in the ANN performance and depended on the problem at hand. **Figure 1** shows a simple ANN and its constituents. In most cases, setting the correct topology was based on a heuristic model. On the other hand, the dimensions of the input and the output spaces generally suggested the number of input and output layer nodes. Selecting the network complexity or regularization was very important [8]. When designing a neural network, there are a number of different parameters that must be decided. Some of these parameters are the number of
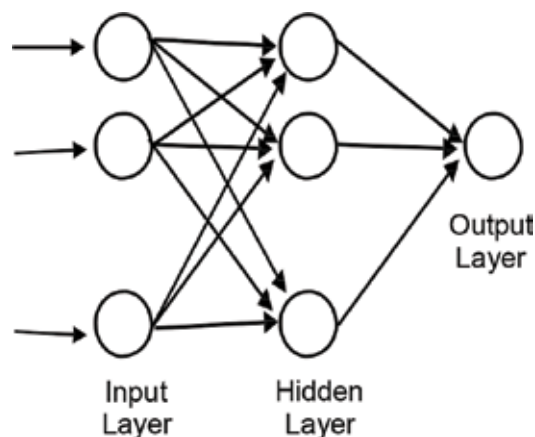
Figure 1. Architecture of a neural network.

training iterations, the number of layers, the learning rate, the number of neurons per layer and the transfer functions, and so on.

The benefit of ANN was that it had the ability to respond to an input pattern in a desirable manner after the learning phase. Previous studies have proved that the efficiency of ANN can predict the faults of machining processes. This technique was found to be very useful as it can be used in industrial automation in a more flexible manner [9]. ANN has been extensively used in health diagnosis of mechanical gear, bearing and rotating machines by using features more from vibration signals and less from the acoustic signals.

There has been an increasing demand for advanced AE analysis tools, which have the capacity to distinguish different sources of AE data. This has resulted in developing modern and more flex-ible pattern recognition software, combining traditional, graphical AE analysis and advanced unsupervised pattern recognition (UPR) and supervised pattern recognition (SPR) analysis. Application of the UPR techniques on AE data during various test cases has also increased the understanding of the damage evolution and the capacity of noise discrimination [10].

The problem of a roller with health monitoring has illustrated the effectiveness of GA for fault classification by using ANNs [8]. In this regard, Al–Balushi and Samanta have suggested a procedure to diagnose the fault of gears by wavelet transformation and ANN for AE signals. These features were taken from wavelet transformation and were used as an input to an ANN based on diagnosing approach [11]. In the fault prognosis systems, the acoustic emission and vibration signal were utilised as an input signal. Additionally, ANN was utilised as a prog-nosis system for rotating machinery failure [12]. In this way, a multiple-layer neural network was successfully used to detect the fault in the gearbox, and classification was used to utilise the supervised learning with an experimentally obtained data. The data were presented as processed vibration and acoustic emission signals [13]. The utilisation of acoustic emission for early detection of the helicopter rotor head dynamic component faults was previously studied. They analysed the stress wave of the flight-test data set by using the wavelet-based techniques for assessing the background operational noise as compared to machinery failure results. The feed-forward neural network was used as a classifier to determine the correct fight regime [14].

For solving the issues of velocity and the time differences, a new approach to AE source localization was described. This new approach to AE source location was documented on the wing spar cut-out of L-39 aircraft, as this method was used to estimate the AE source coordination by using the ANN process which extracted signal parameters [15]. Fog et al. studied the detection of the exhaust valve burn—through a four-cylinder, 500 mm bore and two-stroke marine diesel engine. This investigation comprised of monitoring three different valve conditions (normal, leak and large leak). Vibration and structure-born stress waves (AE) were monitored. The acoustic emission (AE) signal features were extracted by using principal component analysis (PCA). A feed-forward neural classifier was also used for discriminating between the three valve conditions [16].

The AE data collected during a static test of a 12-m FRP wind turbine blade was analysed and classified by using different unsupervised pattern recognition (UPR) techniques, and using

the UPR results, a supervised pattern recognition (SPR) method was trained based on the back-propagation neural network. This was applied to the AE data collection and a subsequent biaxial fatigue loading of the same blade [17].

The neural network gained attention in grinding research due to its functions of learning, interpolation and pattern recognition and classification. Different other examples of the application in the engineering field were also reported [18–22]. Aguiar et al. attempted to attain the classification of burn degrees of the surface-grinding machine, which was utilised for grinding tests with an aluminium oxide-grinding wheel and the utilisation of neural networks. The AE and power signal along with the statistics from the digital signal processing of these signals were used as inputs of the neural networks [9].

Furthermore, the ANN approach was proposed for the detection of work-piece "burn", the unwanted change in metallurgical properties of the material produced by overly aggressive or otherwise inappropriate grinding [19]. The grinding AE signals for 52100 bearing steel were collected and digested to extract feature vectors. These appeared to be more useful for ANN processing. Aguiar et al.'s work was different as it used grinding parameters as an input to the neural networks that were not tested yet in surface roughness prediction by neural networks. In addition, a higher sampling rate data acquisition system was used to get the acoustic emission and cutting power [23].

Goebel and Wright developed hybrid architecture, featuring fuzzy logic and neural networks to cope with weaknesses of traditional methods for monitoring and diagnosing an unattended milling machine. Force, spindle current and acoustic emission data were used as inputs to the neural network after they underwent some signal processing for calculating the membership functions of fuzzy relations. Additionally, fuzzy logic principles were utilised for diagnosing the system's status concerning tool wear and chatter [24]. The findings of it was encouraging to use neural network in detection and classification of work piece "burn" and surface roughness prediction revealed that AE signal from grinding machine [9, 19, 23, 24].

Impact damage is a problem that damages the composite industry. This damage may seem superficial, but it may often have negative effects on the performance of the composite structure. The conventional NDE techniques can detect the locations or the shapes of the impact damage and cannot quantify its effects on the structure. Conversely, AE records the active flaw growth when the structure gets loaded. It also measures the reduction in the structural performance produced by an impact load. AE signal analysis was used to measure the effect of impact damage on burst pressure in 5.75 inch diameter, inert propellant-filled and filament-wound pressure vessels. The AE data were collected from 15 graphite/epoxy pressure vessels featuring 5 damage states and 3 resin systems. A burst pressure prediction model was developed by correlating the AE amplitude (frequency) distribution, generated during the first pressure ramp, to 800 psig to known burst pressures using a four-layered back-propagation neural network [25].

The ANN pattern recognition technique was used for analysing the AE source signals of the pressure vessel in the site. For this purpose, a new quantitative analysis concept for AE

sources of pressure vessel was introduced by using artificial neural network classification along with raising a new method to evaluate the severity of the AE sources [26].

Macías conducted an analysis of the relationship between AE signals and the main parameters of friction stir welding (FSW) process on the basis of ANN. The AE signals were acquired by data acquisition, applied in the welding process, carrying out plates 3 mm thick of aluminium alloy. Wavelet transform (WT) was also used for the statistical and temporal parameters of the decomposition of EA signals as input for the multilayer feed-forward ANN [27].

The partial discharge (PD) detection, signal analysis and pattern identification, using AE measurements and the back-propagation (BP) ANN, were also studied. In this way, the measured signals were processed with three-dimensional patterns and short duration Fourier transforms (SDFT). The findings showed that utilisation of BP ANN with the SDFT components for the classification of the different PD patterns provided excellent results [28].

To determine the quality of feature extraction and for the ANN classifier, performances were also conducted through a series of experimentations. This helped in input data acquisition during AE experiments on the chemical process plant. These input data consisted of a set of AE power spectra. Each source input data file was subjected to preprocessing consisting of additional linear averaging in each input vector and individual amplitude normalisation by removing the mean value and division by the standard deviation of the feature. Three-layer networks using the back-propagation updating scheme were used for assessing their combined feature extraction and classification capabilities, while solving the problem of process stage recognition [29].

Until 2015, there is no study in the literature related to the estimation of oil film thickness through acoustic emission signals, so to predict and monitor oil film thickness of spur gear, a test rig was built and the gearbox was run at different speeds and load conditions. Artificial neural network (ANN) and regression models used to predict the lubricant regime depended on oil temperature, acoustic emission signals and specific film thickness (λ). Both FFBP and Elman network models were used to predict specific oil film thickness with input as AE and temperature data. The results showed that FFBP and Elman models were effective in predicting oil film thickness from acoustic emission signal and temperature, and this suggested technique attained 99.9% success in prediction and classification at high speed during training. The FFBP was better than Elman during testing and gave excellent results in prediction and classification. Thus, the architecture and topology of the network through specific systems can be used for online monitoring of oil film thickness and to predict any causes of failure of spur gear operation [30, 31].

2.2. Spiking neural network

Recently, spiking neural network (SNN) is the third-generation neural network (**Figure 2**) and has gained a lot of interest in the scientific community [32]. The SNNs became famous before the introduction of the sigmoidal or the perceptron neuron [32]. It was observed that the SNNs were very suitable for the parallel implementation in the digital hardware [32] and in the analogue hardware [33, 34].
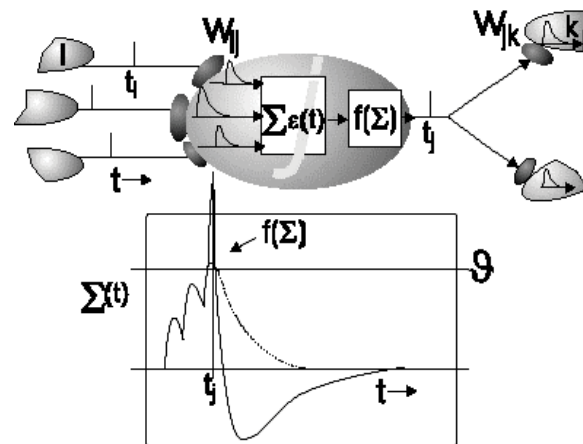
Figure 2. Spiking neural network [32].

The earlier generations of the neural networks used the analogue signals for conveying the data from one neuron to the next. This communication between the neurons in the SNNs used spikes, which was similar to the system used in the actual human neurons. The spikes could be recognized only at those instances when they had occurred. With the help of the weighted sum of the analogue input value, the earlier neuron estimated the value using the sum-specific non-linear function. The value helped in determining the delay in the spike output, which was aimed for the succeeding neuron. Generally, the spiking neuron was viewed as the leaky integrator because the target neuron integrated the spikes for a period of time and accepted the resultant integrated values used as the membrane potential. When the membrane potential value approached a specific threshold value, then, the neuron was seen to send a spike; thereafter, the membrane potential value was reset.

An increased knowledge in the information processing of the biological neurons helped in explaining many additional parameters (like the gene and the protein expression) that needed to be taken into consideration for the neurons to spike [33–35]. The additional parameters included the different physical properties of the connections [32], the likelihood of the spikes being accepted at the synapse and the emitted neurotransmitters or the open-ion channels [36, 37]. Several of the properties were modelled mathematically and were used for studying the biological neuronal system [38, 39]. The SNNs were made of the artificial neurons that communicated using the trains that were considered as the pulse-coded data [40]. The SNN was biologically acceptable, and it was seen to offer a means for the representation of the frequency, time, phase and such other features for the information processing. Moreover, the SNN possessed the ability for training the neurons for converting their spatial-temporal data to spikes (their properties include the spiking rates and spiking time). When one was selecting the neuronal model for an SNN, one needed to consider the computational efficacy and the biological credibility [40]. If it was seen that the computational efficacy was better than the biological plausibility, then the leaky integrate-and-fire (LIF) model needed to be adopted due to its cost effectiveness.

In their study, Silva et al. [41, 42] depicted the applications of the prototype decision support system for monitoring the tool wear depending on the SNN technique. This system consisted of six different components, that is, collection of data, feature extraction, multi-sensor integration, pattern recognition, tool wear estimation and the outlier detection. Their proposed architecture consisted of one built-in self-organizing neural architecture part that was based on the SNN. Their study showed that the modelling process was very efficient for classifying the tool wear level of the tool inserts with the help of the apparent weak features. Their method showed the effectiveness of using the SNN model for the tool condition monitoring, thus implying that the approach was feasible for many industrial applications, wherein a lot of noisy data are obtained. This researcher was the only one who used SNN in condition monitoring; the result showed the capability of spiking neuron networks for tool condition monitoring.

2.3. Genetic algorithm-based fault diagnosis

GA created by John Holland in the 1970s is an evolutionary algorithm which is part of the field of artificial intelligence. A genetic algorithm (GA) is a method for solving both constrained and unconstrained optimization problems based on a natural selection process that mimics biological evolution. The algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm randomly selects individuals from the current population and uses them as parents to produce the children for the next generation. Over successive generations, the population "evolves" towards an optimal solution.

As originally proposed, a simple GA mainly consists of three processes: selection, genetic operation and replacement. Description of a typical GA cycle and its high-level description are provided in **Figure 3**. The population composed of a group of chromosomes, which were the candidates for the solution. The fitness values of all chromosomes were evaluated by an
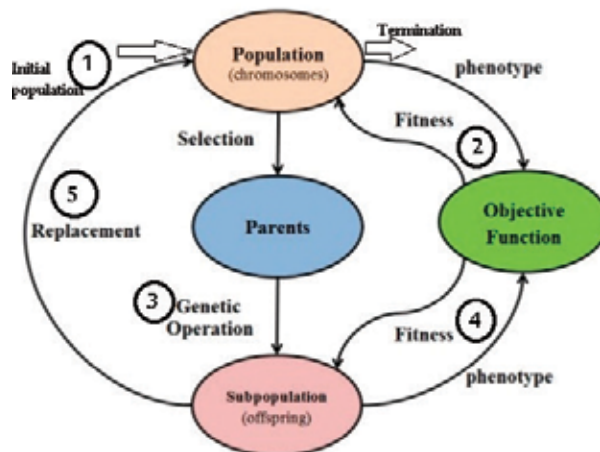
Figure 3. Genetic algorithm cycle.

objective function (performance criteria or a system's behaviour) in a decoded form (phenotype). A particular group of parents was selected from the population for generating offspring on the basis of the defined genetic operations of crossover and mutation.

The fitness of all offsprings was then evaluated using the same criterion. The chromosomes in the current population were then replaced by their offspring on the basis of a certain replacement strategy. Such a GA cycle was repeated until the termination criterion was reached [8]. Using ANNs utilised a simple problem of a roller with health monitoring to illustrate the effectiveness of GA in AE feature selection for fault classification. It revealed that utilising GAs to select an optimal feature set for a classification application of ANNs was a very powerful technique [8]. Ming applies the AE technique for bearing condition monitoring and fault diagnosis. Scales for continuous wavelet transform, wavelet-based waveform parameter selection and optimisation on the basis of genetic algorithm were the proposed selection methods [43].

The AE was monitored by utilising a data acquisition system during the process of conducting the mechanical tests on several materials. Two of the sensors were positioned directly on the specimen. AE signals were thought to be pattern vectors described by a number of writers. In this chapter, "model" data sets were generated to become closer to AE signals that were recorded during the tests. This chapter presented and validated a genetic algorithm-based approach to cluster the AE signals. Its superiority over the k-means algorithm was highlighted by the study of different "model" data sets. The genetic strategy can be characterised by a high stability and a high performance especially to cluster data sets consisting of a minority class, a cluster with signals of extreme features or a set of clusters with very different sizes [44].

2.4. Fuzzy logic-based fault diagnosis

Zadeh introduced the fuzzy logic (FL) in 1965 [45–47]. FL is a multi-valued logic that allows the intermediate values between conventional evaluations like true/false, yes/no, high/low, and so on. The FL helps in providing a variety of ways to solve a control or classification problem. Thus, this method focuses on what the system should do rather than trying to model how it works [48].

Aguiar et al. work is mentioned twice in this chapter because it contains two parts: first part used ANN for the classification of burn degrees of the surface grinding machine; in this part, a methodology was used to predict the surface roughness of advanced ceramics by using an adaptive neuro-fuzzy inference system (ANFIS). For this study, alumina work pieces were pressed and sintered into rectangular bars. The statistical data processed from the AE signal and the cutting power were also used as input data for ANFIS [9]. Cusido et al. provided approaches for a one-board fault detecting system and test program set (TPS) fault detecting system for electromechanical actuators (EMA) ball bearings by analysing the different vibration and AE signals and by using FL inference techniques [49].

Omkar et al. presented the results of fuzzy modelling to discover the problem in grinding through digital processing of the acoustic emission signals produced during the process. Fuzzy C-means (FCM) clustering was utilised in classification of the AE signal to different

sources of signals. FCM was potentially helpful in discovering the cluster among the data, when the boundaries between the subgroups overlap. AE test was conducted by using pulse, pencil and spark signal source on the surface of solid steel block. Four parameters such as event duration, peak amplitude, rise time and ring down count were measured with the help of AET 5000 system. These data were then used in training and validation of the FCM-based classification [50].

Aguiar et al. investigated the burning in the grinding process on the basis of a fuzzy model. The inputs of the models were received from the digital processing of the raw AE and cutting power signals. The parameters obtained and used in this study consisted of the mean-value deviance, grinding power and root mean square (RMS) of the acoustic emission signal [51]. Ren et al. also attempted to introduce the most successful AE model during the continuous cutting periods by using fuzzy modelling. The fuzzy identification method provided a simple way to arrive at a more definite conclusion on the basis of the information collected with the difficulty in understanding the exact physics of the machining process [52].

Recent studies used type-2 fuzzy logic in their research [53–56] because of the need to have extremely fuzzy situations to use type-2 fuzzy. If we were extending the use of FL to a higher order, then it is called type-2 FL. Hence, Ren et al. explained how type-2 TSK [Takagi–Sugeno–Kang (TSK)] fuzzy uncertainty estimation method was implemented to filter the raw AE signal directly from the AE sensor during turning process. This paper specifically focuses on filtering and capturing the uncertainty by type-2 TSK fuzzy approach on the interval of AE signal during one 10 mm cutting length [53].

Ren et al. attempted to find out the relationship between AE and tool wear. They presented an application of type-2 FL on AE signal modelling in precision manufacturing. Type-2 fuzzy modelling was used for distinguishing the AE signal in precision machining. It provided a simple way for arriving at a definite conclusion without understanding the exact physics of the machining process [54].

The knowledge about uncertainty prediction of tool life was highly essential for tool condition investigation. It was also important for taking decisions about how to maintain the machine quality. Ren et al. presented a type-2 fuzzy tool condition monitoring (TCM) system based on AE in micro-milling. In the system, type-2 FLSs were utilised for analysing the AE signal feature (SF) and choosing the most reliable ones for integration to effectively estimate the cutting tool condition through its life. The acquired results show that the type-2 fuzzy tool life estimation is in accordance with the cutting tool wear state during the micro-milling process [55].

A type-2 fuzzy analysis method was utilised to analyse the AE SFs in TCM in micro-milling process. The interval output of type-2 approach provided an interval of uncertainty associated with SFs of AE signal. The SFs with less RMSE and variation were selected to estimate the cutting tool life in the future [56]. The new philosophy for AE source localisation under high background noise was also designed. The algorithm was based on probabilistic and fuzzy-neuro principles, so AE events can be put to classification according to their energy and location probability. AE signals recorded during the stamping processes of a thin metal sheet were used for new algorithm testing [57].

Khalifa and Komarizadeh developed an efficient walnut recognition system through putting together the AE analysis, principle component analysis (PCA) and adaptive neuro-fuzzy inference system (ANFIS) classifier. This new system was tested later and classified walnuts into two classes. In the classification phase, selected statistical features were used as the input for the ANFIS classifier [58].

2.5. Support vector machine-based fault diagnosis

The support vector machine (SVM) approach was utilised in the form of a classification technique on the basis of the statistical learning theory (SLT). It was basically based on the principle of hyperplane classifier or linear separability. The main purpose of SVM was to explore a linear optimal hyperplane for maximizing the margin of separation between the two classes [59, 60].

The SVM was utilised for fault diagnosis of spur bevel gear box. This was considered to be a popular machine learning application due to its higher accuracy and for its generalization capabilities [61]. These studies also examined the fault diagnosis of low-speed bearings based on AE technique and vibration signal. Fault diagnosis was conducted by using the classification technique with the help of relevance vector machine (RVM) and SVM. The classification process provided a comparative study between RVM and SVM for fault diagnosis of low-speed bearing [62, 63].

Yu and Zhou exposed the method to classify the AE signals in composite laminates by utilising SVM. The classifier had built to achieve the identification and classification of AE signals. The results of simulation showed that SVM had the potential to effectively distinguish different acoustic emission signals and noise signals. The classification accuracy rate of grid search parameters was higher than the GA algorithm by this method [64]. Chu-Shu also revealed the method on how to classify the AE signals in composite laminates by using the SVM [65].

On the basis of a thorough review of literature, this study informs about the new approaches on the basis of hierarchical clustering and support vector machines (SVM) and are introduced to cluster AE signals and to detect P-waves for micro-crack location in the presence of noise through inducing the cracks in rock specimens during a surface instability test [66]. Thus, this chapter proposes a novel grinding wheel wear monitoring system based on discrete wavelet decomposition and SVM. The grinding signals were collected by an AE sensor [67].

Elforjani used a model to analyse the output signals of a machine while in operation and accordingly helps to set an early alarm tool that reduces the untimely replacement of components and the wasteful machine downtime. In this work, Elforjani uses three supervised machine learning techniques such as Gaussian process regression (GPR), support vector machine regression (SVMR) and multi-layer artificial neural network (ANN) model to correlate AE features with corresponding natural wear of slow-speed bearings throughout the series of laboratory experiments. Analysis of signal parameters such as root mean square (RMS) and signal intensity estimator (SIE) was done to discriminate the individual types of early damage. It was concluded that neural network models with back-propagation learning algorithm have an advantage over the other models in estimating the remaining useful life (RUL) for slow-speed bearings if the proper network structure is chosen and sufficient data are provided [68].

The development of AI technique shows a promising potential in machine condition monitoring and diagnosis, although only few articles were found in this area. However, ANN based on AE has been successfully applied to many relevant problems. It can be considered that ANN is the most new popular method in condition monitoring with AE signal. The use of fuzzy, GA and SVM in condition monitoring and fault diagnosis based on AE signal analysis still needs additional attention because of the absence of available papers. Finally, the future works will be able to find a novel idea for machine condition monitoring and fault diagnosis using AE signal analysis and AI [69].

3. Conclusion

This chapter presents a survey based on a literature review using AE signal analysis and AI techniques in machine condition monitoring and fault diagnosis. It surveys the articles with a keyword index machine condition monitoring and machine fault diagnosis using AE signal analysis and AI.

We can conclude that the classification of AE signals carries high importance in machine condition monitoring and fault diagnosis. AI has several advantages when compared to the traditional mathematical modelling and statistical analysis. This includes dispensing of the necessity for detailed system behaviour knowledge, which can be replaced by relatively simple computational methods. ANN based on AE has been successfully applied to numerous relevant problems. Therefore, we can consider that ANN is the most new popular method in AE signal analysis.

GA applications with AE signal analysis in machine condition monitoring and fault diagnosis still need more support and attention because of the lack of existing evidence. The experimental results prove that the use of fuzzy logic method is efficient and feasible. The efforts to find a novel idea must be encouraged to give more contributions in robust machine condition monitoring and fault diagnosis.

Finally, the ability to continually change and obtain a novel idea for machine condition monitoring and diagnosis using AE signal analysis and AI will be in future works.

Acknowledgements

Conflict of interest

The author declares that there is no conflict of interest regarding the publication of this chapter.

Author details

Yasir Hassan Ali

Address all correspondence to: yha2006@gmail.com

Technical College Mosul, Northern Technical University, Mosul, Iraq

References

[1] Ali YH, Rahman RA, Hamzah RIR. Acoustic emission signal analysis and artificial intelligence techniques in machine condition monitoring and fault diagnosis: A review. Journal Teknologi. 2014;**69**(2)

[2] Ali YH et al. Acoustic emission technique in condition monitoring and fault diagnosis of gears and bearings. International Journal of Academic Research. 2014;**6**(5)

[3] Mba D, Rao RB. Development of Acoustic Emission Technology for Condition Monitoring and Diagnosis of Rotating Machines; Bearings, Pumps, Gearboxes, Engines and Rotating Structures. 2006

[4] Filippetti F, Franceschini G, Tassoni C. A survey of AI techniques approach for induction machine on-line diagnosis. Proceedings of Power Electronics and Motion Control PEMC. 1996;**2**:314-318

[5] Siddique A, Yadava G, Singh B. 2003. Applications of artificial intelligence techniques for induction machine stator fault diagnostics: Review. Diagnostics for Electric Machines, Power Electronics and Drives. 2003. SDEMPED 2003. 4th IEEE International Symposium on: 29-34

[6] Kok JN, Boers EJ, Kosters WA, van der Putten P, Poel M. Artificial intelligence: Definition, trends, techniques, and cases. Artificial Intelligence. 2009

[7] Pham D, Pham P. Artificial intelligence in engineering. International Journal of Machine Tools and Manufacture. 1999;**39**(6):937-949

[8] Saxena A, Saad A. Genetic algorithms for artificial neural net-based condition monitoring system design for rotating mechanical systems. Applied Soft Computing Technologies: The Challenge of Complexity. Springer. 2006. pp. 135-149

[9] Aguiar PR, Martins CH, Marchi M, Bianchi EC. Digital Signal Processing for Acoustic Emission; 2012

[10] Kouroussis D, Anastassopoulos A, Lenain J, Proust A. Advances in Classification of Acoustic Emission Sources. Reims: Les Journées COFREND; 2001

[11] Al-Balushi K, Samanta B. Gear fault diagnostics using wavelets and artificial neural network. COMADEM 2000. 13th International Congress on Condition Monitoring and Diagnostic Engineering Management. 1001-1010. 2000

[12]   Mahamad AKB. Diagnosis, Classification and Prognosis of Rotating Machine using Artificial Intelligence. Kumamoto University; 2010

[13]   Abu-Mahfouz I. Condition monitoring of a gear box using vibration and acoustic emission based artificial neural network. SAE Transactions. 2001;**110**(6):1771-1781

[14]   Menon S, Schoess JN, Hamza R, Busch D. Wavelet-based acoustic emission detection method with adaptive thresholding. SPIE's 7th Annual International Symposium on Smart Structures and Materials. 2000. pp. 71-77

[15]   Blahacek M, Chlada M, Prevorovský Z. Acoustic emission source location based on signal features. Advanced Materials Research. 2006;**13**:77-82

[16]   Fog TL, Brown E, Hansen H, Madsen L, Sørensen P, Hansen E, Steel J, Reuben R, Pedersen P. Exhaust Valve leakage Detection in Large Marine Diesel Engines. COMADEM′ 98, 11th Int. Conf. on Condition Monitoring and Diagnostic Engineering Management. pp. 269-279

[17]   Kouroussis D, Anastassopoulos A, Vionis P, Kolovos V. Unsupervised Pattern recognition of acoustic emission from full scale testing of a wind turbine blade. Journal of Acoustic Emission (USA). 2000;**18**:217

[18]   Wang J-Z, Wang L-S, Li G-f, Zhou G-H. Prediction of surface roughness in cylindrical traverse grinding based on ALS algorithm. Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on, 1. 2005: 549-554

[19]   Wang Z, Willett P, DeAguiar PR, Webster J. Neural network detection of grinding burn from acoustic emission. International Journal of Machine Tools and Manufacture. 2001;**41**(2):283-309

[20]   Dotto FR, Aguiar PR d, Bianchi EC, Serni PJ, Thomazella R. Automatic system for thermal damage detection in manufacturing process with internet monitoring. Journal of the Brazilian Society of Mechanical Sciences and Engineering. 2006;**28**(2):153-160

[21]   Kwak J-S, Ha M-K. Neural network approach for diagnosis of grinding operation by acoustic emission and power signals. Journal of Materials Processing Technology. 2004; **147**(1):65-71

[22]   Aguiar P, França T, Bianchi E. Roughness and roundness prediction in grinding. Proceedings of the 5th CIRP International Seminar on Intelligent Computation in Manufacturing Engineering (CIRP ICME 06). 2006. pp. 25-28

[23]   Aguiar PR, Cruz CE, Paula WC, Bianchi EC. Predicting Surface Roughness in Grinding using Neural Networks

[24]   Goebel K, Wright PK. Monitoring and Diagnosing Manufacturing Processes Using a Hybrid Architecture with Neural Networks and Fuzzy Logic. EUFIT, Proceedings. 1993. 2

[25]   Walker JL, Russell SS, Workman GL, Hill EV. Neural network/acoustic emission burst pressure prediction for impact damaged composite pressure vessels. Materials Evaluation. 1997;**55**(8):903-907

[26] Shen G, Duan Q, Zhou Y, Li B, Liu Q, Li C, Jiang S. Investigation of artificial neural network pattern recognition of acoustic emission signals for pressure vessels. NDT. 2001;**23**:144-146

[27] Macías EJ, Roca AS, Fals HC, Fernández JB, Muro JC. Neural networks and acoustic emission for modelling and characterization of the friction stir welding process. Revista Iberoamericana de Automática e Informática Industrial RIAI. 2013;**10**(4):434-440

[28] Tian Y, Lewin P, Davies A, Sutton S, Swingler S. Application of acoustic emission techniques and artificial neural networks to partial discharge classification. Electrical Insulation, 2002. Conference Record of the 2002 IEEE International Symposium on: 119-123. 2002

[29] Szyszko S, Payne P. Artificial neural networks for feature extraction from acoustic emission signals. Measurements, Modelling and Imaging for Non-Destructive Testing, IEE Colloquium on. 1991:6/1-6/6

[30] Ali YH, Abd Rahman R, Hamzah RIR. Artificial neural network model for monitoring oil film regime in spur gear based on acoustic emission data. Shock and Vibration. 2015;**2015**

[31] Ali YH, Rahman RA, Hamzah RIR. Regression modeling for spur gear condition monitoring through oil film thickness based on acoustic emission signal. Modern Applied Science. 2015;**9**(8):21

[32] Ahmed FY, Yusob B, Hamed HNA. Computing with spiking neuron networks a review. International Journal of Advances in Soft Computing & Its Applications. 2014:6(1)

[33] Kasabov N. To spike or not to spike: A probabilistic spiking neuron model. Neural Networks. 2010;**23**(1):16-19

[34] Kasabov N. Integrative connectionist learning systems inspired by nature: current models, future trends and challenges. Natural Computing. 2009;**8**(2):199-218

[35] Kojima, H. and S. Katsumata. An analysis of synaptic transmission and its plasticity by glutamate receptor channel kinetics models and 2-photon laser photolysis. In: International Conference on Neural Information Processing. 2008. Springer

[36] Ikegaya Y et al. Statistical significance of precisely repeated intracellular synaptic patterns. PloS One. 2008;**3**(12):e3983

[37] Ahmed FY, Shamsuddin SM, Hashim SZM. Improved SpikeProp for using particle swarm optimization. Mathematical Problems in Engineering. 2013;**2013**

[38] Izhikevich E.. Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting (Computational Neuroscience): The MIT Press. 2006

[39] Izhikevich EM, Edelman GM. Large-scale model of mammalian thalamocortical systems. Proceedings of the National Academy of Sciences. 2008;**105**(9):3593-3598

[40] Kasabov, N. Evolving spiking neural networks and neurogenetic systems for spatio-and spectro-temporal data modelling and pattern recognition. In: IEEE World Congress on Computational Intelligence. 2012. Springer

[41] Silva RG. Condition monitoring of the cutting process using a self-organizing spiking neural network map. Journal of Intelligent Manufacturing. 2010;**21**(6):823-829

[42] Silva RG, Wilcox S, Araújo AA. Multi-sensor condition monitoring using spiking neuron networks. In: IADIS international conference applied computing 2007. 2007

[43] Ming ZX. Application of Acoustic Emission Technique in Fault Diagnostics of Rolling Bearing. Master's thesis. Tsinghua University, Beijing, Haidian; 2006

[44] Sibil A, Godin N, R'Mili M, Maillet E, Fantozzi G. Optimization of acoustic emission data clustering by a genetic algorithm method. Journal of Nondestructive Evaluation. 2012;**31**(2):169-180

[45] Zadeh LA. Fuzzy sets. Information and Control. 1965;**8**(3):338-353

[46] Zadeh LA. Outline of a new approach to the analysis of complex systems and decision processes. IEEE Transactions on Systems, Man and Cybernetics. 1973;(1):28-44

[47] Zadeh LA. Fuzzy algorithms. Information and Control. 1968;**12**(2):94-102

[48] Hellmann M. Fuzzy Logic Introduction" a Laboratoire Antennes Radar Telecom. FRE CNRS. 2272

[49] Cusido J, Delgado M, Navarro L, Sala V, Romeral L. EMA fault detection using fuzzy inference tools. AUTOTESTCON, 2010 IEEE. 2010: 1-6

[50] Omkar S, Suresh S, Raghavendra T, Mani V. Acoustic emission signal classification using fuzzy C-means clustering. Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on, 4. 2002:1827-1831

[51] De Aguiar PR, Bianchi EC, Canarim RC. Monitoring of Grinding Burn by Acoustic Emission

[52] Ren Q, Baron L, Balazinski M. Fuzzy identification of cutting acoustic emission with extended subtractive cluster analysis. Nonlinear Dynamics. 2012;**67**(4):2599-2608

[53] Ren Q, Baron L, Balazinski M. Application of type-2 fuzzy estimation on uncertainty in machining: An approach on acoustic emission during turning process. Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American. 2009: 1-6

[54] Ren Q, Baron L, Balazinski M. Type-2 fuzzy modeling for acoustic emission signal in precision manufacturing. Modelling and Simulation in Engineering. 2011;**2011**:17

[55] Ren Q, Balazinski M, Baron L, Jemielniak K, Botez R, Achiche S. Type-2 fuzzy tool condition monitoring system based on acoustic emission in micromilling. Information Sciences. 2014;**255**:121-134

[56] Ren Q, Baron L, Balazinski M, Jemielniak K. Acoustic emission signal feature analysis using type-2 fuzzy logic system. Fuzzy Information Processing Society (NAFIPS), 2010 Annual Meeting of the North American. 2010: 1-6

[57] Blahacek M, Prevorovsky Z, Krofta J, Chlada M. Neural network localization of noisy AE events in dispersive media. Journal of Acoustic Emission(USA). 2000;**18**:279

[58] Khalifa S, Komarizadeh MH. An intelligent approach based on adaptive neuro-fuzzy inference systems (ANFIS) for walnut sorting. Australian Journal of Crop Science. 2012;**6**(2)

[59] Vapnik V. Statistical Learning Theory New York. NY: Wiley; 1998

[60] Burges CJ. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 1998;**2**(2):121-167

[61] Saravanan N, Kumar Siddabattuni V, Ramachandran K. A comparative study on classification of features by SVM and PSVM extracted using Morlet wavelet for fault diagnosis of spur bevel gear box. Expert Systems with Applications. 2008;**35**(3):1351-1366

[62] Widodo A, Yang B-S, Kim EY, Tan AC, Mathew J. Fault diagnosis of low speed bearing based on acoustic emission signal and multi-class relevance vector machine. Nondestructive Testing and Evaluation. 2009;**24**(4):313-328

[63] Widodo A, Kim EY, Son J-D, Yang B-S, Tan AC, Gu D-S, Choi B-K, Mathew J. Fault diagnosis of low speed bearing based on relevance vector machine and support vector machine. Expert Systems with Applications. 2009;**36**(3):7252-7261

[64] Yu Y, Zhou L. Acoustic emission signal classification based on support vector machine. TELKOMNIKA Indonesian Journal of Electrical Engineering. 2012;**10**(5):1027-1032

[65] Chu-Shu K. A machine learning approach for locating acoustic emission. EURASIP Journal on Advances in Signal Processing. 2010;**2010**

[66] Yang Z, Yu Z. Grinding wheel wear monitoring based on wavelet analysis and support vector machine. The International Journal of Advanced Manufacturing Technology. 2012;**62**(1-4):107-121

[67] Yu Y, Zhou L. Acoustic emission signal classification based on support vector machine. Computer Engineering and Technology (ICCET), 2010 2nd International Conference, 16-18 April Chengdu. 2010;**6**:300-304

[68] Elforjani M, Shanbr S. Prognosis of bearing acoustic emission signals using supervised machine learning. IEEE Transactions on Industrial Electronics. 2017

[69] Ali YH, Ali SM, Rahman RA, Hamzah RIR. Acoustic Emission and Artificial Intelligent Methods in Condition Monitoring of Rotating Machine–A Review. National Conference for Postgraduate Research 2016, Universiti Malaysia Pahang. Malaysia. 2016

Life and Health Sciences

Normal Versus Abnormal ECG Classification by the Aid of Deep Learning

Linpeng Jin and Jun Dong

Additional information is available at the end of the chapter

Abstract

With the development of telemedicine systems, collected ECG records are accumulated on a large scale. Aiming to lessen domain experts' workload, we propose a new method based on lead convolutional neural network (LCNN) and rule inference for classification of normal and abnormal ECG records with short duration. First, two different LCNN models are obtained through different filtering methods and different training methods, and then the multipoint-prediction technology and the Bayesian fusion method are successively applied to them. As beneficial complements, four newly developed disease rules are also involved. Finally, we utilize the bias-average method to output the predictive value. On the Chinese Cardiovascular Disease Database with more than 150,000 ECG records, our proposed method yields an accuracy of 86.22% and 0.9322 AUC (Area under ROC curve), comparable to the state-of-the-art results for this subject.

Keywords: telemedicine, cardiovascular disease, ECG, deep learning, convolutional neural network, rule inference, classification

1. Introduction

As an inexpensive, noninvasive and well-established diagnostic tool for cardiovascular disease, the electrocardiogram (ECG) has been widely applied since the 1980s. In clinics, its main applications are as follows: (1) long-time ECG monitoring, such as 24-hour ambulatory ECG and (2) short-term ECG recording, such as standard 10-second, 12-lead ECG.

With the improvement of living standards, more and more people pay attention to their own health problems and performing an ECG test is the preferred method for preventing cardiovascular disease. Although it is easy to collect ECG now, diagnostic conclusions cannot be

accurately provided since domain experts are in short supply, especially in the basic community medical insurance system (BCMIS) of China. A feasible solution is that ECG records are sent to be interpreted by domain experts in telemedicine centers, who send back their diagnostic conclusions via the Internet. In fact, such institutions are common in China now, for instance, Shanghai Aerial Hospital Network, Henan Telemedicine Center, and so on. But in telemedicine centers, a large number of ECG records need to be interpreted and the workload of domain experts will be very heavy considering the huge possible audience. Since ECG records are mainly collected from people attending physical examinations, their diagnostic conclusions are likely to be "normal." If computer-assisted ECG analysis algorithms filter out most normal records and then domain experts only focus on interpreting the remaining abnormal ones, that is, man–machine integration [1], the diagnostic efficiency will be greatly increased and the social benefits will be significant. The key technical indicator is that the detection precision of normal records must be at least 90% for long-time ECG monitoring and 95% for short-term ECG recording [2].

As heartbeat classification is an important step in computer-based ECG analysis regardless of application scenarios, a considerable amount of research has been dedicated to this subject. In general, the processing flow of these works is such that feature vectors, including physiology characteristics with diagnostic value (such as RR interval and PR interval) [3] and statistical characteristics (such as wavelet transform [4] and independent component analysis [5]), are extracted from heartbeat segments first, and feature selection [6] is conducted when necessary. Afterwards, a number of machine learning algorithms are employed for classification, such as support vector machine (SVM) [7] and Neural Network [8]. The relevant literature can be categorized into two different types based on the adopted evaluation scheme, namely, "intra-patient" and "inter-patient" [9].

The intra-patient evaluation has been adopted by extensive literature and its main characteristic is that the training and testing sets contain heartbeat segments from the same patients. By this scheme that is not in conformity with clinical practice, classification results tend to be overly optimistic, since the human heartbeat can be used for identity recognition [10]. Using the MIT-BIH arrhythmias (MIT-BIH-AR) database [11] and the Advancement of Medical Instrumentation (AAMI) recommendation [12], de Chazal et al. [9] proposed the inter-patient evaluation where the training and testing sets are constructed by heartbeat segments from different ECG records so that inter-individual variation can be taken into account. This scheme adopted by recent works [5, 6, 8, 13, 14] can evaluate the clinical performance of heartbeat classification algorithms in a relatively effective manner.

Note that heartbeat classification is just an intermediate step and what domain experts are concerned about is the classification result of an ECG record provided by computer-assisted ECG analysis algorithms. However, due to limited standard ECG databases, there are less research works relevant to this subject compared with heartbeat classification [15–17]. For the MIT-BIH-AR database, there are a total of 48 two-lead ECG records with approximately 30 min in duration, each containing heartbeat annotations for both R-peak position and disease type. The lead configurations (sensor locations) are not the same at all and only 40 records have both II and VI. Based on this database, we cannot use the obtained classifier for telemedicine

centers. For the Common Standards for Quantitative ECG (CSE) database [18], there are approximately 1000 standard 10-second, 12-lead ECG records, each containing annotations for P, QRS and T positions. However, this database is not free, which makes many scholars unable to carry out research on it. For other databases (such as American Heart Association (AHA) database [19] and ST-T database [20]), there are also some advantages and disadvantages.

Aiming to carry out relevant research for telemedicine centers, we constructed the Chinese Cardiovascular Disease Database (CCDD) [21] containing 193,690 standard 12-lead ECG records with about 10–20 s in duration. As shown in **Figure 1**, the ECG record consists of 6 limb leads of I, II, III, aVR, aVL, aVF and six chest leads of V1, V2, V3, V4, V5, V6 that describe cardiac electrical activity jointly. If it is collected for 10 s at the sampling frequency of 500 Hz, there are 60,000 (=500 $\times$ 10 $\times$ 12) sampling points.

In this study, we develop a method for classification of normal and abnormal ECG records with short duration (record classification), but one that can also be easily extended to other cases since a long-term ECG record can be divided every T s into segments of length T [12]. Based on the CCDD, our research group has tried and proposed some methods for this subject. Using existing methods for heartbeat classification [22, 23] and the average-prediction technology, Zhu [24] obtained an accuracy of 72% with a specificity of 94% and a sensitivity of 25% when testing 11,760 ECG records. Wang [25] proposed a Multi-Way Tree Based Hybrid Classifier (MTHC), including analysis of RR interval, similarity analysis of QRS complex, analysis of physiology characteristics (such as P wave, T wave and PR interval) and statistical learning based on morphological and numerical characteristics. The specificity, sensitivity and accuracy are 54.62, 95.13 and 72.49%, respectively, on the 140,098 ECG records. Zhang [13] proposed a heartbeat classification method which has certain advantages over the relevant literature [5, 6, 8, 9, 14], but the accuracy is about 50–60% when he used it for record classification. For this, our prior work [26] analyzed traditional feature-based methods from the ability to construct nonlinear functions and proposed lead convolutional neural networks (LCNNs) for multi-lead ECGs. Using the explicit training method and the single-point-prediction technology, we
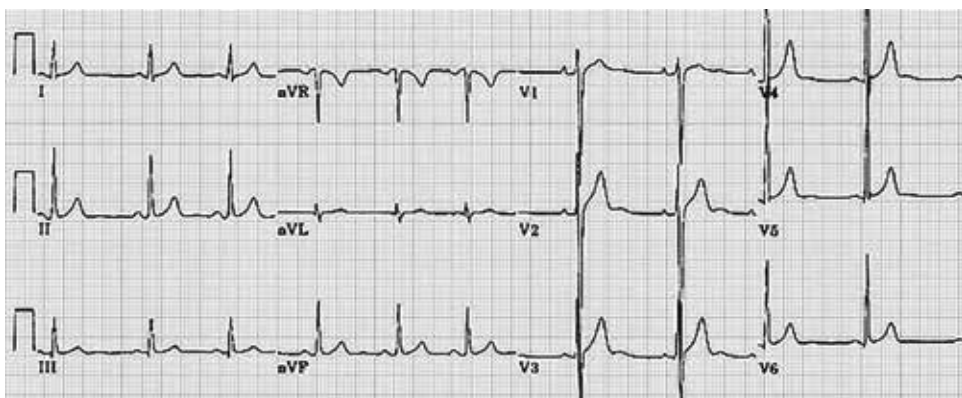
Figure 1. Standard 12-lead synchronous ECG.
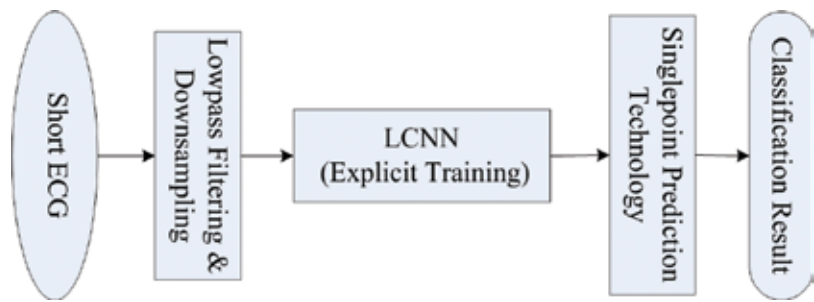
Figure 2. LCNN-based method for record classification.

achieved an accuracy of 83.66% with a specificity of 83.84% and a sensitivity of 83.43% when testing 151,274 records. **Figure 2** depicts the whole process flow.

To improve the classification performance further, this study extends our prior work in the following aspects. (1) Two different LCNN models are obtained through different filtering methods (a low-pass filter and a band-pass filter) and different training methods (the explicit method and the implicit method), and then the multipoint-prediction technology and the Bayesian fusion method are successively applied to them. (2) Four effective disease rules based on R peaks are developed for further analysis. (3) The final classification result is determined by utilizing the bias-average method.

The rest of the chapter is organized as follows: in Section 2, the CCDD used in this study is first introduced; in Section 3, the proposed method is described in detail; then in Section4, the experimental results, as well as comparison with the published results are presented; discussions and conclusions are provided in Section 5 and 6, respectively.

2. Dataset

There are 193,690 standard 12-lead ECG records with about 10–20s in duration and sampling frequency of 500 Hz in the CCDD. These data were successively obtained from hospitals (i.e., real clinical environment) located at different districts in Shanghai, Suzhou and Changsha. For data1–251, all the records have detailed heartbeat annotations including P-onset, P-peaks, P-offset, QRS-onset, R-peaks, QRS-offset, T-onset, T-peaks, T-offset and disease type. For data252–943, all the records only have heartbeat annotations for R-peak position and disease type. Just like the MIT-BIH-AR database, all the records in data1–943 were chosen purposefully and two or more cardiologists provided the annotations. As heartbeat annotations are available, we can carry out research on heartbeat classification. For data944–193,690, all the records only have diagnostic conclusions provided by one cardiologist. Note that a diagnostic conclusion may contain more than one disease type.

As shown in **Figure 3**, we use hexadecimal codes, that is, form of "0×dddddd", to encode disease types divided into three grades. There are 12 first-grade types (i.e., Invalid ECG,

```
0x00        Invalid ECG
0x01        Normal ECG
0x02        Sinus Rhythm                                      //The first-grade type
0x0201         Sinus Rhythm                                   //The second-grade type
0x020101          Sinus Rhythm                                //The third-grade type
0x020102          Normal Sinus Rhythm
0x020103          Heart Rate is Uncertain, May be Sinus Rhythm
0x020104          Heart Rate is Uncertain, May be Sinus Irregularity
                ...
0x0202         Sinus Irregularity
                ...
0x03        Atrial Arrhythmia
0x04        Junctional Rhythm
0x05        Ventricular Arrhythmia
0x06        Conduction Block
0x07        Atrial Hypertrophy
0x08        Ventricular Hypertrophy
0x09        Myocardial Infarction
0x0a        ST-T Change
0x0b        Other Abnormalities
```

Figure 3. Disease types in the CCDD.

Normal ECG, Sinus Rhythm, Atrial Arrhythmia, Junctional Rhythm, Ventricular Arrhythmia, Conduction Block, Atrial Hypertrophy, Ventricular Hypertrophy, Myocardial Infarction, ST-T Change and Other Abnormalities), 72 second-grade types and 335 third-grade types, including all the possible diagnostic conclusions provided by cardiologists in clinic. More details can be seen on our website (http://58.210.56.164:88/ccdd/).

In telemedicine centers, ECG records that are not explicitly diagnosed as "normal" all need to be further interpreted by domain experts, so that the potential cardiovascular disease of a patient can be detected as early as possible. Therefore, in this study we only regard ECG records whose diagnostic conclusions are "0×01" or "0×020101" or "0×020102" as "normal" (denoted as 0-class) and all the others as "abnormal" (denoted as 1-class). Moreover, we throw away some exception data, that is, the diagnostic conclusion is "0×00" or the duration is less than 9.625 s.

The training and testing sets are organized as follows [26]: "data944–25,693" is partitioned into three parts where the numbers of training samples, validation samples and testing samples (i.e., the small-scale testing set) are 12,320, 560 and 11,789, respectively, and the large-scale testing set is composed of 151,274 ECG records from data25694–179,130. Note that we will combine training samples and validation samples together for implicit training. **Table 1**

#	Dataset	Normal	Abnormal	Irregular	Total
Training samples	data 944–25,693	8800	3520	763	12,320
Validation samples	data 944–25,693	280	280	56	560
Small-scale testing set	data 944–25,693	8387	3402	822	11,789
Large-scale testing set	data 25694–179,130	85,141	66,133	9172	151,274

Table 1. Data distribution.

summarizes the detailed information, where "irregular" denotes the abnormal ECG records whose heart rhythms are irregular (their diagnostic conclusions are "0×0202").

3. Methodologies

Figure 4 shows the overall framework of the proposed method. As we can see, it consists of three parts, namely, statistical learning, rule inference as well as summarizing. In the part of statistical learning, the ECG record is first preprocessed by two different methods respectively, and then two probability values are outputted by utilizing LCNNs and the multipoint-prediction technology (LCNN[A] and LCNN[B] have the same architecture and are obtained by "explicit training" and "implicit training", respectively). After that, we use the Bayesian fusion method to incorporate the two outputs. In the part of rule inference, all R-peak positions in the ECG record are detected first [27], and then four disease rules based on them are used for further analysis. Finally, in the part of summarizing, we utilize the bias-average method to determine the classification result, that is, "normal" or "abnormal."
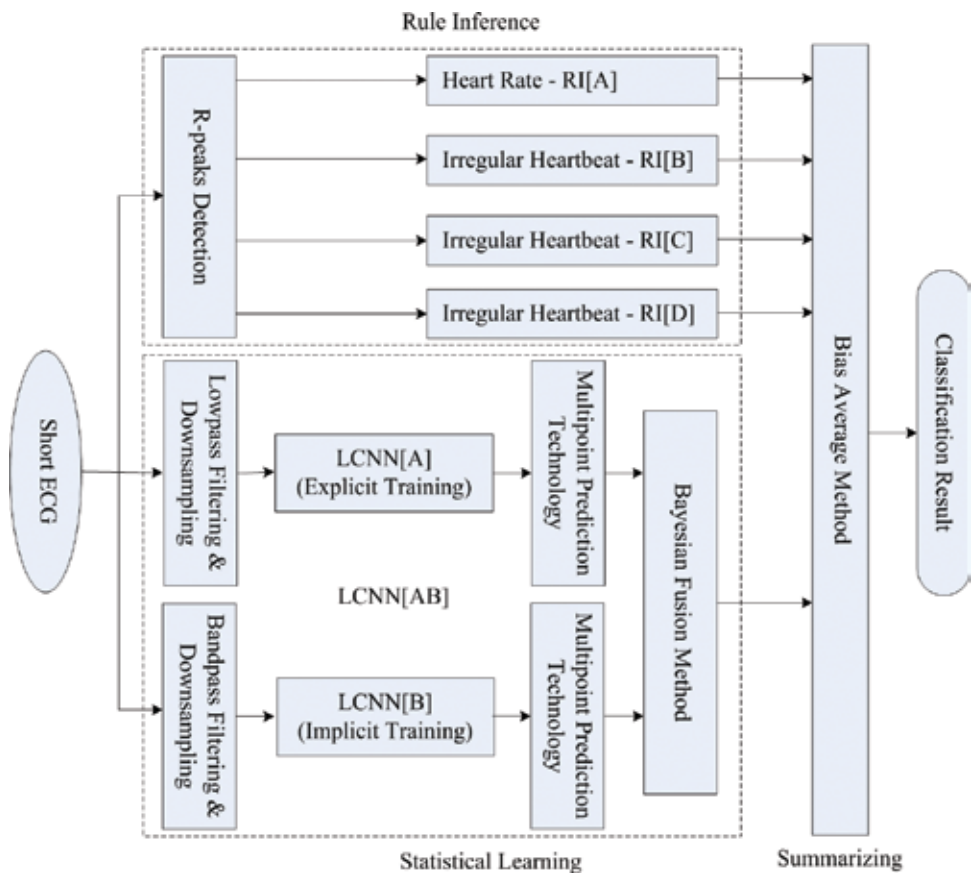
Figure 4. Overall framework of the proposed method.

Essentially, **Figure 4** describes an ensemble method including two homogeneous classifiers (i.e., LCNN[A] and LCNN[B]) and five heterogeneous classifiers (i.e., RI[A], RI[B], RI[C], RI [D], LCNN[AB] consisting of LCNN[A] and LCNN[B]). For the homogeneous ensemble, different base classifiers must be obtained if we want to be able to enhance classification performance. In fact, our generation strategy has certain advantages over some well-known ensemble methods such as Bagging and AdaBoost [28]. For the heterogeneous ensemble, different classifiers should be complementary to each other. Since LCNNs are not good at detecting abnormal heart rate and irregular heart rhythm, we use simple disease rules. Next, we will introduce each of the steps in detail.

3.1. Statistical learning

3.1.1. Preprocessing

Generally speaking, ECG records collected in clinics are often contaminated by several types of interfering noise, such as low-frequency waves (less than 0.5 Hz) caused by breathing movements, high-frequency waves (50 or 60 Hz) caused by electricity, and biological waves (about 33 Hz) caused by physical activities. Although we can use many methods to de-noise, some useful information may be lost after doing that. Since LCNNs which belong to deep learning [29, 30] have the ability to capture useful information but ignore interfering noise after learning from a certain number of training samples, we do not perform special de-noising.

An effective strategy for homogeneous ensemble learning is to make input data different, so we apply a low-pass filter and a band-pass filter of 0.5–40 Hz [31] to the ECG record, respectively. Of course, there is no problem if we exchange the low-pass filter for the band-pass filter; here we just make one path consistent with our prior work. After that, we conduct the downsampling operation (from original 500 to 200 Hz) and extract a data segment of 9.5 s from the incoming ECG record after ignoring the first 0.125 s. Only eight basic leads, namely II, III, V1, V2, V3, V4, V5 and V6 are reserved since the remaining four leads can be linearly derived from them. As a result, each ECG record consists of 8×1900 sampling points.

3.1.2. Lead convolutional neural network

Convolutional neural networks (CNNs) [32] have been successful in the field of ECG signal classification. Zhu used the CNN-based method for both heartbeat classification and record classification, and obtained an accuracy of 99.20% on the MIT-BIH-AR database with 47,190 heartbeats in the "intra-patient" evaluation, and an accuracy of 83.49% and 0.8819 AUC on the CCDD with 11,760 ECG records, respectively [24]. Kiranyaz applied CNNs on 1-lead ECGs for patient-specific heartbeat classification and achieved good results [33, 34]. However, multi-lead ECGs are different from two-dimensional images: data in the horizontal direction (intra-lead) are relevant while data in the vertical direction (inter-lead) are independent. For this, our prior work proposed lead convolutional neural networks (LCNNs) for multi-lead ECGs, which have better classification performance. **Figure 5** shows an example of three-stage LCNN.

In **Figure 5**, CU denotes a convolution unit consisting of a convolutional layer and a sub-sampling (max-pooling) layer; 1DCov and SubSamp denote the one-dimensional convolution
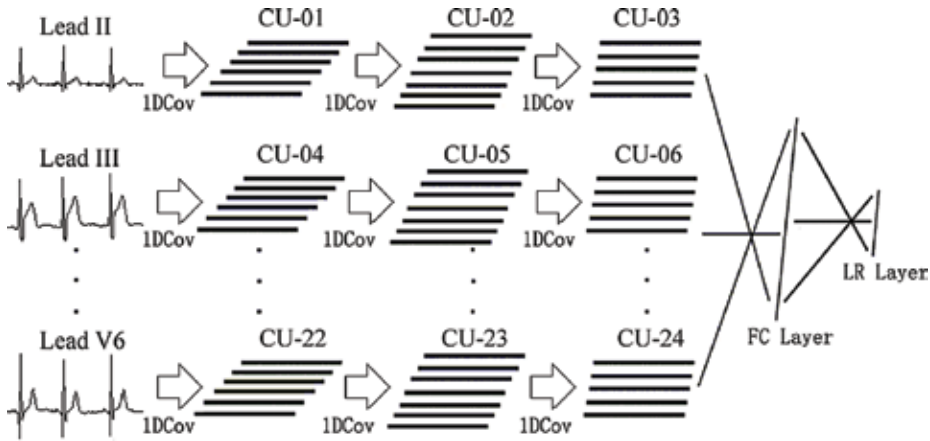
Figure 5. Architecture of a three-stage LCNN.

operation and the sub-sampling operation, respectively. The computational process consists of data for each lead going through three different CUs, and then inputting information from all leads into a fully connected (FC) layer. Finally, the logistic-regression (LR) layer output the predictive value. In fact, we can regard each CU as a feature extractor, and the subsequent multilayer perceptron consisting of the FC layer and the LR layer as a classifier.

To describe the LCNN clearly, we present its computational formula as follows: let $[x_1, x_2, ..., x_8]$ be the incoming ECG record where x_i (a vector, $1 <= i <= 8$) is the data of the i-th lead, then

$$
\begin{cases}
f(x) = g_E \left(g_D \left(\bigcup_{i=1}^{8} g_{C_i} \left(g_{B_i} \left(g_{A_i}(x_i) \right) \right) \right) \right) \\
g_D(x) = \varphi \left(W^T x + b \right) \\
g_E(x) = \dfrac{1}{1 + e^{-\left(W^T x + b \right)}}
\end{cases}
\tag{1}
$$

Here, g_E and g_D are the computational formulas for the LR layer and the FC layer, respectively, W and b are the weights and the biases of the corresponding layer, $\phi(x)$ is the sigmoid function. g_{Ai}, g_{Bi} and g_{Ci} ($1 <= i <= 8$) are the computational formulas for the CUs whose expression are all $f_{sub}(f_{cov}(x))$, the only differences between them exist in weights and biases. For the i-th lead, ECG data is inputted into g_{Ai} first, and then the output of g_{Ai} continues to be inputted into g_{Bi} and g_{Ci}. After that, the outputs of gC1, gC2, ..., gC8 (namely 8 vectors) are concatenated together (namely 1 vector) using the union operation U. Finally, the resulting vector is inputted into g_D and g_E successively and a value ranging from 0 to 1 is outputted. Note that both the input and the output of each function are vectors and the expressions of $f_{sub}(x)$ and $f_{cov}(x)$ are given by

$$
\begin{cases}
f_{\mathrm{cov}}(v) = \bigcup_{j,\,k} v_{ij}^k = \bigcup_{j,\,k} \left(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} w_{ij,\,(i-1)m}^p\, v_{(i-1)m}^{(k+p)} \right) \\[4em]
f_{\mathrm{sub}}(v) = \bigcup_{j,\,k} v_{ij}^k = \bigcup_{j,\,k} \varphi \left(\max \left(\bigcup_{\substack{q=(k-1)Q_i+1, \\ (k-1)Q_i+2, \\ \ldots, \\ kQ_i}} v_{(i-1)j}^q \right) \right)
\end{cases} \tag{2}
$$

Here, P_i and Q_i denote the size of the convolutional kernel and the sub-sampling step in the i-th layer, respectively, U denotes the union operation. v_{ij}^k denotes the output value of the k-th neuron (start from 1) in the j-th feature map (unit) of the i-th layer. $w_{ij,(i-1)m}^p$ is the weight that connects the j-th unit of the i-th layer with the m-th unit of the $(i-1)$-th layer, and b_{ij} is the bias in the j-th unit of the i-th layer.

The setting of each parameter (such as P_i, Q_i and the number of CUs for each lead) will greatly influence the classification performance, and too large or too small parameters all result in unfavorable outcome. In practical applications, they can be determined by a trial-and-error method. To keep things simple and ensure a fair comparison, we use two 3-stages LCNNs with the same architecture, namely LCNN[A] and LCNN[B] in **Figure 4**, and the corresponding parameters are the same as those in our prior work. The number of the neurons in the input layer is 8×1700; the sizes of three convolutional kernels are 21, 13 and 9, respectively; the sizes of three sub-sampling steps are 7, 6, and 6, respectively; the numbers of three feature maps are 6, 7 and 5, respectively; the number of the neurons in the FC layer is 50.

3.1.3. Training method

To obtain different LCNN models, we develop two training methods for LCNN based on mini-batch gradient descent [35] in supervised learning, that is, the explicit method and the implicit method. The main difference between them exists in validation mechanisms. Since gradient descent is a typical local search algorithm, we incorporate "translating starting point" and "adding noise" into it to increase the number of training samples presented to the network, so that classifiers with good generalization performance can be obtained.

The explicit method is commonly used for training neural networks, which utilizes independent validation samples to evaluate the obtained classifier during the training phase. As shown in **Figure 6(a)**, the training process can be described as follows: a random 8×1700 local segment is extracted from the 8×1900 training sample first (translating starting point), and then it is added with random signals whose maximal amplitude is less than 0.15 millivolts (adding noise) with high probability, finally back propagation is invoked. When Batchsize training samples have been presented to the network, the weights will be adjusted and the current LCNN model will be tested by the particular 8×1700 local segments (start from the first position) extracted from 8×1900 validation samples. If the accuracy is the best up to the

```
[Algorithm]:      Explicit Training Method
[Input]:          Training Samples X={(xt[i],yt[i]), i=1,2,...,N1}, Validation Samples V={(xv[i],yv[i]), i=1,2,...,N2}
[Output]:         An LCNN Model
 BestAcc = 0
  while(!MaxEpoch)
    dW = {0,0,...,0}                              //Initialize the weight changes
    for i = 1 to N1                              //Make sure that N1 is divisible by BatchSize
      xt1[i]  = PreprocessA(xt[i])               //Lowpass filtering & downsampling
      xt2[i]  = LocalSegment(xt1[i], rand)       //Extract a 8×1700 local segment starting from a random position
      xt3[i]  = AddRandSignal(xt2[i], rand)      //Add random signals with high probability
      dW      = dW + BackPropagation(xt3[i], yt[i]) //Invoke Backpropagation and accumulate dW
      if (i % BatchSize == 0)                    //BatchSize training samples have been presented to the network
        UpdateNN(dW)                             //Adjust the weights
        dW      = {0,0,...,0}
        CMatrix = [0,0;0,0]                       //Initialize the confusion matrix
        for j = 1 to N2
          xv1[j] = PreProcessA(xv[i])            //Preprocessing
          xv2[j] = LocalSegment(xv1[j], 1)       //Extract a 8×1700 local segment starting from the 1st position
          CMatrix = Test(xv2[j], yv[j])          //Test the current LCNN model
        end
        if(Accuracy(CMatrix) > BestAcc)
          BestAcc = Accuracy(CMatrix)
          SaveNN(Model)
        end
      end
    end
  end
end
```

a

```
[Algorithm]:      Implicit Training Method
[Input]:          Training Samples X={(xt[i],yt[i]), i=1,2,...,N}
[Output]:         An LCNN Model
 BestAcc = 0
  while(!MaxEpoch)
    dW      = {0,0,...,0}                         //Initialize the weight changes
    CMatrix = [0,0;0,0]                           //Initialize the confusion matrix
    for i = 1 to N                                //Make sure that N is divisible by BatchSize
      x1[i]  = PreprocessB(xt[i])                 //Bandpass filtering & downsampling
      x2[i]  = LocalSegment(x1[i], rand)          //Extract a 8×1700 local segment starting from a random position
      xt3[i] = AddRandSignal(x2[i], rand)         //Add random signals with high probability
      dW     = dW + BackPropagation(xt3[i],yt[i]) //Invoke Backpropagation and accumulate dW
      if (i % BatchSize == 0)                     //BatchSize training samples have been presented to the network
        UpdateNN(dW)                              //Adjust the weights
        dW     = {0,0,...,0}
        for j = 1 to BatchSize                    //Samples used between two adjacent weight-updating processes
          xt1[i]     = PreprocessB(xt[i])
          xt2[i-j+1] = LocalSegment(xt1[i-j+1],1) //Extract a 8×1700 local segment starting from the 1st position
          CMatrix    = Test((xt2[i-j+1],yt[i-j+1]))//Test the current LCNN model
        end
      end
    end
    if(Accuracy(CMatrix) > BestAcc)
      BestAcc = Accuracy(CMatrix)
      SaveNN(Model)
    end
  end
end
```

b

Figure 6. (a). Pseudo-code of explicit training method. (b). Pseudo-code of implicit training method.

present, the current LCNN model will be saved. Of course, the training will stop when reaching the maximum number of epochs.

Besides the explicit method, we develop an implicit method for training LCNN, whose main characteristic is that a small number of local segments extracted from training samples are

used for evaluating the obtained model. As shown in **Figure 6(b)**, most of the steps are the same as those shown in **Figure 6(a)**. When Batchsize training samples have been presented to the network, the weights will be adjusted and the current LCNN model will be tested by the particular 8×1700 local segments (start from the first position) extracted from the 8×1900 training samples, which are used between two adjacent weight-updating processes. At the end of each epoch, the current LCNN model will be saved if the total accuracy is the best up to the present.

While some may think this method would result in overfitting, this is a false assumption. We extract random local segments and add random signals to them with high probability for training. Differently, we extract particular local segments and do not add random signals to them for validation. Therefore, the probability of overlap is very small.

To be consistent with our prior work, we employ the back propagation algorithm of inertia moment and variable step [36] and do not use any unsupervised pre-training. The relevant parameters are set as follows: the initial step length is 0.02, the step decay is 0.01 except for the second and the third epoch (set as 0.0505), Batchsize is 560, and the maximum number of training epochs is 500.

3.1.4. Multipoint prediction technology

We can immediately get the classification result after inputting a 8×1700 local segment extracted from the incoming ECG record into the obtained LCNN model. This is the method our prior work adopted, namely the single-point-prediction technology. However, the output of the LCNN is a probability value ranging from 0 to 1 and the classification confidence is low if the value is around 0.5. For this, we develop a new testing method, namely the multipoint-prediction technology.

As shown in **Figure 7**, nine 8×1700 local segments which start from the 1st, 26th, 51st, 76th, 101st, 126th, 151st, 176th and 201st positions are extracted from the ECG record, respectively, then their predictive values outputted by the LCNN are aggregated by the average rule (for simplicity's sake), so that the classification confidence can be enhanced.

3.1.5. Bayesian fusion method

In this study, we employ a Bayesian fusion method [37] to incorporate the outputs of the two LCNNs: given M classifiers and K classes, the predicted class k can be determined based on the final probability estimates $P(y = i | c_1, c_2, ..., c_m)$ and is given by

$$\begin{cases} P(y = i | c_1, c_2, ..., c_m) = \dfrac{1}{M} \sum_{m=1}^{M} P(y = i | c_m) \\ k = \underset{1 \leq i \leq K}{\arg\max} \{ P(y = i | c_1, c_2, ..., c_m) \} \end{cases} \tag{3}$$

Here, $P(y = i | c_m)$ is the probability value predicted on the i-th class by the classifier c_m. What we focus on is a binary classification problem (i.e., "normal" vs. "abnormal"), thus $K = 2$, and

```
[Input]:      Testing Samples X  {x[i],1<  i<  N}, LCNN[A] and LCNN[B]
[Output]:     Classification Result
  for i = 1 to N
     //Multipoint Prediction Technology
     AvgPb1 = 0
     AvgPb2 = 0
     x1[i]    = PreprocessA(x[i])              //Lowpass Filtering & Downsampling
     x2[i]    = PreprocessB(x[i])              //Bandpass Filtering & Downsampling
     for j = 1 to 9
        start    = (j-1) * 25 - 1             //9 starting position
        x1s[i]   = LocalSegment(x1[i], start) //Extract a 8 ×1700 local segment starting from the given position
        x2s[i]   = LocalSegment(x2[i], start)
        AvgPb1   = AvgPb1 + TestA(x1s[i])     //Use LCNN[A] as a classifier
        AvgPb2   = AvgPb2 + TestB(x2s[i])     //Use LCNN[B] as a classifier
     end
     AvgPb1 = AvgPb1/9
     AvgPb2 = AvgPb2/9

     //Bayesian Fusion Method
     nne       (AvgPb1 - AvgPb2)/2

     //Bias Average Method
     if(max(ro1,ro2,ro3,ro4) == 1)            //ro1,ro2,ro3,ro4 are the outputs of four rule based classifiers
        PbEm  = (max(ro1,ro2,ro3,ro4) + nne)/2
     else
        PbEm  = nne
     End

     //Determine the classification result
     if(PbEm < 0.5)
        //x[i] is identified as "normal"
     else
        //x[i] is identified as "abnormal"
     end
  end
```

Figure 7. Pseudo-code of testing process.

obviously, *M* = 2. Of course, other fusion methods can be employed for this purpose. For example, we can train one logistic regression using the concatenation of the outputs of both LCNNs, and it may be able to give better results. The part of statistical learning is denoted as the LCNN[AB] classifier.

3.2. Rule inference

R-peaks detection algorithms have reached a very high level at present [38]. On the other hand, there is no doubt that an ECG record is abnormal if it is identified as a specific disease. Hence, we develop four disease rules based on R peaks to detect abnormal heart rate and irregular heart rhythm.

Let *fs* be the sampling frequency, R_i (1 < = i < = n) be the *i*-th R-peak position in an ECG record, and *std.(.)* be the function of calculating the standard deviation. The formula for calculating the average RR interval is given by

$$AvgRR = \frac{1}{n-1}\sum_{i=2}^{n}(R_i - R_{i-1}) \tag{4}$$

The four disease rules are defined as follows:

1. Heart rate

The formula for calculating heart rate is given by

$$HR = \frac{60 \times fs \times (n-1)}{R_n - R_1} \tag{5}$$

Normal heart rate is defined as 60–100 CPM (i.e., counts per minute) clinically; here we regard 59–101 CPM as normal, otherwise as abnormal. This rule is denoted as the RI[A] classifier.

2. Irregular heart rhythm based on local characteristics.

The first rule for detecting irregular heart rhythm is defined as follows: three successive RR intervals exceed 15% of the average RR interval, that is, $\exists k \in [1, n-3]$, $\forall j \in [1, 3]$ satisfies

$$\left| \frac{(R_{k+j} - R_{k+j-1}) - AvgRR}{AvgRR} \right| > 0.15 \tag{6}$$

We regard an ECG record as abnormal if Eq. (6) is satisfied, otherwise as normal. This rule is denoted as the RI[B] classifier.

3. Irregular heart rhythm based on local and global characteristics.

The second rule for detecting irregular heart rhythm is defined as follows: one RR interval exceeds 15% of the average RR interval and the standard deviation of the rates between two neighboring RR intervals is greater than 0.05, that is, $\exists k \in [1, n-1]$ satisfies

$$\begin{cases} \left| \dfrac{(R_{k+1} - R_k) - AvgRR}{AvgRR} \right| > 0.15 \\ std\left(\left\{ \dfrac{R_{i+2} - R_{i+1}}{R_{i+1} - R_i} \middle| 1 \le i \le n-2 \right\} \right) > 0.05 \end{cases} \tag{7}$$

Likewise, we regard an ECG record as abnormal if Eq. (7) is satisfied, otherwise as normal. This rule is denoted as the RI[C] classifier.

4. Irregular heart rhythm based on global characteristics.

The last rule for detecting irregular heart rhythm is defined as follows: the standard deviation of RR intervals is greater than (0.05 × AvgRR), that is, satisfies

$$std(\{R_{i+1} - R_i | 1 \le i \le n-1\}) > 0.05 \times AvgRR \tag{8}$$

In the same way, we regard an ECG record as abnormal if Eq. (8) is satisfied, otherwise as normal. This rule is denoted as the RI[D] classifier.

Of course, false R-peak detections will influence the classification performance of the four disease rules and result in an unfavorable outcome. In this study, we utilize the Zhu method

[27] to detect R peaks since its accuracy is 99.83% in the MIT-BIH-AR database with 48 ECG records and 99.78% in the CCDD with 251 ECG records, respectively. Note that 0 is outputted if the classification result is "normal", otherwise 1 is outputted.

3.3. Summarizing

As we can see, there are five classifiers, namely RI[A], RI[B], RI[C], RI[D] and LCNN[AB], involved in **Figure 4**. We can regard an ECG record as "abnormal" if any one of the classifiers gives such a result. However, the outputs of the four rule-based classifiers are 0 or 1, while the output of the LCNN[AB] classifier is a probability value ranging from 0 to 1. We hope that the predictive value can be outputted in the form of a probability, so a new fusion method, namely the bias-average method, is developed, given by

$$\begin{cases} ro = \max(ro_1, ro_2, ro_3, ro_4) \\ output = \begin{cases} \dfrac{ro + nno}{2}, & ro == 1 \\ nno, & otherwise \end{cases} \end{cases} \tag{9}$$

Where, ro_1, ro_2, ro_3, ro_4 and nno are the outputs of the RI[A], RI[B], RI[C], RI[D] and LCNN [AB] classifiers, max(.) is to choose the maximum value. If we use the first three disease rules, just replace "$\max(ro_1, ro_2, ro_3, ro_4)$" with "$\max(ro_1, ro_2, ro_3)$." The final output value ranges from 0 to 1 and the classification result is "normal" if it is less than 0.5, otherwise it is "abnormal." **Figure 7** shows the whole testing process including the multipoint-prediction technology, the Bayesian fusion method and the bias-average method.

4. Result

4.1. Performance metrics

In this study, we use several metrics to investigate the performance of different algorithms: Sp (specificity), NPV (negative predictive value), Se (sensitivity) and Acc (accuracy), given by

$$\begin{cases} Sp = \dfrac{TN}{TN + FP} \\ NPV = \dfrac{TN}{TN + FN} \\ Se = \dfrac{TP}{TP + FN} \\ Acc = \dfrac{TP + TN}{TN + FP + FN + TP} \end{cases} \tag{10}$$

Here, TP is the number of abnormal ECG records that are correctly classified as abnormal, TN is the number of normal ECG records that are correctly classified as normal, FP is the number of normal ECG records that are incorrectly classified as abnormal, and FN is the number of

abnormal ECG records that are incorrectly classified as normal. We also use receiver operating characteristic (ROC) curve-related metrics, including area under ROC curve (AUC), true positive value (TPR) (e.g., "Sp" in this study) and false positive value (FPR)(e.g., "1-Se" in this study). A rough suggestion for measuring classification performance using "AUC" is as follows [39]:

1. 0.90–1.00 = excellent classification;

2. 0.80–0.90 = good classification;

3. 0.70–0.80 = fair classification;

4. 0.60–0.70 = poor classification;

5. 0.50–0.60 = failure.

Note that "NPV" in this study is also called the detection precision of normal ECG records.

4.2. Numerical experiment

Our prior work [26] has achieved the best results for record classification on the CCDD up to now, so it is meaningful to compare the proposed method with it. As described in Section 2, the same training, validation and testing samples are used. To show the contribution of each strategy intuitively, we present the corresponding results in **Tables 2** and **3** in turn. The classification results on the small-scale testing set is only used for reference; what we mostly focus on is the results on the large-scale testing set, which can be deemed as a more realistic estimate of potential performance in real applications.

In **Tables 2** and **3**, "0", "1", "2", "3" and "4" denote the LCNN[AB], RI[A], RI[B], RI[C] and RI[D] classifiers, respectively. From the results, we can see that although NPV and Se are slightly lower, "0" outperformed our prior work in all other metrics regardless of the small-scale testing set and the large-scale testing set. On this basis, most metrics continue to increase when "1", "2" and "3" are added, while many metrics start to decrease when "4" is added. The role of "4" is to increase NPV and Se, so that TPR under the condition of NPV being equal to

Model	Sp	NPV	Se	Acc	AUC	FPR = 1%		NPV = 95%		NPV = 90%	
						TPR	NPV	FPR	TPR	FPR	TPR
LCNN [26]	88.84	90.48	76.95	85.41	0.9034	17.5	97.8	8.22	63.3	24.7	90.1
0	91.81	90.04	74.96	86.95	0.9172	23.0	98.3	9.61	74.1	25.2	91.9
0 + 1	91.77	90.11	75.16	86.98	0.9176	23.3	98.3	9.63	74.2	25.2	91.9
0 + 1 + 2	91.62	90.22	75.51	86.97	0.9180	23.3	98.3	9.63	74.2	25.2	91.9
0 + 1 + 2 + 3	90.28	91.11	78.28	86.82	0.9220	24.6	98.4	9.90	76.3	25.3	92.3
0 + 1 + 2 + 3 + 4	80.60	93.09	85.24	81.94	0.9196	25.8	98.5	9.71	74.8	25.1	91.7

Table 2. The results of different classification models on the small-scale testing set.

Model	Sp	NPV	Se	Acc	AUC	FPR = 1%		NPV = 95%		NPV = 90%	
						TPR	NPV	FPR	TPR	FPR	TPR
LCNN [26]	83.84	86.69	83.43	83.66	0.9086	16.0	95.4	1.81	26.7	10.6	73.9
0	87.49	86.30	82.12	85.15	0.9199	22.1	96.6	3.42	50.4	11.3	79.3
0 + 1	87.43	86.75	82.81	85.41	0.9225	24.4	96.9	3.60	53.1	11.5	80.4
0 + 1 + 2	87.20	87.30	83.67	85.66	0.9251	26.2	97.1	3.83	56.5	11.7	81.6
0 + 1 + 2 + 3	86.03	89.11	86.46	86.22	0.9322	34.3	97.8	4.43	65.4	12.1	84.4
0 + 1 + 2 + 3 + 4	80.37	91.04	89.82	84.50	0.9320	36.7	97.9	4.60	67.9	11.8	82.4

Table 3. The results of different classification models on the large-scale testing set.

95% (TPR95) can be increased. Compared with our prior work, the "0 + 1 + 2 + 3" model increased all the metrics and significantly improved the classification performance.

As mentioned previously, the aim of this study is to develop a computer-assisted ECG analysis algorithm for telemedicine centers. Specially, the normal ECG records are filtered out first, and then the remaining abnormal ones are delivered to domain experts for further interpretation, so that their workload can be lessened and diagnostic efficiency can be improved [1]. The key technical indicator is to make TPR95 as high as possible [2]. From **Tables 1** and **3** we can see that, 38.21% (=56.28% × 67.9%) of domain experts' workload can be lessened. The more normal ECG records there are, the less work domain experts do. Although the proposed method increases the computational complexity, we can use the classification system in practical applications. The total computing time for an ECG record is about 125 ms on an Intel Core2 CPU@2.93GHz, 2GB RAM, 32bit Window 7 OS.

Some may think that the self-compared results are not supportive enough to show the effectiveness. In fact, Zhu [24], Wang [25] and Zhang [13] have proposed methods for this subject, but their results are significantly inferior to those of our prior work. Since the heartbeat classification method [5] has achieved the highest accuracy of 99.3% in the "intra-patient" evaluation by now and a good accuracy of 86.4% in the "inter-patient" evaluation, we re-implement it using the CCDD and classify each ECG record with any abnormal heartbeats as "abnormal", but the performance is poor. The open-source software ECG-KIT [11] has recently been made available on Physionet (http://physionet.org/physiotools/ecg-kit/). However, different from the classification standard in this study, ECG-KIT uses the AAMI recommendation to output the result, so it would not necessarily be meaningful to compare our proposed method with ECG-KIT.

Finally, let us consider the industrial applications. The statistical analysis on the classification results given by General Electric Medical System shows that the total accuracy on 2112 ECG records is 88.0%, where the accuracies of interpreting sinus rhythms and non-sinus rhythms are 95 and 53.5%, respectively [40]. Hence, the total accuracy will only be 74.25% if the ratio of the number of sinus rhythms to the number of non-sinus rhythms is 1:1. The similar statistical analysis also shows that on 576 ECG records, the accuracies of Philips Medical Systems and

Draeger Medical Systems are 80 and 75%, respectively, and the average accuracy of non-experts is 85% [41]. Our proposed method achieved an accuracy of 86.22% and 0.9322 AUC on 151,274 ECG records, indicating that it has competitive classification performance.

5. Discussion

Due to inter-individual variation in ECG characteristics and the complexity of clinical data, record classification is a highly difficult problem. As a new research direction in recent years, deep learning has already achieved great success in hard Artificial Intelligence tasks, such as speech recognition [42] and image classification [43]. As an improved deep learning architecture, the LCNN show good performance for record classification. However, one LCNN with limited layers and neurons is not strong enough, so we use an ensemble method based on LCNN in this study and the experimental results show its effectiveness. In fact, using the same training and testing samples (heartbeat segments) from the MIT-BIH-AR database, the LCNN-based ensemble method has achieved an accuracy of 99.46% with a specificity of 99.69% and a sensitivity of 98.73% for detecting ectopic heartbeat, comparable to the state-of-the art results for heartbeat classification in the "intra-patient" evaluation [5].

One may doubt the viability of LCNNs since no feature-extraction operations are involved. In fact, if we let f(x) and g(x) be the corresponding function of feature extraction and classification, respectively, the decision-making function of traditional machine-learning methods can be written as g(f(x)). Therefore, we can directly construct the function g(f(x)) by deep-learning techniques [26, 44]. Nevertheless, there is not a general-purpose method for all problems, and what one needs to do is to choose or develop the most appropriate method for a special problem. For instance, abnormal ECG records whose heart rhythms are irregular should be collected first if we use the LCNN to detect irregular heart rhythm. It works well; however, the LCNN will capture not only characteristics of heart rhythm, but also other characteristics such as morphology and intervals during the training phase. That will result in more types of ECG records to be collected if we want to obtain a good classifier. In fact, on the small-scale and large-scale testing sets, the sensitivities of the LCNN[AB] classifier in the detection of irregular heart rhythm are 58.15 and 58.54%, respectively, while the corresponding results for detecting abnormal ECG records whose heart rhythm are regular are 80.31 and 85.92%. From **Table 1**, we know that the two testing sets only have a small proportion of irregular heart rhythm. If there are more abnormal ECG records whose heart rhythm is irregular, the LCNN[AB] classifier will not achieve such good results. In a word, accurate detection of irregular heart rhythm plays an important role in enhancing classification performance. Fortunately, we can achieve this aim by running the calculation based on R-peak positions. It is also effective for detecting abnormal heart rate, thus four disease rules are developed in this study.

From the perspective of simulating cardiologists' diagnostic thinking, LCNNs and disease rules express experiential knowledge and intuitive knowledge, respectively, and the combination of them is a complete simulation [1]. This is another reason why the proposed method can achieve good classification results.

In fact, our proposed method can be divided into two parts: the classification of "normal" versus "abnormal" serves as a global classifier and the classification of each specific disease serve as a local classifier. The advantage of the global classifier is that the error accumulation due to introducing each local classifier can be avoided, since the detection performance for many cardiovascular diseases is not very good regardless of whether LCNNs or disease rules are used. There are not enough samples used for training LCNNs in many cases (especially the disease with low detection rate in the general population); while for disease rules, the inaccurate detection of fiducial points can result in subsequent misclassification. Nevertheless, to enhance classification performance further, we can integrate classifiers of easily detectable diseases into **Figure 4**. For instance, using LCNNs for atrial fibrillation detection, the sensitivity, specificity and accuracy are 98.93, 98.76 and 98.76%, respectively, on 142,167 ECG records in the CCDD, and the detection performance is still excellent on ECG records collected by Shanghai MicroPort Co. Ltd. (almost 100% accuracy). Likewise, using disease rules, we can effectively identify QRS complexes with abnormal amplitude.

6. Conclusion

In order to lessen the workload of domain experts in telemedicine centers, we present a systematic approach for record classification in this chapter. Based on LCNNs and disease rules, an effective ensemble method including two homogeneous classifiers and five heterogeneous classifiers is developed. On the CCDD, our method yields an accuracy of 86.22% and 0.9322 AUC (excellent), which is a significant improvement on previously reported results [13, 24–26]. Specifically, TPR under the condition of NPV being equal to 95% can reach 67.9%, which means that the workload of domain experts can be decreased by (N% × 67.9%) in a clinically acceptable scope (especially in the BCMIS of China) if the percentage of normal ECG records is N%. In general, at least 70 × 67.9 = 47.53% of domain experts' workload can be reduced since N% can be more than 70%. We have deployed the classification system on the real-time cloud platform of Shanghai Aerial Hospital Network.

Regarding future work, our aim is to develop effective detection algorithms for other common cardiovascular diseases such as premature atrial contraction and premature ventricular contraction, and to integrate them into **Figure 4**. Since both LCNNs and disease rules have advantages and disadvantages, the combination of the two is a preferable research direction.

Author details

Linpeng Jin[1,2] and Jun Dong[1]*

*Address all correspondence to: jdong2010@sinano.ac.cn

1 Suzhou Institute of Nano-tech and Nano-bionics, Chinese Academy of Sciences, Suzhou, China

2 University of Chinese Academy of Sciences, Beijing, China

References

[1] Dong J, Zhang JW, Zhu HH, et al. Wearable ECG monitors and its remote diagnosis service platform. IEEE Intelligent Systems. 2012;**6**(27):36-43. DOI: 10.1109/MIS.2012.4

[2] Liu X. Atlas of Classical Electrocardiograms. 1st ed. Shanghai: Shanghai Science and Technology Press; 2011

[3] de Chazal P, Reilly RB. A patient-adapting heartbeat classifier using ECG morphology and heartbeat interval features. IEEE Transactions on Biomedical Engineering. 2006;**53**(12): 2535-2543. DOI: 10.1109/TBME.2006.883802

[4] Sumathi S, Beaulah HL, Vanithamani R. A wavelet transform based feature extraction and classification of cardiac disorder. Journal of Medical Systems. 2014;**38**(9):1-11. DOI: 10.1007/s10916-014-0098-x

[5] Ye C, Kumar BV, Coimbra MT. Heartbeat classification using morphological and dynamic features of ECG signals. IEEE Transactions on Biomedical Engineering. 2012;**59**(10):2930-2941. DOI: 10.1109/TBME.2012.2213253

[6] Llamedo M, Martinez JP. Heartbeat classification using feature selection driven by database generalization criteria. IEEE Transactions on Biomedical Engineering. 2011;**58**(3):616-625. DOI: 10.1109/TBME.2010.2068048

[7] Osowski S, Hoa LT, Markiewic T. Support vector machine-based expert system for reliable heartbeat recognition. IEEE Transactions on Biomedical Engineering. 2004;**51**(4):582-589. DOI: 10.1109/TBME.2004.824138

[8] Mar T, Zaunseder S, Martinez JP, et al. Optimization of ECG classification by means of feature selection. IEEE Transactions on Biomedical Engineering. 2011;**58**(8):2168-2177. DOI: 10.1109/TBME.2011.2113395

[9] de Chazal P, O'Dwyer M, Reilly RB. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Transactions on Biomedical Engineering. 2004;**51**(7):1196-1206. DOI: 10.1109/TBME.2004.827359

[10] Biel L, Pettersson O, Philipson L, et al. ECG analysis: A new approach in human identification. IEEE Transactions on Instrumentaion and Measurement. 2001;**50**(3):808-812. DOI: 10.1109/19.930458

[11] Goldberger AL, Amaral L, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation. 2000;**101**(23):215-220. DOI: 10.1161/01.CIR.101.23.e215

[12] Testing and Reporting Performance Results of Cardiac Rhythm and ST Segment Measurement Algorithms. ANSI/AAMI EC57:1998, Rev; 2012

[13] Zhang ZC, Dong J, Luo XQ, et al. Heartbeat classification using disease-specific feature selection. Computers in Biology and Medicine. 2014;**46**:79-89. DOI: 10.1016/j. compbiomed. 2013.11.019

[14] Escalona-Moran MA, Soriano MC, Fischer I, et al. Electrocardiogram classification using reservoir computing with logistic regression. IEEE Journal of Biomedical and Health Informatics. 2015;**19**(3):892-898. DOI: 10.1109/JBHI.2014.2332001

[15] Bortolan G, Degani R, Pedrycz W. A fuzzy pattern matching technique for diagnostic ECG classification. In: Proceedings of Computers in Cardiology; 25–28 September 1988; Washington, DC. New York: IEEE; 1988. pp. 551-554

[16] Willems JL, Abreu-Lima C, Arnaud P, et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. The New England Journal of Medicine. 1991;**325**(25):1767-1773. DOI: 10.1056/NEJM199112193252503

[17] Rahman QA, Tereshchenko LG, Kongkatong M, et al. Utilizing ECG-based heartbeat classification for hypertrophic cardiomyopathy identification. IEEE Transactions on Nanobioscience. 2015;**14**(5):505-512. DOI: 10.1109/TNB.2015.2426213

[18] Willems JL, Arnaud P, Bemmel JHV, et al. A reference database for multilead electrocardiographic computer measurement programs. Journal of the American College of Cardiology. 1987;**10**(6):1313-1321. DOI: 10.1016/S0735-1097(87)80136-5

[19] Ameriean Heart Association Database [Internet]. 1985. Available from: http://www.physionet.org/physiobank/other.shtml [Accessed: January 06, 2018]

[20] Taddei A, Distante G, Emdin E, et al. The European ST-T database: Standard for evaluating Systems for the Analysis of ST-T changes in ambulatory electrocardiography. European Heart Journal. 1992;**13**(9):1164-1172. DOI: 10.1093/ oxfordjournals.eurheartj.a060332

[21] Zhang JW, Liu X, Dong J. CCDD: An enhanced standard ECG database with its Management & Annotation Tools. International Journal on Artificial Intelligence Tools. 2012;**21**(5): 1-26. DOI: 10.1142/S0218213012400209

[22] Minhas FU, Arif M. Robust electrocardiogram beat classification using discrete wavelet transform. Physiological Measurement. 2008;**29**(5):555-570. DOI: 10.1109/ ICBBE.2008.796

[23] Martis RJ, Chakraborty C, Ray AK. A two-stage mechanism for registration and classification of ECG using Gaussian mixture model. Pattern Recognition. 2009;**42**(11):2979-2988. DOI: 10.1016/j.patcog.2009.02.008

[24] Zhu HH. Research on ECG recognition critical methods and development on remote multi body characteristic signal monitoring system [Thesis]. Beijing: University of Chinese Academy of Sciences; 2013

[25] Wang LP. Study on Approach of ECG Classification with Domain Knowledge [Thesis]. Shanghai: East China Normal University; 2013

[26] Jin LP, Dong J. Deep learning research on clinical electrocardiogram analysis. Science China Information Sciences. 2015;**45**(3):398-415. DOI: 10.1360/N112014-00060

[27] Zhu HH, Dong J. An R-peak detection method based on peaks of Shannon energy envelope. Biomedical Signal Processing and Control. 2013;**8**(5):466-474. DOI: 10.1016/j.bspc.2013.01.001

[28] Jin LP, Dong J. Ensemble deep learning for biomedical time series classification. Computational Intelligence and Neuroscience, 2016. Article ID: 6212684. DOI: 10.1155/2016/6212684

[29] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;**313**(5786):504-507. DOI: 10.1126/science.1127647

[30] Bengio Y. Learning deep architectures for AI. Foundations and Trends in Machine Learning. 2009;**2**(1):1-127. DOI: 10.1561/2200000006

[31] Thakor NV, Webster JG, Tompkins WJ. Estimation of QRS complex power spectra for design of a QRS filter. IEEE Transactions on Biomedical Engineering. 1984;**31**(11):702-706. DOI: 10.1109/TBME.1984.325393

[32] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998;**86**(11):2278-2324. DOI: 10.1109/5.726791

[33] Kiranyaz S, Ince T, Hamila R, et al. Convolutional neural networks for patient-specific ECG classification. In: The 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 25-29 August 2015, Milan, Italy. New York: IEEE; 2015. pp. 2608-2611

[34] Kiranyaz S, Ince T, Hamila R, et al. Real-time patient-specific ECG classification by 1D convolutional neural networks. IEEE Transactions on Biomedical Engineering. 2016;**63**(3):664-675. DOI: 10.1109/TBME.2015.2468589

[35] Li M, Zhang T, Chen YQ, et al. Efficient mini-batch training for stochastic optimization. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 24–27 August, 2014; 2014. New York: ACM; 2014. pp. 661-670

[36] Vogl TP, Mangis JK, Rigler AK, et al. Accelerating the convergence of the back-propagation method. Biological Cybernetics. 1988;**59**(4):257-263. DOI: 10.1007/ BF00332914

[37] Kittler J, Hatef M, Duin RPW, et al. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998;**20**(3):226-239. DOI: 10.1109/34.667881

[38] Kohler BU, Hennig C, Orglmeister R. The principles of software QRS detection. IEEE Engineering in Medicine and Biology Magazine. 2002;**21**(1):42-57. DOI: 10.1109/51. 993193

[39] Gorunescu F. Data Mining Concepts Techniques and Models. 1st ed. Berlin: Springer; 2011. pp. 185-317. DOI: 10.1007/978-3-642-19721-5

[40] Shah AP, Rubin SA. Errors in the computerized electrocardiogram interpretation of Cardiacrhythm. Journal of Electrocardiology. 2007;**40**(5):385-390. DOI: 10.1016/j. jelectrocard. 2007.03.008

[41] Hakacova N, Tragardh-Johansson E, Wagner GS, et al. Computer-based rhythm diagnosis and its possible influence on nonexpert electrocardiogram readers. Journal of Electrocardiology. 2012;**45**(1):18-22. DOI: 10.1016/j.jelectrocard.2011.05.007

[42] Dahl GE, Dong Y, Li D, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech and Language Processing. 2012;**20**(1):30-42. DOI: 10.1109/TASL.2011.2134090

[43] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems; 3–8 December. Vol. 2012. Vancouver: NIPS; 2012. pp. 153-160

[44] Sutskever I, Hinton GE. Deep narrow sigmoid belief networks are universal Approximators. Neural Computation. 2008;**20**:2629-2636. DOI: 10.1162/neco.2008. 12-07-661

A Quantitative Approach for Web Usability Using Eye Tracking Data

López-Orozco and Florencia-Juárez

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.74562

Abstract

This chapter presents a relatively new approach to show how a web usability classical paradigm can benefit from quantitative data of a nonclassical approach. In the pilot stage, we used experimental eye tracking data acquired from 11 participants faced to a web page to perform three simple tasks. Results show advantages by using eye tracking data to identify and verify some usability problems of such a web page. In this chapter, some hints are presented for people interested in measuring web usability by using such an approach. However, a deeper study should be carried out in order to generalize our results toward the construction of a methodology to be followed by a web developer or interested people in such a field of research.

Keywords: eye tracking, web usability, human-computer interaction, quantitative approach, nonclassical approach

1. Introduction

Nowadays, usability is one of the key factors in the success or failure of a web project. Usability is considered a quality attribute that evaluates the accessibility, readability, navigability and ease of learning of a website or application by the user [1]. Moreover, people do not wish to learn how to use a website, and there is no manual to use a web page when a person does not find the information or products they are looking for within a website. Users should have the ability to understand how a website works immediately after viewing the website. If someone gets lost while browsing a website or has difficulty in reading the information, he or she simply tries not to use that site anymore [2].

Typically, usability has been measured in a qualitative way, using questionnaires, focus groups, interviews, and so on, and, to a lesser extent, in a quantitative manner by measuring time to complete a task, number of errors, number of clicks, and so on [3]. The qualitative methods offer the advantage of having a direct opinion of the users, such as what they like and dislike about the page and how it could be improved, but the opinions of the users could be different from each other. In contrast, the quantitative techniques help us to assign numerical values to statements or observations, in order to study the possible relationships between variables by using statistical methods.

It is important to emphasize that we do not consider any of the techniques mentioned earlier as one "best technique." None of them solve the problem of measuring usability completely. Both approaches have their advantages and disadvantages to consider a limited number of influencing factors [4]. However, techniques that combine the strengths of both are still better and more promising as shown here.

In this chapter, the eye tracking technique was used to obtain quantitative metrics based on what users see on a web page, with the aim of adding these metrics to the traditional qualitative observations, used in the measurement of usability of web pages.

2. Eye tracking technology

The eye tracker is a tool that detects the users' gaze and provides an accurate representation and understanding of an individual's eye movement behavior. An eye tracker detects the two basic eye movements: saccades and fixations. A saccade is the eye movement itself which is needed to place the sight in a different place and a fixation is the period of time when the eye is practically maintained on a single location.

Using an eye tracker allows:

- To see what the user observes on the screen by detecting the user's fixations, both the left and the right eyes on the X and Y coordinates according to the size of the screen as well as duration of each fixation.

- To know the route or scanpath that the user looks at. The eye tracker also provides the saccade route, which shows the fixations made and the movement between them as shown in **Figure 1**.

- To know the diameter of the pupil when making the fixations or at the moment, the user clicks the mouse on a certain position of the web page.

In this chapter, a binocular infrared S2 Eye Tracker Mirametrix Inc. was used. The eye tracker works with an accuracy of 0.5°, data rate of 60 Hz, a head movement box of 25 cm × 11 cm × 30 cm (width × height × depth); with a system setup time less than 5 min, calibration time less than 30 s; and a suggested screen size of 15"–22" monitor, with capacity of blink tracking recovery, fast time to track recovery, weight of 1 lb, eye tracking technique of bright pupil; and low-power computers.
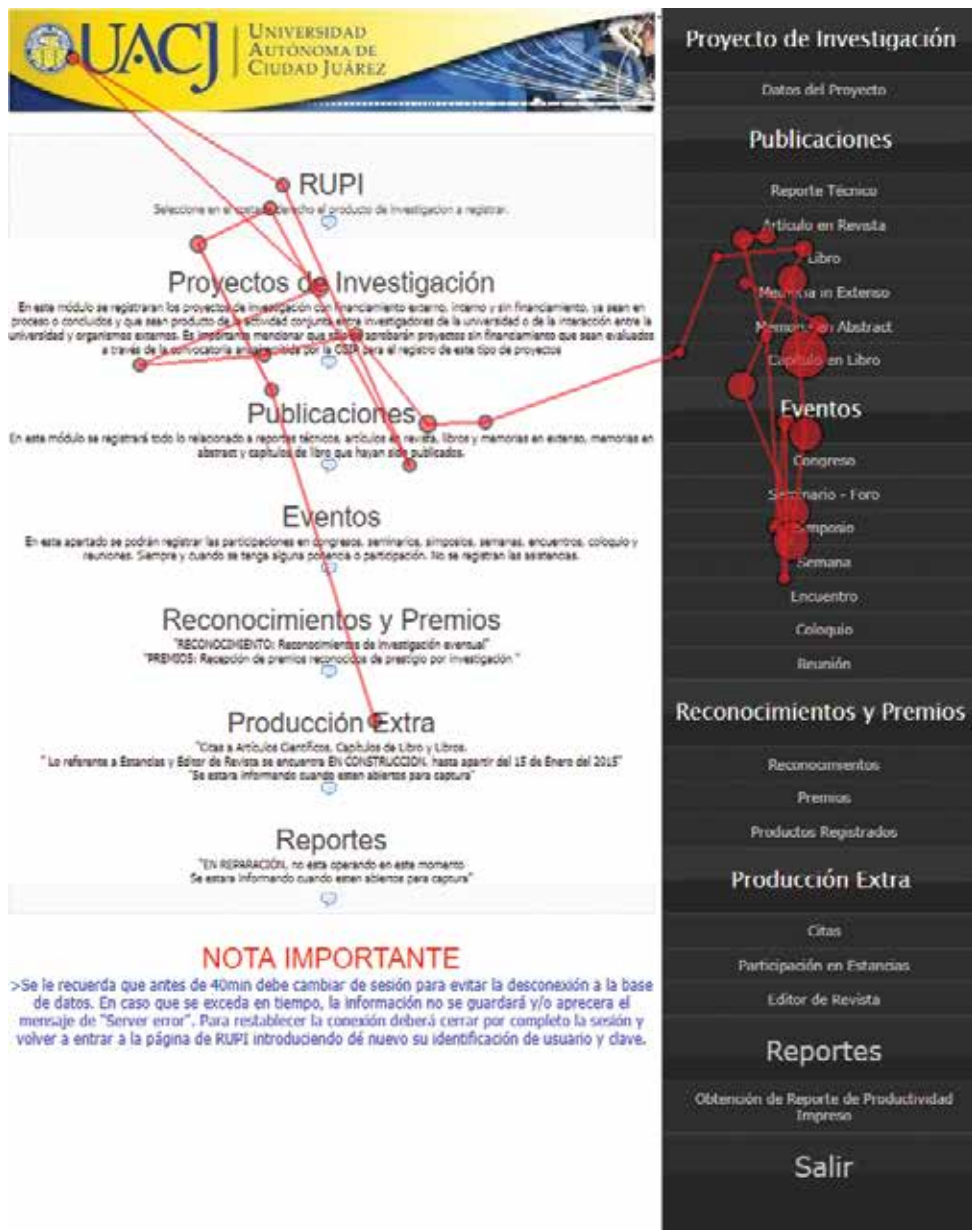
Figure 1. Saccade route generated by the eye tracker.

3. Related work

Some of the most representative research works that have used eye tracking data to solve usability problems in web pages are presented here.

The Centre for Human Computer Interaction Design of the City University of London [5] acquired experimental data by using a Tobii ×50 eye tracker. An ocular follow-up was performed while participants were doing tasks within two web pages. They found sequences of patterns of ocular tracking correlated with the sequences of patterns of usability problems, which are already known.

Similarly, a website for the American Society of Clinical Oncology (ASCO) was proposed, designed and evaluated with an eye tracking methodology [6]. Participants were asked to complete certain tasks, both in an old design of the page and in the new design. The paper describes how eye tracking helped to diagnose errors and identify the best of both designs.

Djamasbi et al. [7] conducted a study where they examined possible gender differences in web preference with the help of an eye tracker. Several studies show that both men and women have different esthetic preferences, for example, men prefer using darker colors, such as black and blue, while women prefer lighter colors. These preferences not only apply in colors but also in the website distribution of images and information. According to the previous studies, two hypotheses were presented: H1—participating women will notice more sections that have images of people more than men and H2—female participants will notice sections with a light-colored background more than men. The results did not show any significant difference between genders with respect to the number of times they were fixed in the sections of the web page. The results of the fixation analysis did not show significant differences between men and women with respect to sections with different background colors. Although the results did not support the hypotheses, the analyses yielded other data on the web page banners, suggesting that men and women can also ignore banner-type objects, regardless of whether these objects are designed considering their preferences.

Szymanski et al. [8] presented a method for evaluating web usability by combining kinetic gestures and eye tracking applications. All this was done by performing three tasks in an application that aims to deliver a list to users of all stores, restaurants and other places within a mall. The tasks performed were to find certain stores in a specific point in the commercial mall.

In the paper presented by Russell et al., at the annual British Human Computer Interaction Conference [9], it shows a study to discover the answer to two design issues both in online documents and on web pages. Research questions were: what is the best font size to read online? and what kind of source font is best for reading online? *Serif* or *Sans Serif*? The experiment was carried out with employees of a computer company, using an eye tracking device. Participants were given instructions to read and retain several texts with different sizes and types of source fonts, followed by answering three multiple-choice questions per text. Results show that larger fonts have a slight but not significant advantage in reading speed over small fonts. Results on the type of source between *Serif* and *Sans Serif* do not show significant differences in ocular follow-up and retention of texts.

Understanding the age-related differences in website navigation is instructive for website design, especially given by the growing number of older people using the Internet. Romano Bergstrom et al. [10] present the usability and data tracking of five independent usability studies of a website that included young and old participants. The results revealed age-dependent

differences in movement and eye functioning during website navigation in some of the sections. In particular, in one study, older participants were less precise, and in two studies, they took longer time to complete tasks compared to younger participants. In two studies, they observed the central part of the screen more frequently than younger participants and also observed the peripheral left side of the screen less frequently. Also, they took more time to cast first look at the peripheral top of the screen than the younger participants in the study. These results highlight the potential of age-related differences in performance while browsing websites and provide motivation for further exploration.

Tag-clouds are a compact list of keywords used in the web page interface, which allow users to browse and navigate through various documents or sections of those pages. In the article written by Montero et al. [11], a study was conducted to investigate the usability of tag-clouds. The experiment was performed in 17 participants, 6 women and 11 men aged between 29 and 49 years. Tasks asked to participants were divided into stages. First was a semantic search, where subjects were asked to search and click on a concept related to a certain topic, such as *Christianity*, where the expected target was the tag *religion*. The second stage consisted of a syntactic search, in which participants were asked to search and click on a tag that answered a certain question that was asked, such as searching for the name of the best-known operating system in the market, where the expected answer was *windows*. The results showed that both the size and the shape of the tag have a clear influence on the visual exploration of the users. They also showed results on the alphabetical order of the tags, which do not imply an improvement in terms of efficiency at the time of the location of the tags.
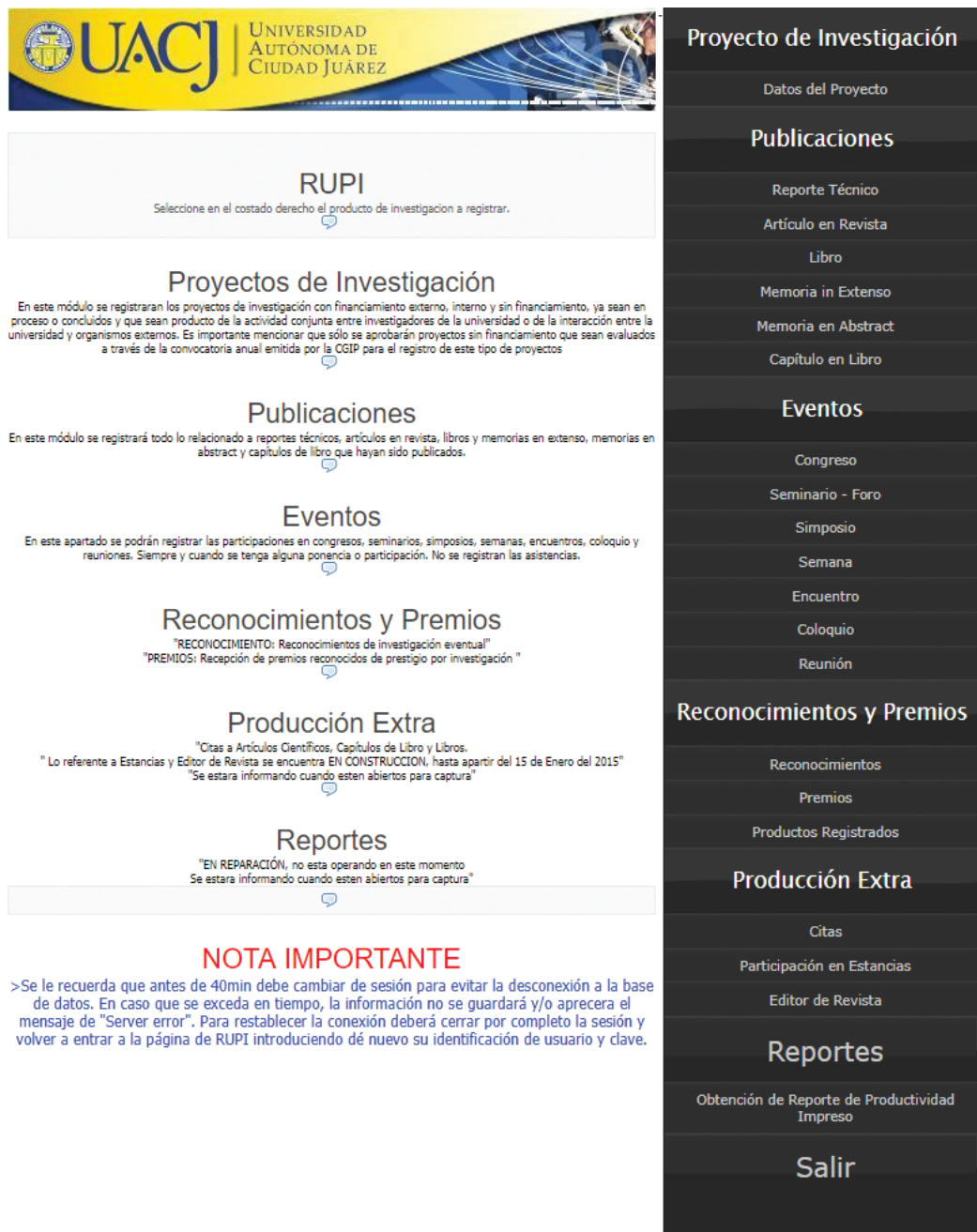
Most of the works reported here try to solve completely the usability problems either by using only eye tracking data or by a qualitative approach. In general, they only solve some aspects. We are convinced that by combining the best of both approaches (qualitative and quantitative) in usability measurements, results should be better. For this reason, this chapter proposes a combined approach of both.

4. Experimental approach

In order to study and get some data, we designed an experiment to face people to a simple task on a web page. The web page **R***egistro* **Ú***nico de* **P***roductos de* **I***nvestigacion* (RUPI) of the Universidad Autónoma de Ciudad Juárez (UACJ) was used as it is shown in **Figure 2**. During the experiment, participants were asked to fill forms with informational data provided to them beforehand. An eye tracker was used to get data from the user's gaze while conducting the experiment.

4.1. Participants

A total of nine full-time and two half-time professors of the Universidad Autonóma de Ciudad Juarez participated in the experiment. They were four women and seven men, aged between 28 and 43 years, and all of them are from the Electrical Engineering and Computer Science

Figure 2. RUPI home page.

Department. All participants had normal or corrected-to-normal visual acuity and had no known neurological disorders. They were naive with respect to the purpose of the study, and they gave written consent before the experiment.

4.2. Apparatus

Eye movements were recorded by using a remote binocular infrared S2 Eye Tracker, Mirametrix Inc., as mentioned in Section 2. The eye tracking system was used in the pupil corneal reflection tracking mode sampling at 60 Hz and 0.5–1 of accuracy. We ran our experiment by using the Viewer Software provided by the eye tracker manufacturer. Due to the fact that the eye tracker generates a special record of eye position and synchronization events, we had to perform a preprocessing stage in order to adequate this output to a more accessible format for our further analysis. A recording of both eyes' movements was done during the experiment; however, we are only concerned with the analysis of movement data of a single eye.

4.3. Training before the testing phase

Before starting the experiment, participants were trained on the tasks they would perform with an example of information to be managed during the testing phase. This step was done in an intuitive way without using the stimulus page of the real experiment. Additionally, some extra indications were given to the participants. Firstly, they were informed about the possibility to enter fictitious data but as similar as possible to real data only if they did not remember the precise information and secondly, the experiment could be interrupted at any time if they could not find the section where the information should be entered or whether they might feel lost within the website.

4.4. Calibration

The participants were asked to sit in front of the PC monitor at approximately 65 cm. Then, a screen with calibration points was presented as shown in **Figure 3**. The participants should look at each of these calibration points. Subsequently, the percentage of calibration error is displayed on the screen. It is noteworthy that the percentage average error was considered acceptable if it was less than 20%. **Figure 3** shows an example of calibration result considered not acceptable (average error: 35.1 > 20%) for our study purposes.

4.5. Experiment

After the training and calibration were completed, the participants were instructed to start the experimental session, in which they should complete the requested task. The initial screen was the home page of RUPI, where the participants had to find the section to enter the information that was given beforehand. The session ended when the participant pressed the button to save their data.

4.6. End of session

Once the session was finished, the participants were given a short survey about their experience with the website. This survey had questions concerning the graphical user interface and navigation on the page. They were also asked for any kind of help for future users of the website. Participants were also informed about the purpose of the experiment, and they had the opportunity to ask any question about our research project.
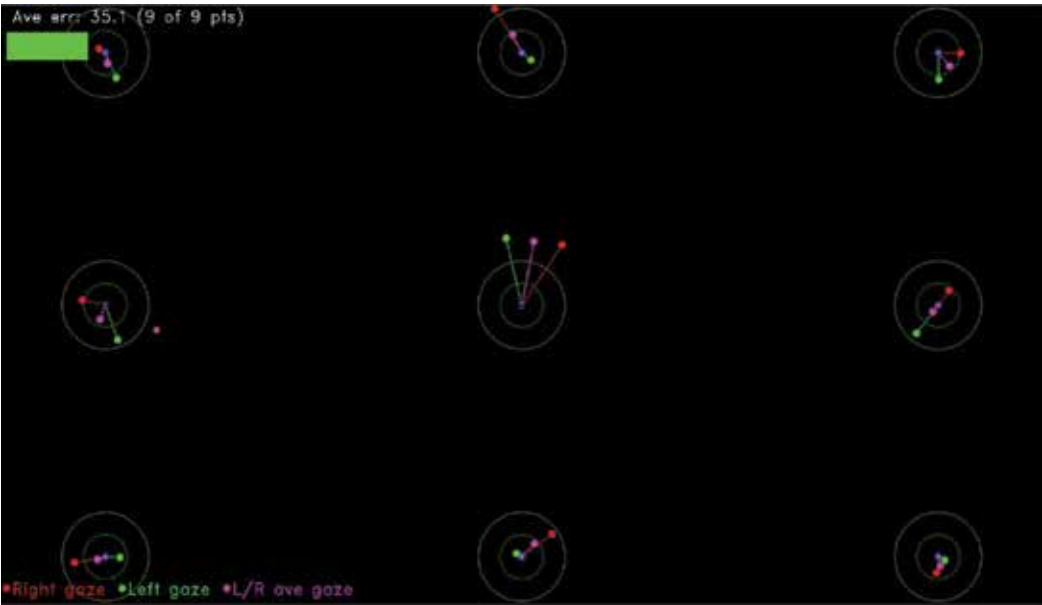
Figure 3. Calibration screen.
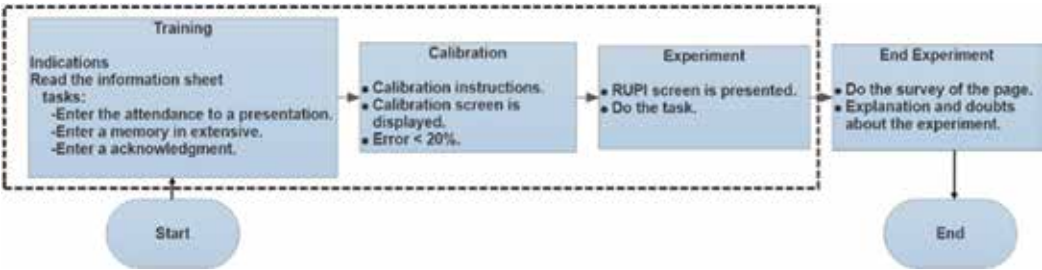
Figure 4. Overall phases performed during the experimental session. Phases bounded by dotted line were repeated three times by participants.

Figure 4 shows all the phases that were performed during the experimental session. Phases that are inside the dotted line are repeated three times for each task performed by the participant.

5. Data analysis

The main idea was to find usability problems by looking for regions or areas on the RUPI web page where participants made a lot of fixations. Our hypothesis was that if participants made a lot of fixation somewhere, this may be an indication of such a region which should be analyzed because people found it as being not clear or similar.

In this section, some statistics were computed from the experimental data. These computations are shown in **Table 1**.

P	Tj	TimeTj (s)	Fix	Order	OverallTime (s)	TotalFix	Fix/TimeTj (fix/s)	TotalFix/OverallTime (fix/s)
P1	1	154.95	9393	1			60.62	
	2	123.40	7487	2	358.93	21762	60.67	60.63
	3	80.58	4882	3			60.59	
P2	1	149.98	8862	2			59.09	
	2	161.40	9675	3	405.29	24242	59.94	59.81
	3	93.91	5705	1			60.75	
P3	1	125.81	7650	3			60.81	
	2	102.24	6212	1	364.03	21907	60.76	60.18
	3	135.98	8045	2			59.16	
P4	1	138.23	8398	2			60.75	
	2	199.90	11838	1	439.13	20236	59.22	46.08
	3	101.00	null	3			null	
P5	1	102.32	6213	2			60.72	
	2	528.84	31563	3	765.91	45966	59.68	60.01
	3	134.75	8190	1			60.78	
P6	1	94.48	5715	3			60.49	
	2	86.76	5192	2	292.33	17618	59.84	60.27
	3	111.09	6711	1			60.41	
P7	1	215.67	13079	1			60.64	
	2	285.65	17059	3	693.57	41827	59.72	60.31
	3	192.25	11689	2			60.80	
P8	1	118.92	7211	2			60.64	
	2	274.46	16635	1	594.09	36008	60.61	60.61
	3	200.71	12162	3			60.59	
P9	1	203.08	12340	1			60.76	
	2	370.39	20587	3	813.76	47172	55.58	57.97
	3	240.29	14245	2			59.28	
P10	1	282.24	17150	1	282.24	17150.00	60.76	60.76
P11	1	145.44	8847	1	145.44	8847.00	60.83	60.83

First task performed by participant

Table 1. Computation of statics with experimental data.

5.1. Statistics

Eqs. (1)–(3) were used to compute the elapsed time and total number of fixations in each task performed by participants and a ratio between such number of fixations and the elapsed time, respectively.

$$OverallTime = \sum_{j=1}^{3} Time_{T_j} \tag{1}$$

where $Time_{T_j}$ is the elapsed time (s) in each task performed by each participant,

$$TotalFix = \sum_{j=1}^{3} Fix_{T_j} \tag{2}$$

where Fix_{T_i} is the number of fixations in each task performed by each participant and

$$\overline{Fix/Time} = \frac{TotalFix}{OverallTime} \qquad (3)$$

giving the amount of fixations made for each participant per time unit. In other words, this variable is an indicator of "reading" speed.

5.1.1. Average time of the participants in completing the tasks

The average time to complete the first task for all participants was of 157.38 s ($\sigma = 55.77$), and for the second task, it was 237.00 s ($\sigma = 144.68$) and the average of the third task was 132.17 s ($\sigma = 72.50$).

It is important to mention that only two participants performed one task. For statistics computation, their data were aggregated to the other participants who did three tasks, so they were assigned as participant ID 10 and 11, respectively.

Additionally, **Figure 5** shows the number of fixations made in each task by participant, and the time of each participant in performing the three tasks was also computed and it is shown in **Figure 6**.

5.1.2. Scanpaths and selection of regions of interest

By using the geometric coordinates *(x, y)* and duration of the fixations, it is possible to draw the saccadic route of a participant in the web page. This saccadic route is known as scanpath. Once all the scanpaths of all participants were plotted, several regions of interest (ROIs) were selected by focusing for those with a high density of fixations.

Figure 7 shows the RUPI home page with the fixations of all scanpaths, and we only decided to analyze two ROIs in this interface: the *help menu* on the middle of the page (ROI$_1$) and the
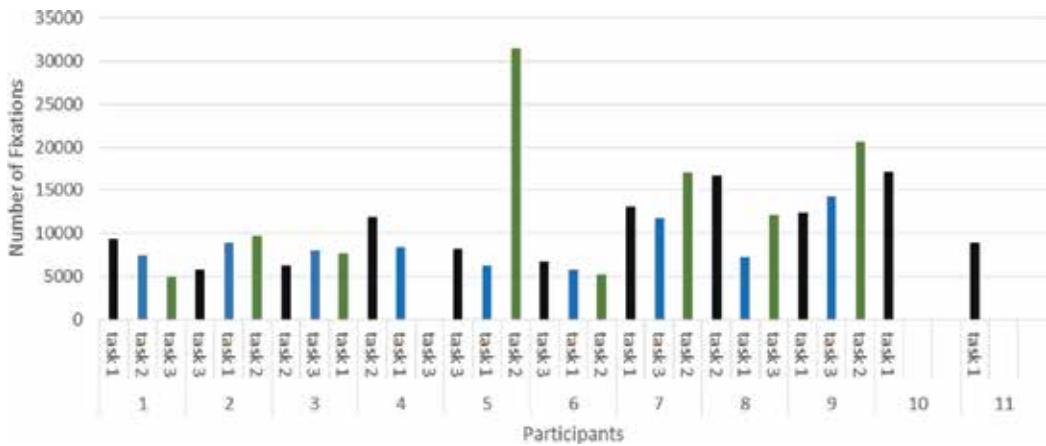
Figure 5. Number of fixations by participant. Tasks are presented in order performed.
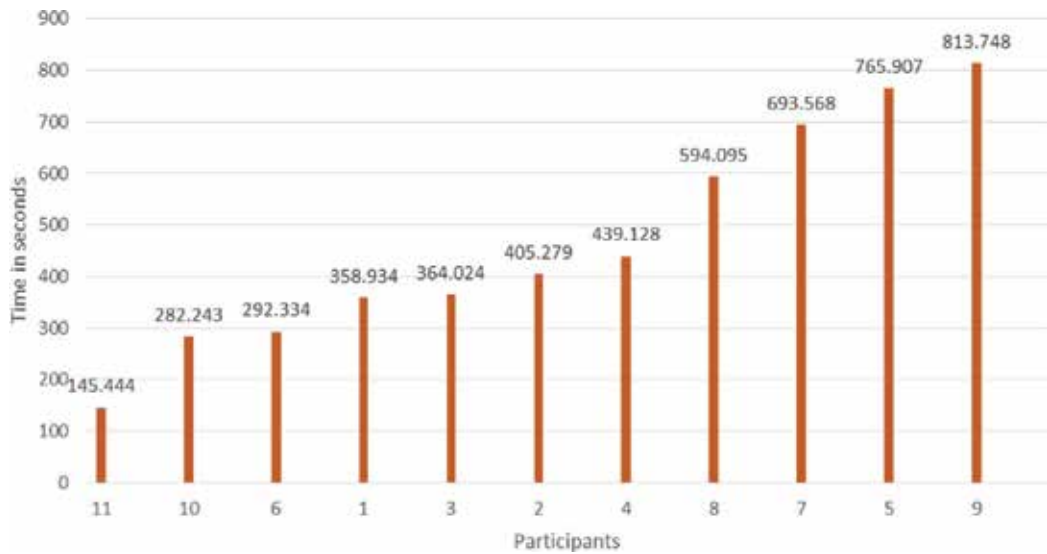
Figure 6. Time in completing the three tasks by participant ordered from lowest to highest.

right menu to enter the data on the right section (ROI_2), since those areas were the two with the highest fixation density as we can notice in such a figure.

Now, the qualitative data obtained from participants and how they could relate to the previous quantitative data are mentioned.

5.2. Qualitative data

After the experimental session, participants were asked to answer a 10-question short survey. This survey was intended in general to have information about the level of familiarization between the participant and RUPI web page, to know their experience during the experiment, to give their opinion about the graphical design of RUPI web page, and so on.

From the data obtained from the participant's survey (qualitative tests), the main results are described as follows:

- Most of the participants (8 of 11) have already once used the RUPI web page

- A total of 54.54% of the participants mentioned that the website's *help* section (ROI_1) to enter the data is not bad, and the other 45.4% mentioned that it is bad. This situation confirms our initial hypothesis that the website's *help* section is not clear, or it is not flashy

- A total of 45.45% of the participants had regular experience in finding the option to enter the information of the tasks, while 36.36% had bad experience. Only one participant could not enter the data and gave indications to end the session. It is important to mention that none of the participants entered the data in the corresponding section, since the RUPI web page is not well designed (ROI_2). For example, any information of a section can be entered in a different section.
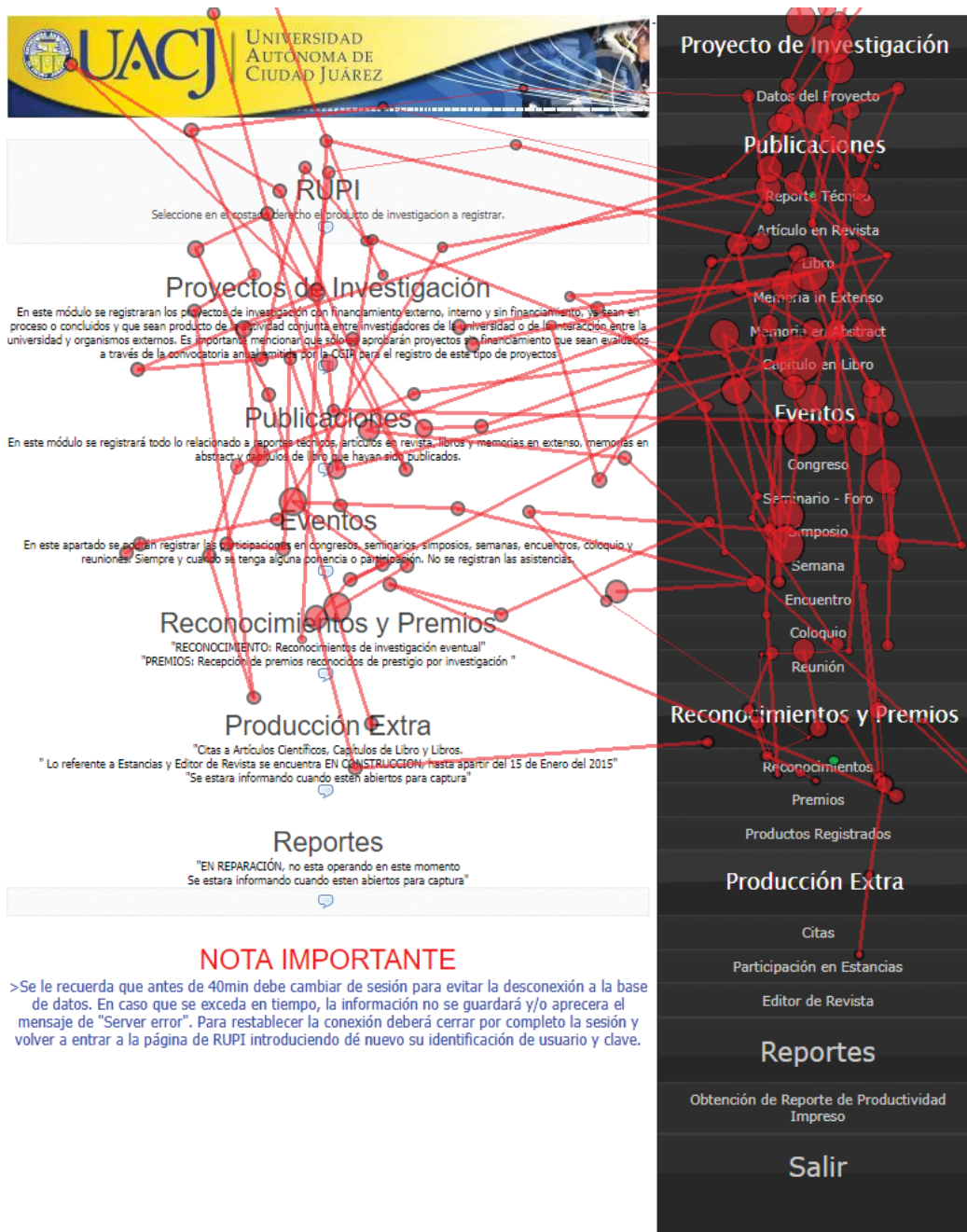
Figure 7. Scanpaths composed of fixations made by participants during the experiment.

5.3. Analysis of participants' behavior

In the ROI_2 located to the right of the web page, most of the participant fixations were observed when performing the very first task. Participants were concentrated in the *menu* area to enter

the options they were asked. Clearly, participants move a lot throughout the designated area as it is shown in **Figure 7**. This can give an indication that users get lost in the *menu* when trying to complete the first task.

Another point to mention is that there are very few fixations in the *help* area (ROI$_1$) of the website, giving way to understand several situations: (1) either this area was ignored often by users because it is not very flashy or (2) users who have already used the web page know that this section does not have clear information about the subsections and the data that should be entered in each one. This last assumption can be reinforced with the qualitative tests applied to the participants, since 54.54% of the participants mentioned that the website's *help* section to enter the data was not bad, and the other 45.45% mentioned that it was bad.

5.4. Redesigning and testing

Based on previous results and to test and confirm our findings quickly, simple modifications to the RUPI web page were proposed. Modifications to the visual and functional design were made. The scanpaths of participants testing our new design are shown in **Figure 8**. Here, fixations were better distributed among the different areas of the web page as it was expected.
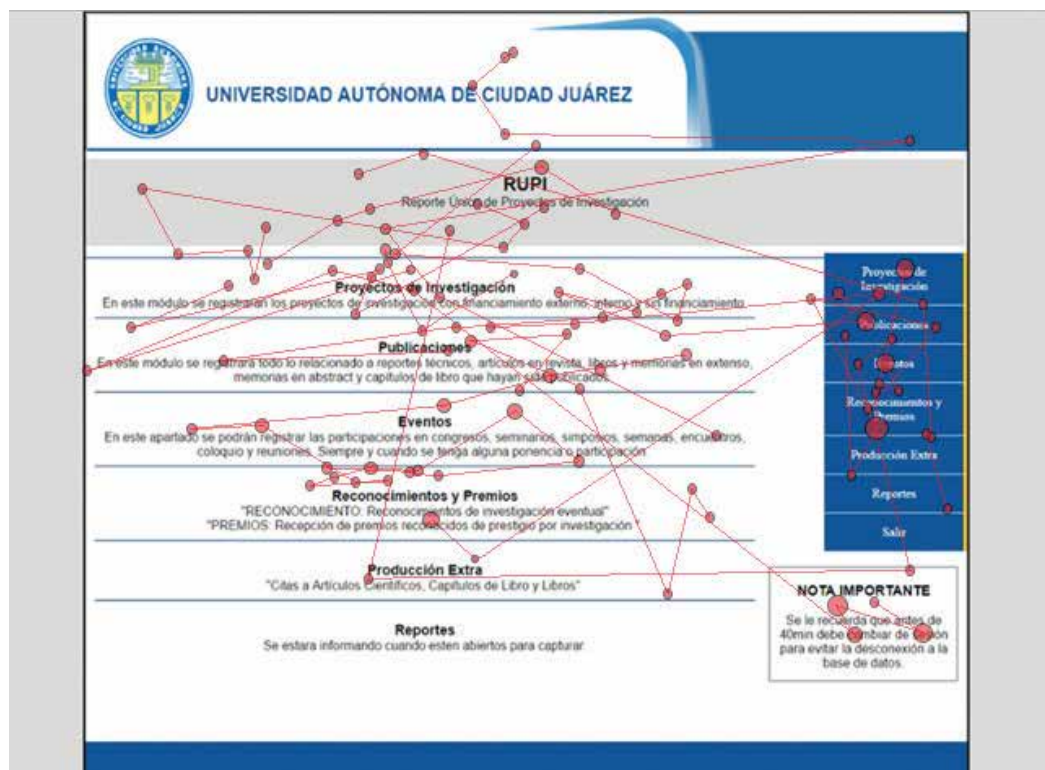
Figure 8. Scanpaths composed of fixations made by all participants during the experiment with the new web page design.

Moreover, 60% of participants rated the help section as good and 40% as regular. A total of 60% of participants entered the data successfully and 80% mentioned that they had not experienced any problem with the web page content. Finally, 60% of participants made plausible the fact that the new design has elements to allow users to go forward and go back while browsing.

6. Conclusions

In this chapter, a relatively new way of measuring usability by trying to combine the qualitative and quantitative techniques is proposed. Analysis of results showed how both techniques can be complemented. For example, an ROI where participants' rate was poor or marked as not helpful was *areas of help*. On the other hand, most of the works that try to solve usability problems by using the eye tracking technique generally do not focus on the qualitative part, as it could be seen in the state of the art, that is, qualitative measurement is not considered. More particularly, we have explored how the visual behavior of users can be combined with qualitative measurements to solve usability problems of the RUPI website. However, due to the complexity of human behavior and their reaction to a problem (in a web page) as well as the complexity of the required research, it is necessary to study this subject in more depth.

Acknowledgements

Thanks

Thanks to the student community involved during the experimental phase, specially to Mario Luna-Maldonado and Ramón G. Ortiz-Bustillos who carried out the experiment as part of their undergraduate project. Also, we thank the teaching staff of UACJ for accepting to participate in our experiments.

Author details

López-Orozco* and Florencia-Juárez

*Address all correspondence to: francisco.orozco@uacj.mx

Autonomous University of Ciudad Juárez, Ciudad Juárez, Mexico

References

[1] Sanchez W. La usabilidad en Ingeniería de Software: definicion y características, Ing-novacion. San Salvador: Universidad Don Bosco; Aug 2011

[2] Nielsen J. Designing web usability: The practice of simplicity. Thousand Oaks, CA, USA: New Riders Publishing; 1999

[3] Nielsen J, Pernice K. Eyetracking web usability. 1st ed. Thousand Oaks, CA, USA: New Riders Publishing; 2010

[4] Conte T, Vaz V, Massolar J, Mendes E, Travassos H. Improving a web usability inspection technique using qualitative and quantitative data from an observational study. In: Software Engineering, 2009. SBES'09. XXIII Brazilian Symposium. Fortaleza, Ceara, Brazil: IEEE; Oct 2009

[5] Ehmke C, Wilson S. Identifying web usability problems from eye-tracking data. In: Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI…But Not as We Know It. Vol. 1. British Computer Society. United Kingdom: University of Lancaster; Sep 2007. pp. 119-128

[6] Bojko A. Using eyetracking to compare web page designs: A case study. Journal of Usability Studies. 2006:**1**(3):112-120

[7] Djamasbi S, Tullis T, Hsu J, Mazuera E, Osberg K, Bosch J. Gender preferences in web design: Usability testing through eye tracking. In: Proceedings of the Thirteenth Americas Conference on Information Systems (AMCIS), Keystone, Colorado; August 2007. pp. 1-8

[8] Szymanski JM, Sobecki J, Chynał P, Anisiewicz J. Eye tracking in gesture based user interfaces usability testing. In: Asian Conference on Intelligent Information and Database Systems. Springer, Cham: Springer International Publishing; Mar 2015. pp. 359-366

[9] Beymer D, Russell D, Orton P. An eye tracking study of how font size and type influence online reading. In: Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction. Vol. 2. British Computer Society; Liverpool, United Kingdom. Sep 2008. pp. 15-18

[10] Romano Bergstrom JC, Olmsted-Hawala EL, Jans ME. Age-related differences in eye tracking and usability performance: Website usability for older adults. International Journal of Human-Computer Interaction. 2013;**29**(8)

[11] Montero YH, Herrero-Solana V, Guerrero-Bote V. Usabilidad de los tag-clouds: estudio mediante eye-tracking. In: Scire: Representacion y organizaci´on del conocimiento. Idioma: español; 2010;**16**(1):31-41. ISSN: 1135-3716

Deep Learning Models for Predicting Phenotypic Traits and Diseases from Omics Data

Md. Mohaiminul Islam, Yang Wang and
Pingzhao Hu

Additional information is available at the end of the chapter

Abstract

Computational analysis of high-throughput omics data, such as gene expressions, copy number alterations and DNA methylation (DNAm), has become popular in disease studies in recent decades because such analyses can be very helpful to predict whether a patient has certain disease or its subtypes. However, due to the high-dimensional nature of the data sets with hundreds of thousands of variables and very small number of samples, traditional machine learning approaches, such as support vector machines (SVMs) and random forests, have limitations to analyze these data efficiently. In this chapter, we reviewed the progress in applying deep learning algorithms to solve some biological questions. The focus is on potential software tools and public data sources for the tasks. Particularly, we show some case studies using deep neural network (DNN) models for classifying molecular subtypes of breast cancer and DNN-based regression models to account for interindividual variation in triglyceride concentrations measured at different visits of peripheral blood samples using DNAm profiles. We show that integration of multi-omics profiles into DNN-based learning methods could improve the prediction of the molecular subtypes of breast cancer. We also demonstrate the superiority of our proposed DNN models over the SVM model for predicting triglyceride concentrations.

Keywords: deep learning, omics data, phenotypic traits, data integration, support vector machine, random forest

1. Introduction

1.1. Omics data

Omics refers to use high-throughput experimental technologies to examine genomics, transcriptomics, metabolomics and proteomics for understanding biological and disease mechanisms. The omics data generated from these technologies are high-dimensional and correlated. Different computational and statistical analyses of these data can be used to identify risk factors for different diseases or to build autonomous disease prediction models. The technological development allows researchers to have a huge amount of high-dimensional biological data. The omics technologies generate such high-throughput data by detecting numerous alterations in molecular components [1]. These technologies also generate additional biological data to comprehend different types of correlations and dependencies among the molecular components. Bioinformatics is a discipline which emerges to perform computational analysis with the high-throughput biological data. Bioinformatics offers tools and methodologies for analyzing different omics data to understand the underlying information about different diseases. Such analyses will help physicians to provide early and patient-specific treatment. Schneider and Orchard [2] listed the state-of-the-art available technologies to generate omics data. They also provided the list of different available bioinformatics resources to analyze omics data and discussed the bioinformatics challenges to handle the high-throughput data.

For example, one of the key questions in medical science is to identify the specific mutations for a particular disease in individual patients. To do this, we need to first collect blood samples from the patients with the disease and healthy individuals without diseases. DNA will be then extracted from these blood samples. DNA microarray (also named as DNA chip) or DNA sequencing technologies will be then used to generate genome-wide genomic data for identifying mutated genes for the disease [2]. Similar technologies and procedures can also be used to generate other types of omics data such as RNA gene expression data and DNA methylation (DNAm) data.

1.2. Phenotypic traits prediction

A living biological organism can show a number of observable characteristics such as morphology, growth and behavior of the organism. Phenotypes are the product of different genetic expressions of an organism. These expressions are known as the genotype of that organism. However, phenotypic traits are the alternatives of a phenotype of a particular organism. For example, hair is a phenotype but different hair colors are the phenotypic traits. Study of the phenotypic traits prediction is very important as it gives us the knowledge about how genotype impacts upon an individual's diseases or traits. Lippert et al. [3] used the whole-genome sequencing data to identify individuals by predicting their biometric traits. Genome sequencing data were also used by Chen et al. [4] to build a probabilistic Bayesian model to predict the dichotomous traits (e.g., glaucoma, Corn's disease, prostate cancer). This model incorporates annotated information about different variant genotypes and genes, which are associated with diseases. There are other phenotypic trait prediction models such as eye color [5, 6], skin color [6] or facial structures [7].

1.2.1. Breast cancer and its molecular subtypes

Cancer is a collection of diseases that are characterized by uncontrolled cell growth in organs, that is, the sites the cells originate from. For example, breast cancer begins in the breast tissue and may start in the duct or lobe of the breast. When the breast cells are not working properly, they divide continually and a tumor is formed. Breast cancer is a complex and heterogeneous disease with differing prognostic and clinical outcomes.

In clinical practice, estrogen-receptor expression can be used to classify breast cancer patients into estrogen-receptor positive (ER+) or negative (ER−). A patient is classified as ER+ if the stimulation of growth of her cancer cells depends on the receptor for hormone estrogen. Stratification of breast cancer patients into ER+ or ER− is very important as physicians will use the information to determine whether the patients need chemotherapy or hormonal treatments. Statistics show that approximately 67% of breast cancer tests are positive for hormone receptors [8].

Perou et al. discovered the intrinsic subtypes of breast cancer using gene expression profiles of frozen tissue samples through unsupervised analysis [9]. They classified breast cancers into five groups such as normal-like, luminal A and B (Lumb A and B), basal-like and Her2-enriched (Her2). They established a polymerase chain reaction (PCR)-based test for 50 genes, and these genes are well-known as PAM50 signature. Gene expression levels of these 50 genes can be used to classify patients into one of these four subtypes: luminal A and B, basal-like and Her2-enriched. These subtypes are also known as PAM50 subtypes. This classification has been shown to be prognostically independent of clinicopathologic factors and can identify patients who are more likely to benefit from adjuvant chemotherapy [10]. It can also help identifying the fundamental differences among the PAM50 subtypes at the molecular level [11].

1.2.2. Prediction of triglyceride concentration in blood

Triglyceride is a type of fat in the human blood. Having a high concentration of triglycerides in human blood can increase our risk of heart diseases, stroke and other disorders. Many genetic loci have been identified by genome-wide association studies, but only a small proportion of interindividual variability of triglycerides has been explained by the genetic determinants. It is known that the level of triglycerides is heritable. Consequently, the development of new high-throughput genomic technologies makes it natural to extend these phenotypic prediction models to complex traits such as triglyceride. Using DNAm profiles to predict disease phenotypic courses has not yet been explored in detail.

1.3. Machine learning for disease phenotype prediction

Artificial intelligence (AI) is an area of computer science which demonstrates its necessity in our everyday life by machine learning (ML) methods. ML methods can automate the data analysis and can find the hidden intrinsic patterns from big data, which is impossible for a human being. ML methods use these patterns to build predictive models without any explicit programming. These predictive ML models are improving our daily life in various ways such as recommendations of different products during online shopping based on our searches of products, stock price prediction, classification of different objects from images, real-time language translation, and so on.

Conventional ML methods, such as support vector machine (SVM), random forest (RF), Bayesian network (BN), and so on, are dependent on the well-defined, engineered and robust hand-tuned features (or feature vectors) as inputs from the raw input data to make reasonable predictions. A domain human expertise is required to develop these engineered features. However, real-time biomedical data are often high-dimensional and noisy. These conventional ML methods are not capable enough to provide suitable techniques to handle such natural raw data (i.e., normalized gene expression data).

Different ML-based methods were developed to classify breast cancer patients into one of the PAM50 subtypes using gene expression profiles [12]. However, a new class of ML methods called deep learning (DL) can handle such high-dimensional, noisy and natural raw data by following representation learning or hierarchical data-driven approaches.

2. Deep learning and its application in bioinformatics

DL is a family of artificial neural network (ANN)-based ML methods which have been inspired by the working principles of a human brain. In a DL network architecture, a series of hidden layers are connected in a cascade fashion between input and output of the network. Each of these layers takes input from its previous layer and transforms the data into a more abstract form. Nonlinear layers allow DL methods to model complex relations between input and output of the network like shallow ANNs. DL is a representation learning method which means it can be fed with raw data and then, it will automatically extract necessary representation for predictions. A DL network provides representations at different levels. The output of each of hidden layers is considered as the representation at that level. The higher layers the data belong, the more abstract representations we get for these data. In different studies, these higher-level representations of raw data prove to be very effective for classification or detection problems. The most important thing here is that these representations, alternatively called feature vectors, are not learned by human engineering rather from the raw input data directly.

Unlike other ML methods, DL methods have been shown to efficiently handle high-dimensional and noise data in many domains such as computer vision, language processing (**Table 1**). These qualities of DL attract biomedical researchers to use DL instead of conventional ML methods because biomedical data (e.g., omics data) often suffer from high-dimensionality and noisiness.

2.1. How deep learning evolved?

Technological advancement and availability of large data sets allow researchers to rekindle their interest in deep neural networks. Recently, DNN-based models have achieved the state-of-the-art prediction performances at the cost of immense computational power. For example, Krizhevsky et al. [13] trained a deep convolutional neural network (DCNN)-based model with 1.2 million images to classify 1000 different classes, which took approximately 6 days to complete the training. They built this model by optimizing about 60 million learnable parameters. They won the ILSVRC-2012 competition with their state-of-the-art prediction performances in image classification.

Characteristics	Conventional machine learning (including shallow neural network)	Deep learning
Feature engineering	Handcrafted features created by experts are required	Perform automatic feature extraction
Problem-solving approach	Break the problems into small parts and solve them separately. Combine the results for the final output.	Can solve the problems in an end-to-end fashion
Execution time	Relatively much less time is required (ranging from a few seconds to a few hours)	Training requires long time for optimizing its thousands of parameters.
Interpretability	Relatively easy to interpret the reasoning	Hard to interpret the reasoning
High-throughput data	It is challenging to handle high-throughput data directly	Techniques like dropout enable high-throughput data processing relatively possible

Table 1. Comparison of conventional machine learning and deep learning.

In computer vision, there has always been a growing need to train visual recognition systems more generically so that a system trained on one visual recognition task (e.g., classification) could be easily adapted to another task (e.g., detection). To handle such a challenge of adapting the source domain to different target domains, many domain adaptation methods have been proposed [14–16]. For example, Donahue et al. [17] extracted "deep features" from a deep neural network trained on one computer vision task and shown the state-of-the-art performance on a variety of other tasks. The DCNN-based features are very powerful, generic and, thus, can be well adapted to different visual domains, which was also evidenced by Sharif Razavian et al. [18]. They proposed a DCNN-based pretrained model called OverFeat, which was applied to different tasks for which OverFeat was not trained. They achieved state-of-the-art prediction performances for object detection [19]. After having feature descriptors extracted from the first fully connected layer of OverFeat, they applied a linear SVM classifier to these features for image classification, scene recognition, fine-grained recognition and attribute detection on different data sets. A pretrained CNN is usually followed by domain-specific fine-tuning on data from the target domains, especially when training data are scarce. Following this approach, Girshick et al. [20] fine-tuned a CNN pretrained on ILSVRC2012 classification data set and achieved substantially better object detection performance on PASCAL VOC as compared to the standard models based on simple hand-engineered features.

DL-based models trained using a small data set usually show poor prediction performances over the test data. To tackle such kind of overfitting problem, a technique called dropout was proposed by Hinton et al. [21]. Dropout is a regularization technique, which randomly drops out a predefined ratio of neuron connections within a DL-based network's layer. This helps the model to learn a general weight for each of the neuron connections of that layer. This procedure allows the network to perform well over the test data. In addition, another technique called DropConnect was proposed as a generalization of dropout [22]. It randomly drops out of neuron connections from the network instead of randomly dropping neural connection from a layer. DropConnect provided improved prediction performances for different image recognition benchmark data sets over the dropout [22].

There is another class of DNN known as deep belief network (DBN). A DBN is composed of multiple layers where different layers have neural connections between them but not within the hidden units of a layer. Each of these layers is a restricted Boltzmann machines (RBMs). DBN follows an unsupervised layer-wise pretraining approach for these RBMs using a method called contrastive divergence [23]. DBN is useful to fine-tune a DNN when the number of training samples is small. In this case, a DBN is first trained and the optimized weights from this DBN are then used to fine-tune a DNN. Therefore, this DNN will start the training with these learned weights instead of training from the scratch. This leads the network to be convergent at early stage and improve prediction performances because these weights are neighboring to the best value of a converged model [24].

Stacked autoencoder is another type of DL-based method, which is used to create a vigorous high-level representation of its input using unsupervised ML approaches [24]. A stacked autoencoder can be built using a DNN by stacking layers on top of others. With this approach, we can increase or decrease the dimension of the input data. Vincent et al. showed that we can feed a high-level representation of input data learned from a stacked autoencoder into a SVM with improved prediction performance [25].

DL-based methods have already achieved state-of-the-art prediction performances in diverse fields such as image classification [26], object detection [13], speech recognition [27], and so on. However, DL methods also allow us to build state-of-the-art prediction models for sequential data [28, 29]. A series of functionally powerful deep learning tools have been implemented (**Table 2**), which can be publicly downloaded to perform these applications.

Software	Website	Open Source	Interface	Deep learning algorithms
Caffe	http://caffe.berkeleyvision.org/	Yes	Python, MATLAB	CNN, RNN
Deeplearning4j	https://deeplearning4j.org/	Yes	Java, Python	CNN, RNN, RBN, DBN
Keras	https://keras.io/	Yes	Python, R	CNN, RBN, DBN
Theano	http://deeplearning.net/software/theano	Yes	Python	CNN, RNN, RBN, DBN
Torch	http://torch.ch/	Yes	Lua, LuaJIT, C, OpenCL	CNN, RNN, RBN, DBN
Tensorflow	https://www.tensorflow.org/	Yes	Python, Java	CNN, RNN, RBN, DBN
MatConvNet	http://www.vlfeat.org/matconvnet/	Yes	C++, MATLAB	CNN, RNN, RBN, DBN
MXNet	https://mxnet.incubator.apache.org/	Yes	R, C++, Python, Julia, Scala, MATLAB	CNN, RNN, RBN, DBN

Table 2. A list of commonly used deep learning tools.

CNN, convolutional neural network; RNN, recurrent neural network; RBN, radial basis network; and DBN, deep brief network.

2.2. Application of deep learning to solve different bioinformatics applications

In recent years, deep learning has been successfully applied to answer many biological questions using diverse biological data sources (**Figure 1**). We briefly review these advancements in different bioinformatics applications in the following paragraphs.

Maienschein-Cline et al. used a conventional unsupervised machine learning approach to identify targeted genes which were regulated by transcription factors (TFs) by integrating high-throughput binding data and gene expression data [30]. Their experimental results showed that the conventional method faced challenges to handle such high-throughput data even though they may be very helpful for clinical diagnosis. Therefore, Dananee et al. used a DL-based method called stacked denoising autoencoder (SADE) [25] for such kind of analysis [31]. They first took a high-level representation of the input gene expression profiles. They then fed this higher-level representation of the data to a shallow neural network and a SVM. They demonstrated the improved prediction performances than the baseline principal component analysis (PCA) or kernel principal component analysis (KPCA) approaches. They also identified a set of genes from the connectivity matrices of their DL-based model. These genes can be investigated further to improve clinical diagnosis.

Somatic point mutation-based cancer classification (SMCC) becomes an attractive research problem as DNA sequencing technology allows us to generate a huge volume of sequencing data. Efficient analysis of the somatic point mutation data can lead to a better patient-specific personal therapy to cancer patients. This kind of somatic point mutation data usually suffer from large sparsity. Therefore, previous SMCC approaches were not able to provide clinically acceptable cancer classification results. However, recently a DL-based classifier known as DeepGene tried to overcome this limitation [32]. This model first filtered the gene data by mutation rate to remove irrelevant genes from them. Then, it indexed the gene data by their nonzero elements which let DeepGene overcome the data sparsity problem. Finally, the outputs of these two steps were fed into a DNN which performed automatic extraction of features for SMCC. DeepGene achieved ~67% prediction accuracy which is much better than the prediction performances of most baseline classifiers (i.e., SVM ~67%, k-nearest neighbors (KNN) ~42% and naïve Bayes (NB) ~9%).
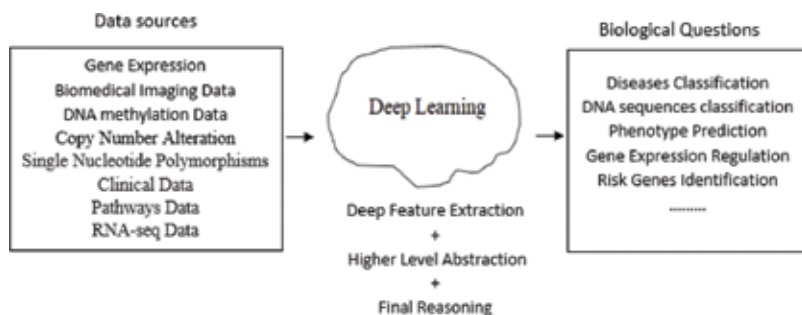
Figure 1. Scopes of deep learning in bioinformatics.

DBN was used to cluster cancer patients by integrating their gene expression and clinical data [33]. This model can capture intra- and cross-modality correlations (i.e., correlation among genomic data from different platforms) and learn a unified representation of the input. Therefore, this model outperformed existing methods in clustering cancer patients. This model can also be used to predict missing values in the data and identify key target genes of miRNAs responsible for different cancer subtypes. Moreover, preliminary clinical screening of a patient with skin disease usually begins with a visual diagnosis by a dermatologist. Since this is a very common malignancy in a human being [34, 35], an automatic system to classify skin diseases will be very helpful for the clinical purpose. Kelley et al. introduced a DCNN model to learn the functional activity of DNA sequences for 164 cell-specific DNA accessibility multitask prediction, and this model achieved the best result among earlier methods [36]. Recently, Esteva et al. collected 129,450 clinical images of skin diseases and built a DCNN model to classify the disease [37]. This model achieved better prediction performance than dermatologists. Nonetheless, a large number of nuclei and the variability in their sizes in histopathological images of breast cancer pose a great difficulty to build an automated system for nucleus detection. Xu et al. overcame this challenge by using a deep learning approach called stacked sparse autoencoder (SSAE). This model outperformed nine previous state-of-the-art nuclear detection methods [38].

Conventional machine learning approaches have been applied to analyze high-content microscopy data to protein subcellular localization from yeast cell images [39]. However, these approaches were not able to perform such analysis without human expert's intervention and yet did not provide accurate classification. Kraus et al. [40] came up with a model called DeepLoc which is a DCNN-based approach to overcome these limitations. DeepLoc outperforms the model ensLOC [39] by 71.4% according to mean average precision using fewer number of images. However, ensLOC uses binary SVM ensemble approach to assign single cells to subcellular compartment classes. Kraus et al. [40] also investigated the reason behind their success over ensLOC by performing 2D visualization of their network's components. They found out that DeepLoc generates a unique signal for different inputs. The structure of a protein and its functions can be studied further by protein contact map prediction from sequences. Wang et al. [41] treated this problem as a pixel-level labeling by considering a protein contact as an image. They proposed a novel deep learning-based protein contact map prediction model with extremely unbalanced positive and negative labels. Their model integrates two evolutionary couplings (EC) and sequence conservation information into their network. Their model gives the state-of-the-art performance result in protein contact map prediction. Furthermore, the predicted protein contacts by this model can generate an improved 3D structure model than the previous best models: CCMpred [42] and MetaPSICOV [43]. Besides, many biological processes such as signal transduction and cellular organization can be affected by different protein-protein interactions (PPI). Hence, it is very important to build a PPI prediction model in order to provide a better design for the therapy of a disease. Sun et al. [44] were the first one to build a deep learning-based model that is a stacked autoencoder for the sequence-based PPI prediction. They achieved an accuracy of 97.19% with 10-fold cross-validation which is better than any existing PPI predictors.

Genomics becomes rich with many different types of functional genomic data because of the latest sequencing technology. Eser et al. introduced flexible integration of data with deep learning

(FIDDLE), which is an open source DCNN-based data integration framework [45]. FIDDLE can predict yeast transcription start site sequencing (TSS-seq) [46] by the integration of heterogeneous genomic data such as RNA-seq and DNA sequence data. The excellent performance of FIDDLE demonstrated that a model built on integrating multiple data sets can provide better prediction performances than the models which were built using only one data set.

Chen et al. [47] proposed a deep learning system (D-GEX) which takes a gene's expression profile as input and infers the expression profile of a target gene. D-GEX has the ability to show cross-platform generalization. This model archives 15.33% improvement in gene expression prediction than a linear regression approach. D-GEX proves its cross-platform generalization when the learned D-GEX is used in RNA-seq-based database for gene expression prediction for each target gene and still outperforms LR by 6.57%.

Existing methods for the classification of cellular phenotypes from cellular images consist of multiple steps. Each of these steps is required with manual modifications and the tuning of different parameter settings. Godinez et al. [48] introduced a new multiscale CNN (M-CNN) network which uses microscopic images to classify them into phenotypes. The prediction performances of the M-CNN in terms of accuracy over eight benchmark data sets are significantly higher than the previous state-of-the-art methods including CNN-based approaches.

Gene expression can be regulated using transcription factors (TFs). Therefore, the cell-specific TF binding predictions using gold standard Chip-seq data are very important. Qin and Feng [49] introduced a DNN model termed TFImpute to achieve the abovementioned goal. TFImpute can determine whether a specific TF would bind to a given DNA sequence in a specific cell line. The prediction performance of TFImpute proves its superiority from the comparison with another latest DNN-based approach called DeepBind [50]. Therefore, biologists can use TFImpute to understand how TF binding can be included by a specific cell line.

Zhou et al. [51] were the first one to propose a DCNN-based approach to predict the effects of noncoding-variants from large-scale chromatin-profiling data and achieved state-of-the-art predictive performance. They call their method as deep learning-based sequence analyzer (DeepSea). Experimental results show that DeepSea can also precisely predict the consequence of specific SNPs on TF binding.

Obtaining precise knowledge about a patient's health condition is crucial to provide early and better treatment. Discovery of good imaging biomarkers can lead clinical research into achieving this goal. Oakden-Rayner et al. [52] provided a proof-of-concept research which proves that computer-based cross-sectional chest CT image analysis is able to predict 5 year mortality in adult (age > 60 years) person. Their framework includes deep learning model, and the predictive performances of this model are better than those who use human-generated features. Besides, visualization of different components of this deep learning-based model can provide an explanation about the better prediction performances [53].

Gene expression can be controlled by enhancer elements and cis-acting DNA regulatory elements [54]. However, existing enhancer predictors face a challenge, that is, the lack of availability of huge and experimentally confirmed enhancers for humans or other species. Yang et al. [55] developed a DNN-based hybrid architecture termed as BiRen which takes only DNA sequence as input to predict enhancers. Experimental results proved that BiRen can

predict common enhances more accurately than previous state-of-the-art methods, which are based on DNA sequence only.

Analysis of high-dimensional single-cell RNA-seq data is very important to answer several biological questions such as the amount of heterogeneity of cells in a population, the discovery of a biomarker for explicit cells and retrieving analogous cell types. Lin et al. [56] introduced an NN-based model to address all these queries without integrating any prior knowledge into the model. This method can deduce cell type more properly using a database of tens of thousands of single cell profiles than any existing methods.

Although the significant advancements have been made in applying DNN models to different bioinformatics applications as described earlier, here, we introduce several DNN-based classification frameworks which take either CNA profiles or gene expression profiles or both of them as input for the prediction of molecular subtypes of breast cancer. In addition, we also present a DNN-based regression model which takes high-dimensional DNAm data as input to predict triglyceride concentrations (before and after treatment) in the human blood.

3. Case studies

We discussed three case studies where deep learning-based approaches were used to address two biological problems such as classification of molecular subtypes of breast cancer (**Figure 2(a)** and **(b)**) and prediction of triglyceride concentration in human blood (**Figure 2(c)**).
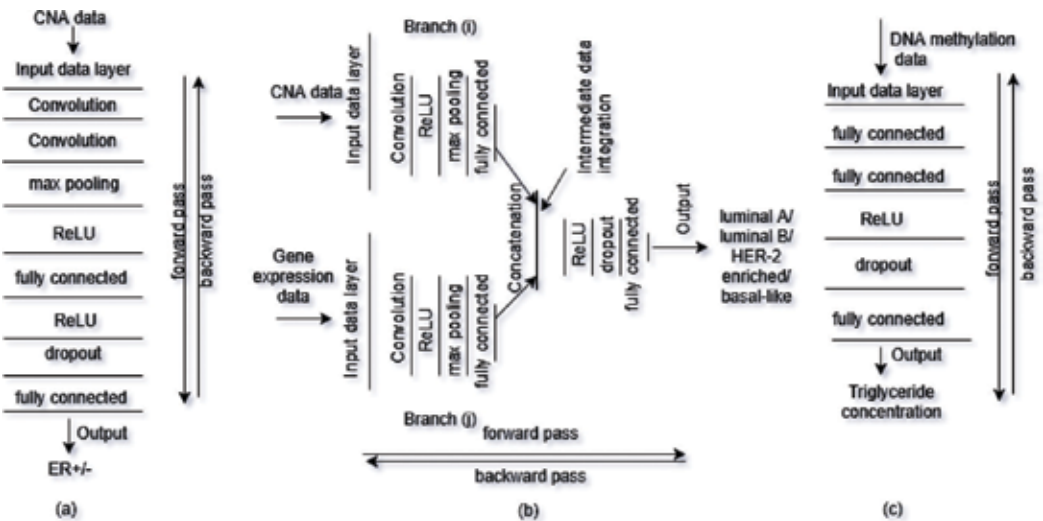
Figure 2. Deep learning architectures of three different case studies: (a) classification of molecular subtypes of breast cancer using single data source; (b) classification of molecular subtypes of breast cancer using multiple data sources; and (c) prediction of triglyceride concentration in human blood.

3.1. Prediction of molecular subtypes of breast cancer

Classification of molecular subtypes of breast cancer using high-throughput genomics data, such as gene expressions and copy number alterations (CNAs), is a challenging task because we have much smaller number of training samples than the number of genomic features. Many traditional machine learning methods often get overfitted over training data to handle such a problem. Here, we explored to use DNN methods to overcome this problem.

We conducted our experiments with two data sets: gene expression and CNA. Both data sets had approximate 2000 patients and larger than 16,000 genes. We collected these data sets from a project called Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [57]. We built a patient by gene mutation matrix from the CNA data where we had only three different values to represent three different copy number statuses of a gene in a patient: 1 = copy number gain, −1 = copy number loss and 0 = no change.

Using the CNA profiles of 991 and 984 patients as training and test sets, respectively, which was separated in original study [56] based on the patients collected at different periods, we classified the status of estrogen-receptor as a supervised binary classification problem. Furthermore, using the CNA profiles of 935 and 842 patients as training and test sets, respectively, we classified the status of PAM50 subtypes (luminal A, luminal B, HER-2 enriched and basal-like) as a supervised multiclass classification problem (it should be noted that some patients have no status of the PAM50 subtypes). Our DCNN-based deep learning architecture used for the tasks is shown in **Figure 2(a)** [58].

Our network took a CNA profile with approximate 16,000 genes of a breast cancer patient as the input in the input data layer. The input layer passed this input data to a convolutional layer. In the convolutional layer, we passed a kernel with weights over the input data and performed an element-wise multiplication with the input data. Then, we took the sum of the elements from this operation to represent the input data. This operation helped us capture the potential correlations among the neighbor genes. Output from the convolutional layers went as an input into max pooling layer. Here, we also passed a kernel without any weights over the input of this layer. We then took the maximum value from the input data under the kernel to represent the input data. This operation helped the model to achieve three invariant properties such as scaling, translation, and rotation. We introduced nonlinearity into our model by passing the output from max pooling layer through a rectified linear unit (ReLu) layer. ReLu layer performed a thresholding operation by converting all values less than zero to zero. At this point, we passed the output from the ReLu layer to a fully connected layer to get the higher level abstraction. We then passed the output from this fully connected to a dropout layer where we used another ReLu between them. Dropout layer randomly dropped out a number of neurons from the network. Dropout strategy enforced the network to learn a general weight for each of the neurons of the network. This helped the model to achieve better generalization ability over the test data as well as prevented the model from the overfitting problem. Then, we used another fully connected layer to get the prediction score for each of the classes. We converted this class score into class probability using softmax function. We trained our network in backpropagation style.

We evaluated the prediction performances of our network by two metrics such as overall accuracy and area under the curve (AUC). We achieved 84.1% accuracy and 0.904 AUC for binary classification. On the other hand, for multiclass classification, we achieved 58.19% accuracy and 0.79 AUC. We compared our experimental outcomes with other two supervised machine learning methods such as support vector machine (SVM) and random forest (RF). Prediction performances of SVM for the binary classification were 76.5% accuracy and 0.702 AUC while 45.0% accuracy and 0.78 AUC for multiclass classification. With RF, we had 82.7% accuracy and 0.817 AUC for binary classification and 49.5% accuracy and 0.729 AUC for multiclass classification. Our experimental outcomes showed that deep learning-based models have better performance than the traditional machine learning methods such as SVM and RF.

We also explored to build DCNN-based network (**Figure 2(b)**) to classify the status of PAM50 subtypes using both gene expression and CNA data. One of these data sets may have biological knowledge which is absent in another one. Our network shown in **Figure 2(b)** had two branches: i and j branches which took CNA data and the gene expression data of the same patient as inputs, respectively. Each of these branches performed the same series of operations over the input data. Outputs of the last fully connected layers from both branches were concatenated to be fed into the third part of the network which provided the final predictions. Here, we also used overall accuracy and AUC as the performance measurement metric. In terms of prediction performances, our model achieved 79.2% accuracy and 0.85 AUC. We compared our prediction performances with two baseline models: SVM and RF. Their prediction performances in terms of overall accuracy were 69.5 and 70.1%, respectively. Besides, our baseline models provided AUC 0.804 and 0.781, respectively. From these comparisons, we can clearly see that our deep learning-based data integration model (**Figure 2(b)**) achieved improved prediction performances than all of our baseline models. Our data integration model also outperformed the models (**Figure 2(a)**) which were built using only one data source as input to the architecture.

3.2. Prediction of triglyceride concentration

Triglyceride is a kind of fat which is found in human blood. Excessive triglyceride concentration can cause different heart-related diseases such as stroke. In this experiment, we built a DNN-based regression model [59], which took epigenome-wide DNA methylation profiles as input to predict triglyceride concentration at multiple draw of peripheral human blood samples.

For our experiments, we collected the epigenome-wide DNAm profiles and triglyceride concentrations (mg/dL) measured at the baseline level (pretreatment) of visit 2 and changes in response to treatment with fenofibrate (posttreatment) at visit 4 from Genetic Analysis Workshop 20 (GAW20). The DNAm profiles were generated using the Illumina Infinium HumanMethylation450 BeadChip array. The beta value measuring the methylation level was expressed between 0 and 1 in 993 samples of the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study. It should be noted that there were only 499 samples with the posttreatment DNAm data. Here, we built three regression models to solve three tasks: (1) we used pretreatment data to predict triglyceride concentration before the medication. We built this model using 900 randomly selected training samples, and we had 93 test samples for

the purpose of testing the model's prediction performances; (2) we used posttreatment data to predict triglyceride concentration after the medication. In this case, we had 400 randomly selected training samples and 99 test samples; (3) we used pretreatment data to predict triglyceride concentration after the medication. Here, we had 620 randomly selected training samples and 94 test samples.

Our DNN-based regression model (**Figure 2(c)**) is a fully connected neural network with four hidden layers between input and output layers. This network took a vector of size larger than 450,000 which represented an epigenome-wide DNA methylation profile of a sample. This vector went through two full connected layers, and the output is the higher level abstraction of the input. Then, we used a ReLU layer to introduce nonlinearity into the model, and the output of ReLU went into a dropout layer to achieve the generalization ability to test the data. The output of dropout layer then went into the final output layer and provided the predicted triglyceride concentration.

We used root mean square error (RMSE) and Pearson correlation coefficient (PCC) to measure our prediction performances. Prediction performances of our DNN-based regression models are as follows: for Task 1, we had RMSE 88.5 and PCC 0.27; for Task 2, we had RMSE 48.1 and PCC 0.22 and for Task 3, we had RMSE 47.4 and PCC 0.29. We used SVM as our baseline model. SVM provided prediction performances are as follows: for Task 1, the RMSE was 90.3 and PCC was 0.13; for Task 2, the RMSE was 48.7 and PCC was 0.19 and for Task 3, the RMSE was 46.9 and PCC was 0.13.

Experimental results showed that our DNN-based triglyceride concentration regression models provided improved prediction performances for all three tasks than the baseline SVM-based regression method. Both our DNN-based and SVM-based regression models achieve best prediction performances when we used the pretreatment data to predict triglyceride concentration after the medication. This outcome shows that there is a long-term epigenetic effect on the phenotypic traits.

4. Conclusion

Classification of molecular subtypes of a disease using omics profiles is a challenging problem since the data sets are quite high-dimensional and highly correlated. The curse of high-dimensionality also affects the performance of predicting a phenotype using DNAm data. Traditional machine learning algorithms, such as SVM and RF, have potential challenges to handle high-dimensional and highly correlated data sets. Recently, DNN learning has been demonstrated advantages over these methods since it does not require any hand-crafted features. DNN learning automatically extracts features from the raw data and efficiently analyzes high-dimensional and correlated data. In this chapter, we reviewed the status of applying DNN in bioinformatics and showed some case studies which introduced several DNN frameworks for classifying molecular subtypes of breast cancer using only one data source or two heterogeneous data sources. In addition, we also presented a DNN-based regression framework which took epigenome-wide DNAm data as input to predict triglyceride concentrations

in human blood. In summary, our and others' works have demonstrated that DNN is a promising tool to predict phenotypic traits and diseases from genome-wide omics data.

Acknowledgements

Conflict of interest

The authors declare that they have no competing interests.

Author details

Md. Mohaiminul Islam[1,2,3], Yang Wang[2] and Pingzhao Hu[1,2,3,4]*

*Address all correspondence to: pingzhao.hu@umanitoba.ca

1 Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, Manitoba, Canada

2 Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada

3 George and Fay Yee Centre for Healthcare Innovation, University of Manitoba, Winnipeg, Manitoba, Canada

4 Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Manitoba, Canada

References

[1] Knasmüller S, Nersesyan A, Mišík M, Gerner C, Mikulits W, Ehrlich V, Hoelzl C, Szakmary A, Wagner KH. Use of conventional and-omics based methods for health claims of dietary antioxidants: A critical overview. British Journal of Nutrition. 2008;**99**(E-S1):ES3-E52. DOI: 10.1017/S000711450896575

[2] Schneider MV, Orchard S. Omics technologies, data and bioinformatics principles. In: Bioinformatics for Omics Data: Methods and Protocols. 2011. pp. 3-30

[3] Lippert C, Sabatini R, Maher MC, Kang EY, Lee S, Arikan O, Harley A, Bernal A, Garst P, Lavrenko V, Yocum K. Identification of individuals by trait prediction using whole-genome sequencing data. Proceedings of the National Academy of Sciences. 2017 Sep 19;**114**(38):10166-10171

[4] Chen YC, Douville C, Wang C, Niknafs N, Yeo G, Beleva-Guthrie V, Carter H, Stenson PD, Cooper DN, Li B, Mooney S. A probabilistic model to predict clinical phenotypic traits from genome sequencing. PLoS Computational Biology. 2014;**10**(9):e1003825. DOI: 10.1371/journal.pcbi.1003825

[5] Liu F, Wen B, Kayser M. Colorful DNA polymorphisms in humans. Seminars in Cell & Developmental Biology. 2013;**24**(6):562-575. DOI: 10.1016/j.semcdb.2013.03.013

[6] Hart KL, Kimura SL, Mushailov V, Budimlija ZM, Prinz M, Wurmbach E. Improved eye- and skin-color prediction based on 8 SNPs. Croatian Medical Journal. 2013;**54**(3):248-256. DOI: 10.3325/cmj.2013.54.248

[7] Claes P, Liberton DK, Daniels K, Rosana KM, Quillen EE, Pearson LN, McEvoy B, Bauchet M, Zaidi AA, Yao W, Tang H. Modeling 3D facial shape from DNA. PLoS Genetics. 2014;**10**(3):e1004224. DOI:/10.1371/journal.pgen.1004224

[8] Breast Cancer Information and Awareness. Available from: http://www.breastcancer. org. [Accessed: 2017-10-20]

[9] Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge Ø. Molecular portraits of human breast tumours. Nature. 2000;**406**(6797):747-752. DOI: 10.1038/35021093

[10] Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of Clinical Oncology. 2009;**27**(8):1160-1167. DOI: 10.1200/JCO.2008.18.1370

[11] Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences of the United States of America. 2003;**100**(14):8418-8423. DOI: 10.1073/pnas.0932692100

[12] Milioli HH, Vimieiro R, Tishchenko I, Riveros C, Berretta R, Moscato P. Iteratively refining breast cancer intrinsic subtypes in the METABRIC dataset. BioData Mining. 2016;**9**(1):2. DOI: 10.1186/s13040-015-0078-9

[13] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. 2012. pp. 1097-1105

[14] Khosla A, Zhou T, Malisiewicz T, Efros AA, Torralba A. Undoing the damage of dataset bias. In: European Conference on Computer Vision. 2012. pp. 158-171. DOI: 10.1109/CVPR.2011.5995347

[15] Saenko K, Kulis B, Fritz M, Darrell T. Adapting visual category models to new domains. Computer Vision–ECCV. 2010;**2010**:213-226

[16] Aytar Y, Zisserman A. Tabula rasa: Model transfer for object category detection. In: IEEE International Conference on Computer Vision (ICCV); 2011 Nov 6; IEEE. pp. 2252-2259

[17] Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T. Decaf: A deep convolutional activation feature for generic visual recognition. In: International Conference on Machine Learning; 2014; p. 647-655

[18] Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: An astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014. pp. 806-813

[19] Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229. 2013

[20] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014; p. 580-587

[21] Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580; 2012

[22] Wan L, Zeiler M, Zhang S, Cun YL, Fergus R. Regularization of neural networks using dropconnect. In: Proceedings of the 30th International Conference on Machine Learning (ICML-13). 2013. pp. 1058-1066

[23] Hinton GE. Training products of experts by minimizing contrastive divergence. Training. Neural Computation. 2006;**14**(8):1771-1800. DOI: 10.1162/089976602760128018

[24] Larochelle H, Erhan D, Courville A, Bergstra J, Bengio Y. An empirical evaluation of deep architectures on problems with many factors of variation. In: Proceedings of the 24th International Conference on Machine Learning 2007. pp. 473-480

[25] Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research. 2010;**11**:3371-3408

[26] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC. Imagenet large scale visual recognition challenge. International Journal of Computer Vision. 2015;**115**(3):211-252. DOI: 10.1007/s11263-015-0816-y

[27] Abdel-Hamid O, Mohamed AR, Jiang H, Deng L, Penn G, Yu D. Convolutional neural networks for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2014;**22**(10):1533-1545. DOI: 10.1109/TASLP.2014.2339736

[28] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;**39**(6):1137-1149. DOI: 10.1109/TPAMI.2016.2577031

[29] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. Journal of Machine Learning Research. 2011; **12**:2493-2537

[30] Maienschein-Cline M, Zhou J, White KP, Sciammas R, Dinner AR. Discovering transcription factor regulatory targets using gene expression and binding data. Bioinformatics. 2011;**28**(2):206-213. DOI: 10.1038/srep20649

[31] Danaee P, Ghaeini R, Hendrix DA. A deep learning approach for cancer detection and relevant gene identification. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. 2016;**22**:219. DOI: 10.1142/9789813207813_0022

[32] Yuan Y, Shi Y, Li C, Kim J, Cai W, Han Z, Feng DD. DeepGene: An advanced cancer type classifier based on deep learning and somatic point mutations. BMC Bioinformatics. 2016;**17**(17):476. DOI: 10.1186/s12859-016-1334-9

[33] Liang M, Li Z, Chen T, Zeng J. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2015;**12**(4):928-937. DOI: 10.1109/TCBB.2014.2377729

[34] American Cancer Society. Cancer facts & figures 2016. Atlanta: American Cancer Society; 2016. Available from: http://www.cancer.org/acs/groups/content/@research/documents/document/acspc047079.pdf. [Accessed: 2017-10-10]

[35] Stern RS. Prevalence of a history of skin cancer in 2007: Results of an incidence-based model. Archives of Dermatology. 2010;**146**(3):279-282. DOI: 10.1001/archdermatol.2010.4

[36] Kelley DR, Snoek J, Rinn JL. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Research. 2016;**26**(7):990-999. DOI: 10.1101/gr.200535.115

[37] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;**542**(7639):115-118. DOI: 10.1038/nature21056

[38] Xu J, Xiang L, Liu Q, Gilmore H, Wu J, Tang J, Madabhushi A. Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. IEEE Transactions on Medical Imaging. 2016;**35**(1):119-130. DOI: 10.1109/TMI.2015.2458702

[39] Chong YT, Koh JL, Friesen H, Duffy SK, Cox MJ, Moses A, Moffat J, Boone C, Andrews BJ. Yeast proteome dynamics from single cell imaging and automated analysis. Cell. 2015;**161**(6):1413-1424. DOI: 10.1016/j.cell.2015.04.051

[40] Kraus OZ, Grys BT, Ba J, Chong Y, Frey BJ, Boone C, Andrews BJ. Automated analysis of high-content microscopy data with deep learning. Molecular Systems Biology. 2017;**13**(4):924. DOI: 10.15252/msb.20177551

[41] Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Computational Biology. 2017;**13**(1):e1005324

[42] Seemayer S, Gruber M, Söding J. CCMpred—Fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics. 2014;**30**(21):3128-3130. DOI: 10.1371/journal.pcbi.1005324

[43] Jones DT, Singh T, Kosciolek T, Tetchner S. MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics. 2014;**31**(7):999-1006. DOI: 10.1093/bioinformatics/btu79

[44] Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinformatics. 2017;**18**(1):277. DOI: 10.1186/s12859-017-1700-2

[45] Eser U, Churchman LSFIDDLE. An integrative deep learning framework for functional genomic data inference. Cold Spring Harbor Laboratory. Biorxiv. 2016. DOI: 10.1101/081380

[46] Malabat C, Feuerbach F, Ma L, Saveanu C, Jacquier A. Quality control of transcription start site selection by nonsense-mediated-mRNA decay. eLife. 2015;**4**:e06722. DOI: 10.7554/eLife.06722

[47] Chen Y, Li Y, Narayan R, Subramanian A, Xie X. Gene expression inference with deep learning. Bioinformatics. 2016;**32**(12):1832-1839. DOI: 10.1093/bioinformatics/btw074

[48] Godinez WJ, Hossain I, Lazic SE, Davies JW, Zhang X. A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics. 2017;**33**(13):2010-2019. DOI: 10.1093/bioinformatics/btx069

[49] Qin Q, Feng J. Imputation for transcription factor binding predictions based on deep learning. PLoS Computational Biology. 2017;**13**(2):e1005403. DOI: 10.1371/journal.pcbi.1005403

[50] Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology. 2015;**33**(8):831-838. DOI: 10.1038/nbt.3300

[51] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nature Methods. 2015;**12**(10):931-934. DOI: 10.1038/nmeth.3547

[52] Oakden-Rayner L, Carneiro G, Bessen T, Nascimento JC, Bradley AP, Palmer LJ. Precision radiology: Predicting longevity using feature engineering and deep learning methods in a radiomics framework. Scientific Reports. 2017;**7**(1):1648. DOI: 10.1038/s41598-017-01931-w

[53] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. 2014. pp. 818-833

[54] Calo E, Wysocka J. Modification of enhancer chromatin: What, how, and why? Molecular Cell. 2013;**49**(5):825-837. DOI: 10.1016/j.molcel.2013.01.038

[55] Yang B, Liu F, Ren C, Ouyang Z, Xie Z, Bo X, Shu W. BiRen: Predicting enhancers with a deep-learning-based model using the DNA sequence alone. Bioinformatics. 2017 Feb 17;**33**(13):1930-1936. DOI: 10.1093/bioinformatics/btx105

[56] Lin C, Jain S, Kim H, Bar-Joseph Z. Using neural networks for reducing the dimensions of single-cell RNA-Seq data. Nucleic Acids Research. 2017;**45**(17):e156. DOI: 10.1093/nar/gkx681

[57] Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;**486**(7403):346-352. DOI: 10.1038/nature10983

[58] Islam MM, Ajwad R, Chi C, Domaratzki M, Wang Y, Hu P. Somatic copy number alteration-based prediction of molecular subtypes of breast cancer using deep learning model. 30th Canadian Conference on Artificial Intelligence. 2017 May 16:57-63

[59] Islam MM, Tian Y, Cheng Y, Wang Y, Hu P. A deep neural network regression model for triglyceride concentrations prediction using epigenome-wide methylation profiles. BMC Proceedings. 2018 (In press)

Can Reinforcement Learning Be Applied to Surgery?

Masakazu Sato, Kaori Koga, Tomoyuki Fujii and
Yutaka Osuga

Additional information is available at the end of the chapter

Abstract

Background: Remarkable progress has recently been made in the field of artificial intelligence (AI).

Objective: We sought to investigate whether reinforcement learning could be used in surgery in the future.

Methods: We created simple 2D tasks (Tasks 1–3) that mimicked surgery. We used a neural network library, Keras, for reinforcement learning. In Task 1, a Mac OS X with an 8 GB memory (MacBook Pro, Apple, USA) was used. In Tasks 2 and 3, a Ubuntu 14. 04LTS with a 26 GB memory (Google Compute Engine, Google, USA) was used.

Results: In the task with a relatively small task area (Task 1), the simulated knife finally passed through all the target areas, and thus, the expected task was learned by AI. In contrast, in the task with a large task area (Task 2), a drastically increased amount of time was required, suggesting that learning was not achieved. Some improvement was observed when the CPU memory was expanded and inhibitory task areas were added (Task 3).

Conclusions: We propose the combination of reinforcement learning and surgery. Application of reinforcement learning to surgery may become possible by setting rules, such as appropriate rewards and playable (operable) areas, in simulated tasks.

Keywords: hysterectomy, robotic surgical procedures, deep learning, artificial intelligence, deep Q network (DQN), reinforcement learning

1. Introduction

Remarkable progress has recently been made in the field of artificial intelligence (AI), and new advancements are made almost every day [1–3]. Although it is not possible to discuss

all developments in AI, such as image recognition, automatic driving or trials of the game of Go, a particularly promising AI task is game playing [4–6]. AI can be trained to play games, such as the game Breakout, and can perform as well as an experienced human subject. In the context of game play, reinforcement learning is a common strategy [6, 7]. For instance, the objective of the game Breakout is to break blocks to score points. While the performance of the player at the instance of breaking a block should not necessarily be evaluated, the performance immediately prior to successfully breaking a block, such as the player's ability to direct the ball, should be evaluated. With this aim, the AI calculates the expected values of the scores (reward) that could be obtained by various actions and learns how to achieve the maximum reward via reinforcement learning [6]. Thus, the AI succeeds in breaking more blocks and obtaining higher scores.

The concept of reinforcement learning has a commonality with surgical procedures. A hysterectomy, which is one of the most basic gynecologic procedures [8], can be considered as an example. The aim of a hysterectomy is to enucleate the uterus; however, there are no specific guidelines regarding precisely how the knife should be used at the moment that the uterus is being removed. Rather, resecting the uterine arteries in advance, for instance, may reduce total blood loss, and this possibility should be evaluated. In addition, the resection of ligaments should also be evaluated, as this step precedes extraction of the uterus.

Altogether, we simulated and investigated whether reinforcement learning could be applied to surgery and whether AI could be possibly used to perform surgeries in the future.

In this study, we created simple 2D tasks that mimicked surgery (such as a hysterectomy simulation) and investigated whether the surgery could be performed as expected via reinforcement learning. During the task, the player (an imaginary knife) moved around the task areas. The player scored when passing target areas, such as imaginary ligaments and arteries, and lost points for other actions. The task was over when the player reached the uterus or moved outside the task area. We established rules and observed the process during which movement of the imaginary knife was learned and improved. In the task with a relatively small task area (Task 1), the knife finally passed all target areas, and the expected learning was achieved. In contrast, in the task with a large task area (Task 2), significantly more time was required to complete the task, suggesting that learning was not achieved. We addressed this problem by expanding the CPU memory and adding inhibitory task areas (Task 3), and some improvements were observed.

In this study, we applied the concept of reinforcement learning to surgical procedures and identified some commonalities between reinforcement learning and the way surgeons approach an operation. We found that various aspects of the techniques of efficient learning should be developed for applying reinforcement learning to surgery. These aspects include choosing a model that closely reproduces the surgical scene using a high-performance computer for deep learning and tuning of neural networks. Surgeons will need to understand the rules, such as when a reward can be earned and where the agent is allowed to move. Efficient learning would be possible by integrating these rules into the AI by engineers.

2. Materials and methods

2.1. 2D tasks

To create the 2D task, a publicly available game code for the game "Snake" was referenced and tuned (https://github.com/farizrahman4u/qlearning4k). The code for the operation task is provided in the **S7 Text**. The snake game was used for this investigation because it could be easily modified to the task we needed. Therefore, in particular, the original code should not necessarily be the snake game if one could create the task which mimics surgery.

2.2. Task 1

We created 9 × 9 squares of a simple 2D task representing a surgery (a hysterectomy) as shown in **Figure 1A**. The goal of the task is to remove the object (uterus, yellow); indirect goals, such as cutting ligaments and arteries, were also established. In the context of surgery, there are preferred cutting points that consider the densities of the arteries and the distance from the object, and these points were set as target points (green). Thus, the objective was for the tip of the knife (orange) to pass (cut) all the target points. In other words, if surgeons see the mass (tumors or uterus what so ever) to remove like yellow areas in **Figure 1A**, then they would cut the ligament (green areas) by using knife as drawing a circle, without approaching the area where the bleeding is expected to occur (red). Rewards, such as +1 for passing green areas, −1 for passing red areas, and 0 for passing blue areas (peritoneum), were defined. In addition, −1
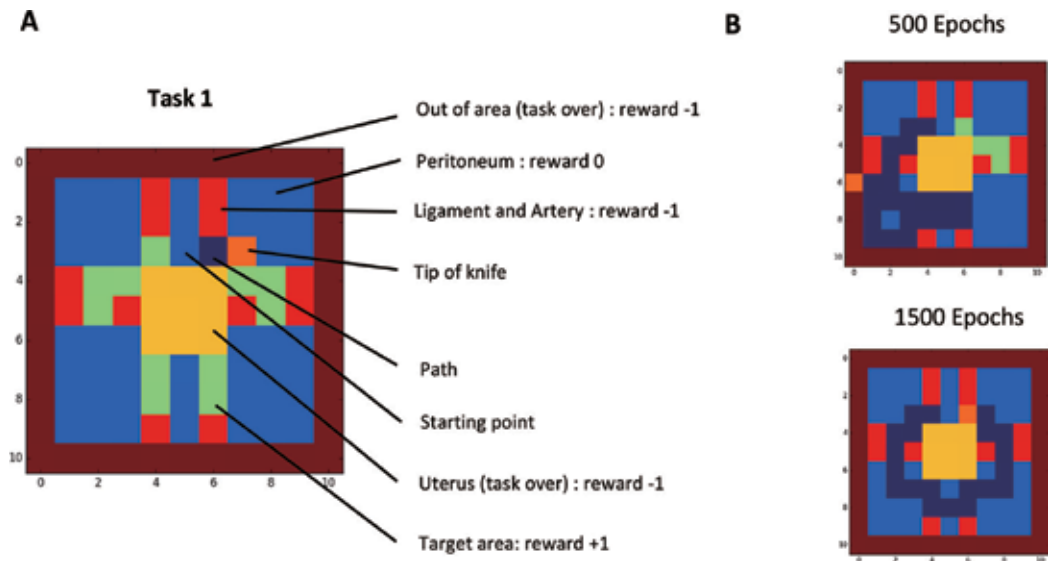
Figure 1. Creation of task. (A) Rules of the task. The name and reward of each area were as follows: orange, tip of knife; dark blue, path where tip of knife had passed; red, ligament or artery, −1; green, target area, +1; yellow, uterus, −1 (task over); blue, peritoneum, 0; brown, out of area, −1 (task over). Starting point was located in the upper middle of area. (B) Results of learning. The results after 500 epochs of learning (upper) and 1500 epochs of learning (lower) are shown. The knife passed all the green points in the shortest amount of time after learning.

was given and the task was defined as over if the knife passed over the yellow area (uterus) or if the knife passed outside the task area. Finally, −1 was given for passing areas were previously passed.

2.3. Task 2

We extended the task to 48 × 48 squares. This imposed significantly heavier burdens on the computer compared to Task 1, and we extended the CPU memory as described below.

2.4. Task 3

We added inhibitory areas (which would trigger the task to be over) to Task 2, and named it Task 3.

2.5. Deep Q network

We used a neural network library, Keras, for the reinforcement learning (https://keras.io) and TensorFlow as its backend (https://www.tensorflow.org) [9]. The codes for the agent can be found at https://github.com/farizrahman4u/qlearning4k.

The code simply obey the rules' Q(S, a) = r + gamma * Q(S', a')'.

Q(S, a) means the maximum score the agent will get by the end of the game, if it does action a, when the game is in state S. On performing action a, the game will jump to a new state S', giving the agent an immediate reward r.

In short, Q(S, a) = Immediate reward from this state + Max-score from the next state onwards.

Gamma is a discounting factor to give more weight to the immediate reward.

We are gynecologists and understand the very minimum of deep q learning, and we just tuned nb_frames (the number of frames should the agent remeber), batch_size (batch size for training), gamma (discounting factor) and nb_eopch (number of epochs to train). The original neural networks contained two convolutional layers and two dense layers. We slightly tuned the network by adding convolutional layers or changing the size of frames, but we did not see much improvement. Thus, the readers might find a more efficient way of learning. However, the objective of the present study is to propose the combination of reinforcement learning and surgery, and we did not investigate the tuning any further after obtaining conclusions.

2.6. Development environment

The development environments used in this study were Python 2.7.12, Keras 1.1.0, TensorFlow 0.8.0, and Matplotlib1.5.3. In Task 1, a Mac OS X with an Intel Core i5 processor and 8 GB memory (MacBook Pro, Apple, USA) was used. In Tasks 2 and 3, a Ubuntu 14. 04LTS with an Intel Xenon E5 v2 processor and 26 GB memory (Google Compute Engine, Google, USA) was used.

3. Results

3.1. Learning process and trials for appropriate learning

We created 9 × 9 squares of a simple 2D task representing the surgical scene (**Figure 1A**). We provide representative movies after 10 epochs, 500 epochs and 1500 epochs of learning (**Figure 1B** and **Movies 1–3**). The knife passed all the green points in the shortest amount of time after learning.

We next extended the task to 48 × 48 squares (**Movies 4** and **5**). This imposed significantly heavier burdens on the computer than in Task 1, and therefore, we expanded the CPU memory as described in the Section 2. The knife was located in limited areas after 500 epochs of learning (**Movie 4**), and the area of movement was extended after 1500 epochs of learning (**Movie 5**). Even after extending the CPU memory, the task required a long amount of time, and thus, learning was not necessarily achieved. Thus, we modified Task 2 to Task 3.

In Task 2, the knife moved to areas where no reward could be obtained (upper right area), which was hypothesized to prevent appropriate learning (**Movie 5**). We considered that some areas are not significant in an actual hysterectomy and added inhibitory areas (Task 3) that would trigger the task to end (dark blue areas in **Movie 6**). By adding these inhibitory areas, the knife reached the lower areas where it had never reached after 1500 epochs of learning (**Movie 6**). However, even with these limitations, a long time for learning was required within the simulated environments. These results suggest that efficient learning can be achieved by setting appropriate rules such as adding inhibitory areas. All video materials referenced in this section are available at: https://www.intechopen.com/download/index/process/160/authkey/f5524cabe6c92eff86119b2f0c58ba26

4. Discussion

In this study, we simulated and investigated whether reinforcement learning could be applied to surgery in the future and evaluated the types of hurdles that may exist.

Progress in AI has been remarkable, and AI is currently an essential part of the technology industry. Recent progress in AI can be attributed to the development of deep learning. Deep learning is a form of machine learning featured by the characteristic that the user does not have to choose features or representations while inputting data [10]. Deep learning is currently widely used in image recognition and audio recognition. While playing games with reinforcement learning has also been investigated, reinforcement learning has recently been combined with deep learning, resulting in drastic improvements in performance [4–6].

Thus, we sought to investigate whether reinforcement learning could be used in surgery in the future and developed appropriate simulations. Our team is composed of clinicians rather than engineers; however, we performed this study after studying deep learning and reinforcement learning.
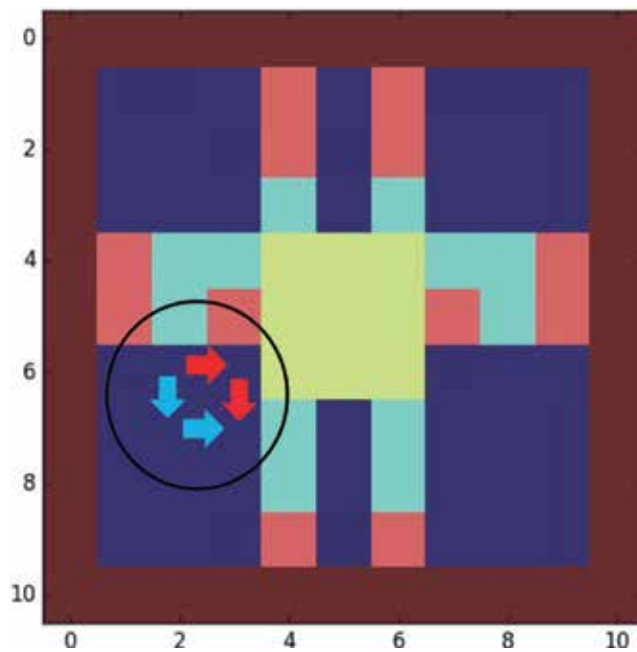
Figure 2. Choice of the shortest route. Blue and red arrows in the circle represent the shortest routes. Only the route of the blue arrows was chosen after several trials.

In reinforcement learning, it is preferable to obtain a reward during the immediately following action rather than having expectations to obtain rewards in distant future actions [6]. This can be understood by the results of Task 1, in which the knife passed all the green areas in the shortest possible time, even though a rule never inhibited a more round-about approach (**Figure 1B**). This can be expected to as a way to shorten the surgical time. Furthermore, choosing the shortest route was also of interest. The blue and red arrows in the circle in **Figure 2** represent the shortest routes; the route of the blue arrows was chosen after several trials (**Figure 1B**). The method used to decide the next action in reinforcement learning is maximizing the expected values that would be obtained after subsequent actions [6]. Therefore, actions that have greater expected values are likely to be preferred. Therefore, we expected that the red arrows were not preferred because points would be deducted (red area) or the task would be ended (uterus) with subsequent actions. This approach is also considered superior in surgery. In other words, avoiding areas associated with point deductions would result in retaining the appropriate margins during surgery. Therefore, the characteristics of reinforcement learning seemed to be compatible with typical surgical approaches.

The other results obtained were expected. It was concluded that reinforcement learning could solve a simple 2D task; 3D models that more closely replicate surgery should be further considered. In addition, increasing the playable area provides significantly more options for actions, and high-performance computers and tuning of neural networks will be needed for more complex tasks. 3D operation models are currently being developed and are used in

practice for endoscopic surgery simulation [11–14]. The speed at which high-performance computers are being developed is astonishing, and deep learning is being thoroughly investigated by engineers both in academia and business [15–17]. Combining the progress in both fields may provide designs allowing AI to realistically perform surgery. In such areas, clinicians can contribute the essential aspects of the surgery, that is, the playable task areas and the appropriate scores. Although increasing reward areas or limiting playable areas would speed learning time, such restrictions could also prevent the AI from finding alternative paths. Thus, clinicians and engineers should work cooperatively define rules.

In this study, we applied the concept of reinforcement learning to surgical procedures and identified common points between reinforcement learning and the way surgeons approach an operation. Reinforcement learning is now exclusively used and studied in the context of game play; however, it could also be applied to performing surgeries now that robotic surgery is widely available. Although there are many hurdles to overcome, AI could be applied to surgery by setting the appropriate rules, such as defining rewards and playable (operable) areas. To realize this goal, it is important for clinicians to further study deep learning and reinforcement learning strategies.

Acknowledgements

Conflict of interest

Authors declare there is no conflict of interests.

Supporting information

Movie 1. Result of Game 1 after 10 epochs of learning.

Movie 2. Result of Game 1 after 500 epochs of learning.

Movie 3. Result of Game 1 after 1500 epochs of learning.

Movie 4. Result of Game 2 after 500 epochs of learning.

Movie 5. Result of Game 2 after 1500 epochs of learning.

Movie 6. Result of Game 3 after 1500 epochs of learning.

https://www.intechopen.com/download/index/process/160/authkey/f5524cabe6c92eff86119b2f0c58ba26

S7 Text. The game codes for Game 1, 2 and 3 (python script). The code of agent.py, memory. py and game.py could be obtained in the link in Section 2.

```python
# game1_test.py

from keras.models import Sequential
from keras.layers import *
from game1 import OP
from keras.optimizers import *
from agent import Agent

grid_size = 11
nb_frames = 5
nb_actions = 5

model = Sequential()
model.add(Convolution2D(16, 2, 2, activation='relu', input_shape=(nb_frames,
grid_size, grid_size)))
model.add(Convolution2D(32, 3, 3, activation='relu'))
model.add(Flatten())
model.add(Dense(256, activation='relu'))
model.add(Dense(nb_actions))
model.compile(RMSprop(), 'MSE')

operation = OP(grid_size)

agent = Agent(model=model, memory_size=-1, nb_frames=nb_frames)
agent.train(operation, batch_size=64, nb_epoch=500, gamma=0.8)
agent.play(operation)



# game1.py

__original_author__ = "Fariz Rahman"
# https://github.com/farizrahman4u/qlearning4k

# Modified by "Masakazu Sato" on 10th December, 2016.
import numpy as np
from game import Game


actions = {0:'left', 1:'right', 2:'up', 3:'down', 4:'idle'}
forbidden_moves = [(0, 1), (1, 0), (2, 3), (3, 2)]

class OP(Game):

    def __init__(self, grid_size=11, knife_length=1):
        self.grid_size = grid_size
        self.knife_length = knife_length
        self.reset()
        self.state_changed = True
```

```python
    @property
    def name(self):
        return "OP"
    @property
    def nb_actions(self):
        return 5

    def play(self, action):
        assert action in range(5), "Invalid action."
        self.plus_scored = False
        self.minus_scored = False
        self.non_scored = False
        self.move_knife(action)
        tip = self.knife[0]

        if self.hit_artery():
            self.minus_scored = True
            self.art_border.remove(tip)
        elif self.hit_marker():
            self.plus_scored = True
            self.mark_border.remove(tip)
        elif self.hit_peritoneum():
            self.non_scored = True
            self.per_border.remove(tip)
        elif self.hit_border():
            self.game_over = True
        else:
            self.minus_scored = True




    def arrange(self):
        self.artery = (1,4),(1,6),(2,4),(2,6),(4,1),(4,9),(5,1),(5,3),(5,7),
(5,9),(9,4),(9,6)
        self.marker = (3,4),(3,6),(4,2),(4,3),(4,7),(4,8),(5,2),(5,8),(7,4),
(7,6),(8,4),(8,6)
        self.uterus = (4,4),(4,5),(4,6),(5,4),(5,5),(5,6),(6,4),(6,5),(6,6)
    self.peritoneum = (1,1),(1,2),(1,3),(1,5),(1,7),(1,8),(1,9),(2,1),(2,2),
(2,3),(2,5),(2,7),(2,8),(2,9),(3,1),(3,2),(3,3),(3,5),(3,7),(3,8),(3,9),
(6,1),(6,2),(6,3),(6,7),(6,8),(6,9),(7,1),(7,2),(7,3),(7,5),(7,7),(7,8),
(7,9),(8,1),(8,2),(8,3),(8,5),(8,7),(8,8),(8,9),(9,1),(9,2),(9,3),(9,5),
(9,7),(9,8),(9,9)

    def new_arrange(self):
        self.artery = (1,4),(1,6),(2,4),(2,6),(4,1),(4,9),(5,1),(5,3),(5,7),
(5,9),(9,4),(9,6)
        self.marker = (3,4),(3,6),(4,2),(4,3),(4,7),(4,8),(5,2),(5,8),(7,4),
(7,6),(8,4),(8,6)
        self.uterus = (4,4),(4,5),(4,6),(5,4),(5,5),(5,6),(6,4),(6,5),(6,6)
        self.peritoneum = (1,1),(1,2),(1,3),(1,5),(1,7),(1,8),(1,9),(2,1),(2,
2),(2,3),(2,5),(2,7),(2,8),(2,9),(3,1),(3,2),(3,3),(3,5),(3,7),(3,8),(3,9),
```

```
(6,1),(6,2),(6,3),(6,7),(6,8),(6,9),(7,1),(7,2),(7,3),(7,5),(7,7),(7,8),(7,
9),(8,1),(8,2),(8,3),(8,5),(8,7),(8,8),(8,9),(9,1),(9,2),(9,3),(9,5),(9,7),
(9,8),(9,9)

        self.art_border   []
        self.art_border +- self.artery
        self.mark_border = ''
        self.mark_border += self.marker
        self.per_border = []
        self.per_border += self.peritoneum

    def move_knife(self, action):
        if action == 4 or (action, self.previous_action) in forbidden_moves:
            action = self.previous_action
        else:
            self.previous_action = action
        head = self.knife[0]
        if action == 0:
            p = (head[0] - 1, head[1])
        elif action == 1:
            p = (head[0] + 1, head[1])
        elif action == 2:
            p = (head[0], head[1] - 1)
        elif action == 3:
            p = (head[0], head[1] + 1)
        self.knife.insert(0, p)
        self.knife.pop()


    def get_state(self):
        canvas = np.ones((self.grid_size, ) * 2)
        canvas[1:-1, 1:-1] = 0.0
        for seg in self.knife:
            canvas[seg[0], seg[1]] = 0.8
        for seg in self.art_border:
            canvas[seg[0],seg[1]] = 0.9
        for seg in self.mark_border:
            canvas[seg[0], seg[1]] = 0.5
        for seg in self.at_border:
            canvas[seg[0],seg[1]] = 0.7
        for seg in self.per_border:
            canvas[seg[0],seg[1]] = 0.2
        return canvas

    def get_score(self):
        if self.game_over:
            score = -3
        elif self.plus_scored:
            score = +3
        elif self.minus_scored:
            score = -2
        elif self.non_scored:
            score = +1
```

```
    else:
        score = -1
    return score

def reset(self):
    grid_size = self.grid_size
    knife_length = self.knife_length
    self.knife = [(3,5)]
    self.game_over = False
    self.plus_scored = False
    self.minus_scored = False
    self.non_scored = False
    self.arrange()
    if np.random.randint(2) == 0:
        self.previous_action = 0
    else:
        self.previous_action = 1
    self.border = []
    for z in range(grid_size):
      self.border += [(z, 0), (z, grid_size - 1), (0, z), (grid_size - 1, z)]
    self.ut_border = []
    self.ut_border += self.uterus
    self.art_border = []
    self.art_border += self.artery
    self.mark_border = []
    self.mark_border += self.marker
    self.per_border = []
    self.per_border += self.peritoneum

def left(self):
    self.play(0)
def right(self):
    self.play(1)
def up(self):
    self.play(2)
def down(self):
    self.play(3)
def idle(self):
    self.play(4)


def hit_border(self):
    return self.knife[0] in self.border or self.knife[0] in self.ut_border
def hit_artery(self):
    return self.knife[0] in self.art_border
def hit_marker(self):
    return self.knife[0] in self.mark_border
def hit_peritoneum(self):
    return self.knife[0] in self.per_border

def is_over(self):
    return self.hit_border()
def is_won(self):
    return len(self.mark_border) < 1
```

```
# game2_test.py

from keras.models import Sequential
from keras.layers import *
from game2 import OP
from keras.optimizers import *
from agent import Agent

grid_size = 50
nb_frames = 5
nb_actions = 5

model = Sequential()
model.add(Convolution2D(16, 2, 2, activation='relu', input_shape=(nb_frames,
grid_size, grid_size)))
model.add(Convolution2D(32, 3, 3, activation='relu'))
model.add(Flatten())
model.add(Dense(256, activation='relu'))
model.add(Dense(nb_actions))
model.compile(RMSprop(), 'MSE')

operation = OP(grid_size)

agent = Agent(model=model, memory_size=5, nb_frames=nb_frames)
agent.train(operation, batch_size=64, nb_epoch=500, gamma=0.80)
agent.play(operation)


# game2.py

__original_author__ = "Fariz Rahman"
# Modified by "Masakazu Sato" on 10th December,2016.
import numpy as np
from game import Game


actions = {0:'left', 1:'right', 2:'up', 3:'down', 4:'idle'}
forbidden_moves = [(0, 1), (1, 0), (2, 3), (3, 2)]

class OP(Game):

    def __init__(self, grid_size=50, knife_length=1):
        self.grid_size = grid_size
        self.knife_length = knife_length
        self.reset()
        self.state_changed = True

    @property
    def name(self):
        return "OP"
    @property
    def nb_actions(self):
        return 5
```

```python
def play(self, action):
    assert action in range(5), "Invalid action."
    self.plus_scored = False
    self.minus_scored = False
    self.non_scored = False
    self.move_knife(action)
    tip = self.knife[0]

    if self.hit_artery():
        self.minus_scored = True
        self.art_border.remove(tip)
    elif self.hit_marker():
        self.plus_scored = True
        self.mark_border.remove(tip)
    elif self.hit_peritoneum():
        self.non_scored = True
        self.per_border.remove(tip)
    elif self.hit_border():
        self.game_over = True
    else:
        self.minus_scored = True


def arrange(self):
    self.artery = (1,21),(1,22),(2,21),(2,22),(3,21),(3,22),(4,21),(4,22),
    (5,21),(5,22),(6,21),(6,22),(7,21),(7,22),(8,21),(8,22),(1,27),(1,28),(2,
    27),(2,28),(3,27),(3,28),(4,27),(4,28),(5,27),(5,28),(6,27),(6,28),(7,27),
    (7,28),(8,27),(8,28),(11,21),(11,22),(11,23),(12,21),(12,22),(12,23),(13,21),
    (13,22),(13,23),(14,21),(14,22),(14,23),(15,21),(15,22),(15,23),(16,21),(16,
    22),(16,23),(11,26),(11,27),(11,28),(12,26),(12,27),(12,28),(13,26),(13,27),
    (13,28),(14,26),(14,27),(14,28),(15,26),(15,27),(15,28),(16,26),(16,27),(16,
    28),(21,1),(21,2),(22,1),(22,2),(23,1),(23,2),(24,1),(24,2),(25,1),(25,2),
    (26,1),(26,2),(27,1),(27,2),(28,1),(28,2),(23,3),(23,4),(23,5),(24,3),(24,4),
    (24,5),(25,3),(25,4),(25,5),(26,3),(26,4),(26,5),(24,6),(24,7),(24,8),(24,9),
    (25,6),(25,7),(25,8),(25,9),(23,12),(23,13),(23,14),(24,12),(24,13),(24,14),
    (25,12),(25,13),(25,14),(26,12),(26,13),(26,14),(21,15),(21,16),(22,15),(22,
    16),(23,15),(23,16),(24,15),(24,16),(25,15),(25,16),(26,15),(26,16),(27,15),
    (27,16),(28,15),(28,16),(21,33),(21,34),(22,33),(22,34),(23,33),(23,34),(24,
    33),(24,34),(25,33),(25,34),(26,33),(26,34),(27,33),(27,34),(28,33),(28,34),
    (23,35),(23,36),(23,37),(24,35),(24,36),(24,37),(25,35),(25,36),(25,37),(26,
    35),(26,36),(26,37),(24,40),(24,41),(24,42),(24,43),(25,40),(25,41),(25,42),
    (25,43),(23,44),(23,45),(23,46),(24,44),(24,45),(24,46),(25,44),(25,45),(25,
    46),(26,44),(26,45),(26,46),(21,47),(21,48),(22,47),(22,48),(23,47),(23,48),
    (24,47),(24,48),(25,47),(25,48),(26,47),(26,48),(27,47),(27,48),(28,47),(28,
    48),(33,21),(33,22),(33,23),(34,21),(34,22),(34,23),(35,21),(35,22),(35,23),
    (40,21),(40,22),(40,23),(41,21),(41,22),(41,23),(42,21),(42,22),(42,23),(43,
    21),(43,22),(43,23),(44,21),(44,22),(44,23),(45,21),(45,22),(45,23),(46,21),
    (46,22),(46,23),(47,21),(47,22),(47,23),(48,21),(48,22),(48,23),(33,26),(33,
    27),(33,28),(34,26),(34,27),(34,28),(35,26),(35,27),(35,28),(40,26),(40,27),
    (40,28),(41,26),(41,27),(41,28),(42,26),(42,27),(42,28),(43,26),(43,27),(43,
    28),(44,26),(44,27),(44,28),(45,26),(45,27),(45,28),(46,26),(46,27),(46,28),
    (47,26),(47,27),(47,28),(48,26),(48,27),(48,28)
```

```
        self.marker = (9,21),(9,22),(10,21),(10,22),(9,27),(9,28),(10,27),(10,
28),(24,10),(24,11),(25,10),(25,11),(24,38),(24,39),(25,38),(25,39),(36,21),
(36,22),(36,23),(37,21),(37,22),(37,23),(38,21),(38,22),(38,23),(39,21),(39,
22),(39,23),(36,26),(36,27),(36,28),(37,26),(37,27),(37,28),(38,26),(38,27),
(38,28),(39,26),(39,27),(39,28)
        self.uterus = (21,17),(22,17),(23,17),(24,17),(25,17),(26,17),(27,17),
(28,17),(20,18),(21,18),(22,18),(23,18),(24,18),(25,18),(26,18),(27,18),(28,
18),(29,18),(19,19),(20,19),(21,19),(22,19),(23,19),(24,19),(25,19),(26,19),
(27,19),(28,19),(29,19),(30,19),(18,20),(19,20),(20,20),(21,20),(22,20),(23,
20),(24,20),(25,20),(26,20),(27,20),(28,20),(29,20),(30,20),(31,20),(17,21),
(17,22),(17,23),(17,24),(17,25),(17,26),(17,27),(17,28),(18,21),(18,22),(18,
23),(18,24),(18,25),(18,26),(18,27),(18,28),(19,21),(19,22),(19,23),(19,24),
(19,25),(19,26),(19,27),(19,28),(20,21),(20,22),(20,23),(20,24),(20,25),(20,
26),(20,27),(20,28),(21,21),(21,22),(21,23),(21,24),(21,25),(21,26),(21,27),
(21,28),(22,21),(22,22),(22,23),(22,24),(22,25),(22,26),(22,27),(22,28),(23,
21),(23,22),(23,23),(23,24),(23,25),(23,26),(23,27),(23,28),(24,21),(24,22),
(24,23),(24,24),(24,25),(24,26),(24,27),(24,28),(25,21),(25,22),(25,23),(25,
24),(25,25),(25,26),(25,27),(25,28),(26,21),(26,22),(26,23),(26,24),(26,25),
(26,26),(26,27),(26,28),(27,21),(27,22),(27,23),(27,24),(27,25),(27,26),(27,
27),(27,28),(28,21),(28,22),(28,23),(28,24),(28,25),(28,26),(28,27),(28,28),
(29,21),(29,22),(29,23),(29,24),(29,25),(29,26),(29,27),(29,28),(30,21),(30,
22),(30,23),(30,24),(30,25),(30,26),(30,27),(30,28),(31,21),(31,22),(31,23),
(31,24),(31,25),(31,26),(31,27),(31,28),(32,21),(32,22),(32,23),(32,24),(32,
25),(32,26),(32,27),(32,28),(18,29),(19,29),(20,29),(21,29),(22,29),(23,29),
(24,29),(25,29),(26,29),(27,29),(28,29),(29,29),(30,29),(31,29),(19,30),(20,
30),(21,30),(22,30),(23,30),(24,30),(25,30),(26,30),(27,30),(28,30),(29,30),
(30,30),(20,31),(21,31),(22,31),(23,31),(24,31),(25,31),(26,31),(27,31),(28,
31),(29,31),(21,32),(22,32),(23,32),(24,32),(25,32),(26,32),(27,32),(28,32)
        self.peritoneum = (22,3),(22,4),(22,5),(27,3),(27,4),(27,5),(22,35),
(22,36),(22,37),(27,35),(27,36),(27,37),(22,44),(22,45),(22,46),(27,44),(27,
45),(27,46),(6,17),(6,18),(6,19),(6,20),(7,17),(7,18),(7,19),(7,20),(8,17),
(8,18),(8,19),(8,20),(9,17),(9,18),(9,19),(9,20),(10,17),(10,18),(10,19),(10,
20),(11,17),(11,18),(11,19),(11,20),(12,17),(12,18),(12,19),(12,20),(13,17),
(13,18),(13,19),(13,20),(14,17),(14,18),(14,19),(14,20),(15,17),(15,18),(15,
19),(15,20),(16,17),(16,18),(16,19),(16,20),(17,7),(17,8),(17,9),(17,10),(17,
11),(17,12),(17,13),(17,14),(17,15),(17,16),(18,7),(18,8),(18,9),(18,10),
(18,11),(18,12),(18,13),(18,14),(18,15),(18,16),(19,7),(19,8),(19,9),(19,10),
(19,11),(19,12),(19,13),(19,14),(19,15),(19,16),(20,7),(20,8),(20,9),(20,10),
(20,11),(20,12),(20,13),(20,14),(20,15),(20,16),(17,17),(18,17),(19,17),(20,
17),(17,18),(18,18),(19,18),(17,19),(18,19),(17,20),(21,3),(21,4),(21,5),
(21,6),(21,7),(21,8),(21,9),(21,10),(21,11),(21,12),(21,13),(21,14),(22,6),
(22,7),(22,8),(22,9),(22,10),(22,11),(22,12),(22,13),(22,14),(23,6),(23,7),
(23,8),(23,9),(23,10),(23,11),(26,6),(26,7),(26,8),(26,9),(26,10),(26,11),
(27,6),(27,7),(27,8),(27,9),(27,10),(27,11),(27,12),(27,13),(27,14),(28,3),
(28,4),(28,5),(28,6),(28,7),(28,8),(28,9),(28,10),(28,11),(28,12),(28,13),(28,
14),(29,7),(29,8),(29,9),(29,10),(29,11),(29,12),(29,13),(29,14),(29,15),(29,
16),(30,7),(30,8),(30,9),(30,10),(30,11),(30,12),(30,13),(30,14),(30,15),(30,
16),(31,7),(31,8),(31,9),(31,10),(31,11),(31,12),(31,13),(31,14),(31,15),(31,
16),(32,7),(32,8),(32,9),(32,10),(32,11),(32,12),(32,13),(32,14),(32,15),(32,
```

16),(33,17),(33,18),(33,19),(33,20),(34,17),(34,18),(34,19),(34,20),(35,17),
(35,18),(35,19),(35,20),(36,17),(36,18),(36,19),(36,20),(37,17),(37,18),(37,
19),(37,20),(38,17),(38,18),(38,19),(38,20),(39,17),(39,18),(39,19),(39,20),
(40,17),(40,18),(40,19),(40,20),(29,17),(30,17),(31,17),(32,17),(30,18),(31,
18),(32,18),(31,19),(32,19),(32,20),(6,23),(6,24),(6,25),(6,26),(7,23),(7,24),
(7,25),(7,26),(8,23),(8,24),(8,25),(8,26),(9,23),(3,24),(9,25),(9,26),(10,23),
(10,24),(10,25),(10,26),(11,24),(11,25),(12,24),(12,25),(13,24),(13,25),(14,
24),(14,25),(15,24),(15,25),(16,24),(16,25),(33,24),(33,25),(34,24),(34,25),
(35,24),(35,25),(36,24),(36,25),(37,24),(37,25),(38,24),(38,25),(39,24),(39,
25),(40,24),(40,25),(6,29),(6,30),(6,31),(6,32),(7,29),(7,30),(7,31),(7,32),
(8,29),(8,30),(8,31),(8,32),(9,29),(9,30),(9,31),(9,32),(10,29),(10,30),(10,
31),(10,32),(11,29),(11,30),(11,31),(11,32),(12,29),(12,30),(12,31),(12,32),
(13,29),(13,30),(13,31),(13,32),(14,29),(14,30),(14,31),(14,32),(15,29),(15,
30),(15,31),(15,32),(16,29),(16,30),(16,31),(16,32),(33,29),(33,30),(33,31),
(33,32),(34,29),(34,30),(34,31),(34,32),(35,29),(35,30),(35,31),(35,32),(36,
29),(36,30),(36,31),(36,32),(37,29),(37,30),(37,31),(37,32),(38,29),(38,30),
(38,31),(38,32),(39,29),(39,30),(39,31),(39,32),(40,29),(40,30),(40,31),(40,
32),(17,33),(17,34),(17,35),(17,36),(17,37),(17,38),(17,39),(17,40),(17,41),
(17,42),(18,33),(18,34),(18,35),(18,36),(18,37),(18,38),(18,39),(18,40),(18,
41),(18,42),(19,33),(19,34),(19,35),(19,36),(19,37),(19,38),(19,39),(19,40),
(19,41),(19,42),(20,33),(20,34),(20,35),(20,36),(20,37),(20,38),(20,39),(20,
40),(20,41),(20,42),(29,33),(29,34),(29,35),(29,36),(29,37),(29,38),(29,39),
(29,40),(29,41),(29,42),(30,33),(30,34),(30,35),(30,36),(30,37),(30,38),(30,
39),(30,40),(30,41),(30,42),(31,33),(31,34),(31,35),(31,36),(31,37),(31,38),
(31,39),(31,40),(31,41),(31,42),(32,33),(32,34),(32,35),(32,36),(32,37),(32,
38),(32,39),(32,40),(32,41),(32,42),(21,35),(21,36),(21,37),(21,38),(21,39),
(21,40),(21,41),(21,42),(21,43),(21,44),(21,45),(21,46),(22,38),(22,39),(22,
40),(22,41),(22,42),(22,43),(23,38),(23,39),(23,40),(23,41),(23,42),(23,43),
(26,38),(26,39),(26,40),(26,41),(26,42),(26,43),(27,38),(27,39),(27,40),(27,
41),(27,42),(27,43),(28,35),(28,36),(28,37),(28,38),(28,39),(28,40),(28,41),
(28,42),(28,43),(28,44),(28,45),(28,46),(17,32),(18,32),(19,32),(20,32),(17,
31),(18,31),(19,31),(17,30),(18,30),(17,29),(29,32),(30,32),(31,32),(32,32),
(30,31),(31,31),(32,31),(31,30),(32,30),(32,29),(1,1),(1,2),(1,3),(1,4),(1,5),
(1,6),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),
(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),(6,1),
(6,2),(6,3),(6,4),(6,5),(6,6),(7,1),(7,2),(7,3),(7,4),(7,5),(7,6),(8,1),(8,2),
(8,3),(8,4),(8,5),(8,6),(9,1),(9,2),(9,3),(9,4),(9,5),(9,6),(10,1),(10,2),(10,
3),(10,4),(10,5),(10,6),(11,1),(11,2),(11,3),(11,4),(11,5),(11,6),(12,1),(12,
2),(12,3),(12,4),(12,5),(12,6),(13,1),(13,2),(13,3),(13,4),(13,5),(13,6),(14,
1),(14,2),(14,3),(14,4),(14,5),(14,6),(15,1),(15,2),(15,3),(15,4),(15,5),(15,
6),(16,1),(16,2),(16,3),(16,4),(16,5),(16,6),(17,1),(17,2),(17,3),(17,4),(17,
5),(17,6),(18,1),(18,2),(18,3),(18,4),(18,5),(18,6),(19,1),(19,2),(19,3),(19,
4),(19,5),(19,6),(20,1),(20,2),(20,3),(20,4),(20,5),(20,6),(1,7),(1,8),(1,9),
(1,10),(1,11),(1,12),(1,13),(1,14),(1,15),(1,16),(2,7),(2,8),(2,9),(2,10),(2,
11),(2,12),(2,13),(2,14),(2,15),(2,16),(3,7),(3,8),(3,9),(3,10),(3,11),(3,12),
(3,13),(3,14),(3,15),(3,16),(4,7),(4,8),(4,9),(4,10),(4,11),(4,12),(4,13),(4,
14),(4,15),(4,16),(5,7),(5,8),(5,9),(5,10),(5,11),(5,12),(5,13),(5,14),(5,15),
(5,16),(6,7),(6,8),(6,9),(6,10),(6,11),(6,12),(6,13),(6,14),(6,15),(6,16),(7,
7),(7,8),(7,9),(7,10),(7,11),(7,12),(7,13),(7,14),(7,15),(7,16),(8,7),(8,8),

(8,9),(8,10),(8,11),(8,12),(8,13),(8,14),(8,15),(8,16),(9,7),(9,8),(9,9),(9,
10),(9,11),(9,12),(9,13),(9,14),(9,15),(9,16),(10,7),(10,8),(10,9),(10,10),
(10,11),(10,12),(10,13),(10,14),(10,15),(10,16),(11,7),(11,8),(11,9),(11,10),
(11,11),(11,12),(11,13),(11,14),(11,15),(11,16),(12,7),(12,8),(12,9),(12,10),
(12,11),(12,12),(12,13),(12,14),(12,15),(12,16),(13,7),(13,8),(13,9),(13,10),
(13,11),(13,12),(13,13),(13,14),(13,15),(13,16),(14,7),(14,8),(14,9),(14,10),
(14,11),(14,12),(14,13),(14,14),(14,15),(14,16),(15,7),(15,8),(15,9),(15,10),
(15,11),(15,12),(15,13),(15,14),(15,15),(15,16),(16,7),(16,8),(16,9),(16,10),
(16,11),(16,12),(16,13),(16,14),(16,15),(16,16),(1,17),(1,18),(1,19),(1,20),
(2,17),(2,18),(2,19),(2,20),(3,17),(3,18),(3,19),(3,20),(4,17),(4,18),(4,19),
(4,20),(5,17),(5,18),(5,19),(5,20),(1,23),(1,24),(1,25),(1,26),(2,23),(2,24),
(2,25),(2,26),(3,23),(3,24),(3,25),(3,26),(4,23),(4,24),(4,25),(4,26),(5,23),
(5,24),(5,25),(5,26),(1,29),(1,30),(1,31),(1,32),(2,29),(2,30),(2,31),(2,32),
(3,29),(3,30),(3,31),(3,32),(4,29),(4,30),(4,31),(4,32),(5,29),(5,30),(5,31),
(5,32),(1,33),(1,34),(1,35),(1,36),(1,37),(1,38),(1,39),(1,40),(1,41),(1,42),
(1,43),(1,44),(1,45),(1,46),(1,47),(1,48),(2,33),(2,34),(2,35),(2,36),(2,37),
(2,38),(2,39),(2,40),(2,41),(2,42),(2,43),(2,44),(2,45),(2,46),(2,47),(2,48),
(3,33),(3,34),(3,35),(3,36),(3,37),(3,38),(3,39),(3,40),(3,41),(3,42),(3,43),
(3,44),(3,45),(3,46),(3,47),(3,48),(4,33),(4,34),(4,35),(4,36),(4,37),(4,38),
(4,39),(4,40),(4,41),(4,42),(4,43),(4,44),(4,45),(4,46),(4,47),(4,48),(5,33),
(5,34),(5,35),(5,36),(5,37),(5,38),(5,39),(5,40),(5,41),(5,42),(5,43),(5,44),
(5,45),(5,46),(5,47),(5,48),(6,33),(6,34),(6,35),(6,36),(6,37),(6,38),(6,39),
(6,40),(6,41),(6,42),(6,43),(6,44),(6,45),(6,46),(6,47),(6,48),(7,33),(7,34),
(7,35),(7,36),(7,37),(7,38),(7,39),(7,40),(7,41),(7,42),(7,43),(7,44),(7,45),
(7,46),(7,47),(7,48),(8,33),(8,34),(8,35),(8,36),(8,37),(8,38),(8,39),(8,40),
(8,41),(8,42),(8,43),(8,44),(8,45),(8,46),(8,47),(8,48),(9,33),(9,34),(9,35),
(9,36),(9,37),(9,38),(9,39),(9,40),(9,41),(9,42),(9,43),(9,44),(9,45),(9,46),
(9,47),(9,48),(10,33),(10,34),(10,35),(10,36),(10,37),(10,38),(10,39),(10,40)
,(10,41),(10,42),(10,43),(10,44),(10,45),(10,46),(10,47),(10,48),(11,33),(11,
34),(11,35),(11,36),(11,37),(11,38),(11,39),(11,40),(11,41),(11,42),(11,43),
(11,44),(11,45),(11,46),(11,47),(11,48),(12,33),(12,34),(12,35),(12,36),(12,
37),(12,38),(12,39),(12,40),(12,41),(12,42),(12,43),(12,44),(12,45),(12,46),
(12,47),(12,48),(13,33),(13,34),(13,35),(13,36),(13,37),(13,38),(13,39),(13,
40),(13,41),(13,42),(13,43),(13,44),(13,45),(13,46),(13,47),(13,48),(14,33),
(14,34),(14,35),(14,36),(14,37),(14,38),(14,39),(14,40),(14,41),(14,42),(14,
43),(14,44),(14,45),(14,46),(14,47),(14,48),(15,33),(15,34),(15,35),(15,36),
(15,37),(15,38),(15,39),(15,40),(15,41),(15,42),(15,43),(15,44),(15,45),(15,
46),(15,47),(15,48),(16,33),(16,34),(16,35),(16,36),(16,37),(16,38),(16,39),
(16,40),(16,41),(16,42),(16,43),(16,44),(16,45),(16,46),(16,47),(16,48),(17,
43),(17,44),(17,45),(17,46),(17,47),(17,48),(18,43),(18,44),(18,45),(18,46),
(18,47),(18,48),(19,43),(19,44),(19,45),(19,46),(19,47),(19,48),(20,43),(20,
44),(20,45),(20,46),(20,47),(20,48),(29,1),(29,2),(29,3),(29,4),(29,5),(29,6),
(30,1),(30,2),(30,3),(30,4),(30,5),(30,6),(31,1),(31,2),(31,3),(31,4),(31,5),
(31,6),(32,1),(32,2),(32,3),(32,4),(32,5),(32,6),(33,1),(33,2),(33,3),(33,4),
(33,5),(33,6),(34,1),(34,2),(34,3),(34,4),(34,5),(34,6),(35,1),(35,2),(35,3),
(35,4),(35,5),(35,6),(36,1),(36,2),(36,3),(36,4),(36,5),(36,6),(37,1),(37,2),
(37,3),(37,4),(37,5),(37,6),(38,1),(38,2),(38,3),(38,4),(38,5),(38,6),(39,1),
(39,2),(39,3),(39,4),(39,5),(39,6),(40,1),(40,2),(40,3),(40,4),(40,5),(40,6),
(41,1),(41,2),(41,3),(41,4),(41,5),(41,6),(42,1),(42,2),(42,3),(42,4),(42,5),

(42,6),(43,1),(43,2),(43,3),(43,4),(43,5),(43,6),(44,1),(44,2),(44,3),(44,4),
(44,5),(44,6),(45,1),(45,2),(45,3),(45,4),(45,5),(45,6),(46,1),(46,2),(46,3),
(46,4),(46,5),(46,6),(47,1),(47,2),(47,3),(47,4),(47,5),(47,6),(48,1),(48,2),
(48,3),(48,4),(48,5),(48,6),(33,7),(33,8),(33,9),(33,10),(33,11),(33,12),(33,
13),(33,14),(33,15),(33,16),(34,7),(34,8),(34,9),(34,10),(34,11),(34,12),(34,
13),(34,14),(34,15),(34,16),(35,7),(35,8),(35,9),(35,10),(35,11),(35,12),(35,
13),(35,14),(35,15),(35,16),(36,7),(36,8),(36,9),(36,10),(36,11),(36,12),(36,
13),(36,14),(36,15),(36,16),(37,7),(37,8),(37,9),(37,10),(37,11),(37,12),(37,
13),(37,14),(37,15),(37,16),(38,7),(38,8),(38,9),(38,10),(38,11),(38,12),(38,
13),(38,14),(38,15),(38,16),(39,7),(39,8),(39,9),(39,10),(39,11),(39,12),(39,
13),(39,14),(39,15),(39,16),(40,7),(40,8),(40,9),(40,10),(40,11),(40,12),(40,
13),(40,14),(40,15),(40,16),(41,7),(41,8),(41,9),(41,10),(41,11),(41,12),(41,
13),(41,14),(41,15),(41,16),(42,7),(42,8),(42,9),(42,10),(42,11),(42,12),(42,
13),(42,14),(42,15),(42,16),(43,7),(43,8),(43,9),(43,10),(43,11),(43,12),(43,
13),(43,14),(43,15),(43,16),(44,7),(44,8),(44,9),(44,10),(44,11),(44,12),(44,
13),(44,14),(44,15),(44,16),(45,7),(45,8),(45,9),(45,10),(45,11),(45,12),(45,
13),(45,14),(45,15),(45,16),(46,7),(46,8),(46,9),(46,10),(46,11),(46,12),(46,
13),(46,14),(46,15),(46,16),(47,7),(47,8),(47,9),(47,10),(47,11),(47,12),(47,
13),(47,14),(47,15),(47,16),(48,7),(48,8),(48,9),(48,10),(48,11),(48,12),(48,
13),(48,14),(48,15),(48,16),(41,17),(41,18),(41,19),(41,20),(42,17),(42,18),
(42,19),(42,20),(43,17),(43,18),(43,19),(43,20),(44,17),(44,18),(44,19),(44,
20),(45,17),(45,18),(45,19),(45,20),(46,17),(46,18),(46,19),(46,20),(47,17),
(47,18),(47,19),(47,20),(48,17),(48,18),(48,19),(48,20),(41,24),(41,25),(42,
24),(42,25),(43,24),(43,25),(44,24),(44,25),(45,24),(45,25),(46,24),(46,25),
(47,24),(47,25),(48,24),(48,25),(41,29),(41,30),(41,31),(41,32),(42,29),(42,
30),(42,31),(42,32),(43,29),(43,30),(43,31),(43,32),(44,29),(44,30),(44,31),
(44,32),(45,29),(45,30),(45,31),(45,32),(46,29),(46,30),(46,31),(46,32),(47,
29),(47,30),(47,31),(47,32),(48,29),(48,30),(48,31),(48,32),(33,33),(33,34),
(33,35),(33,36),(33,37),(33,38),(33,39),(33,40),(33,41),(33,42),(34,33),(34,
34),(34,35),(34,36),(34,37),(34,38),(34,39),(34,40),(34,41),(34,42),(35,33),
(35,34),(35,35),(35,36),(35,37),(35,38),(35,39),(35,40),(35,41),(35,42),(36,
33),(36,34),(36,35),(36,36),(36,37),(36,38),(36,39),(36,40),(36,41),(36,42),
(37,33),(37,34),(37,35),(37,36),(37,37),(37,38),(37,39),(37,40),(37,41),(37,
42),(38,33),(38,34),(38,35),(38,36),(38,37),(38,38),(38,39),(38,40),(38,41),
(38,42),(39,33),(39,34),(39,35),(39,36),(39,37),(39,38),(39,39),(39,40),(39,
41),(39,42),(40,33),(40,34),(40,35),(40,36),(40,37),(40,38),(40,39),(40,40),
(40,41),(40,42),(41,33),(41,34),(41,35),(41,36),(41,37),(41,38),(41,39),(41,
40),(41,41),(41,42),(42,33),(42,34),(42,35),(42,36),(42,37),(42,38),(42,39),
(42,40),(42,41),(42,42),(43,33),(43,34),(43,35),(43,36),(43,37),(43,38),(43,
39),(43,40),(43,41),(43,42),(44,33),(44,34),(44,35),(44,36),(44,37),(44,38),
(44,39),(44,40),(44,41),(44,42),(45,33),(45,34),(45,35),(45,36),(45,37),(45,
38),(45,39),(45,40),(45,41),(45,42),(46,33),(46,34),(46,35),(46,36),(46,37),
(46,38),(46,39),(46,40),(46,41),(46,42),(47,33),(47,34),(47,35),(47,36),(47,
37),(47,38),(47,39),(47,40),(47,41),(47,42),(48,33),(48,34),(48,35),(48,36),
(48,37),(48,38),(48,39),(48,40),(48,41),(48,42),(29,43),(29,44),(29,45),(29,
46),(29,47),(29,48),(30,43),(30,44),(30,45),(30,46),(30,47),(30,48),(31,43),
(31,44),(31,45),(31,46),(31,47),(31,48),(32,43),(32,44),(32,45),(32,46),(32,
47),(32,48),(33,43),(33,44),(33,45),(33,46),(33,47),(33,48),(34,43),(34,44),
(34,45),(34,46),(34,47),(34,48),(35,43),(35,44),(35,45),(35,46),(35,47),(35,

```
48),(36,43),(36,44),(36,45),(36,46),(36,47),(36,48),(37,43),(37,44),(37,45),
(37,46),(37,47),(37,48),(38,43),(38,44),(38,45),(38,46),(38,47),(38,48),(39,
43),(39,44),(39,45),(39,46),(39,47),(39,48),(40,43),(40,44),(40,45),(40,46),
(40,47),(40,48),(41,43),(41,44),(41,45),(41,46),(41,47),(41,48),(42,43),(42,
44),(42,45),(42,46),(42,47),(42,48),(43,43),(43,44),(43,45),(43,46),(43,47),
(43,48),(44,43),(44,44),(44,45),(44,46),(44,47),(44,48),(45,43),(45,44),(45,
45),(45,46),(45,47),(45,48),(46,43),(46,44),(46,45),(46,46),(46,47),(46,48),
(47,43),(47,44),(47,45),(47,46),(47,47),(47,48),(48,43),(48,44),(48,45),(48,
46),(48,47),(48,48)


    def move_knife(self, action):
        if action == 4 or (action, self.previous_action) in forbidden moves:
            action = self.previous_action
        else:
            self.previous_action = action
        head = self.knife[0]
        if action == 0:
            p = (head[0] - 1, head[1])
        elif action == 1:
            p = (head[0] + 1, head[1])
        elif action == 2:
            p = (head[0], head[1] - 1)
        elif action == 3:
            p = (head[0], head[1] + 1)
        self.knife.insert(0, p)
        self.knife.pop()

    def get_state(self):
        canvas = np.ones((self.grid_size, ) * 2)
        canvas[1:-1, 1:-1] = 0.0
        for seg in self.knife:
            canvas[seg[0], seg[1]] = 0.8
        for seg in self.art_border:
            canvas[seg[0],seg[1]] = 0.9
        for seg in self.mark_border:
            canvas[seg[0], seg[1]] = 0.5
        for seg in self.ut_border:
            canvas[seg[0],seg[1]] = 0.7
        for seg in self.per_border:
            canvas[seg[0],seg[1]] = 0.1
        return canvas

    def get_score(self):
        if self.game_over:
            score = -1
        elif self.plus_scored:
            score = +1
        elif self.minus_scored:
            score = -1
```

```python
        elif self.non_scored:
            score = 0
        else:
            score = -1
        return score

    def reset(self):
        grid_size = self.grid_size
        knife_length = self.knife_length
        self.knife = [(10,24)]
        self.game_over = False
        self.plus_scored = False
        self.minus_scored = False
        self.non_scored = False
        self.arrange()
        if np.random.randint(2) == 0:
            self.previous_action = 0
        else:
            self.previous_action = 1
        self.border = []
        for z in range(grid_size):
          self.border += [(z, 0), (z, grid_size - 1), (0, z), (grid_size - 1, z)]
        self.ut_border = []
        self.ut_border += self.uterus
        self.art_border = []
        self.art_border += self.artery
        self.mark_border = []
        self.mark_border += self.marker
        self.per_border = []
        self.per_border += self.peritoneum

    def left(self):
        self.play(0)
    def right(self):
        self.play(1)
    def up(self):
        self.play(2)
    def down(self):
        self.play(3)
    def idle(self):
        self.play(4)

    def hit_border(self):
        return self.knife[0] in self.border or self.knife[0] in self.ut_border
    def hit_artery(self):
        return self.knife[0] in self.art_border
    def hit_marker(self):
        return self.knife[0] in self.mark_border
```

```
    def hit_peritoneum(self):
        return self.knife[0] in self.per_border

    def is_over(self):
        return self.hit_border()
    def is_won(self):
        return len(self.mark_border) < 1


# game3_test.py

from keras.models import Sequential
from keras.layers import *
from game3 import OP
from keras.optimizers import *
from agent import Agent

grid_size = 50
nb_frames = 5
nb_actions = 5

model = Sequential()
model.add(Convolution2D(16, 2, 2, activation='relu', input_shape=(nb_frames,
grid_size, grid_size)))
model.add(Convolution2D(32, 3, 3, activation='relu'))
model.add(Flatten())
model.add(Dense(256, activation='relu'))
model.add(Dense(nb_actions))
model.compile(RMSprop(), 'MSE')

operation = OP(grid_size)

agent = Agent(model=model, memory_size=5, nb_frames=nb_frames)
agent.train(operation, batch_size=64, nb_epoch=500, gamma=0.80)
agent.play(operation)


# game3.py

__original_author__ = "Fariz Rahman"
# Modified by "Masakazu Sato" on 10th December,2016.
import numpy as np
from game import Game


actions = {0:'left', 1:'right', 2:'up', 3:'down', 4:'idle'}
forbidden_moves = [(0, 1), (1, 0), (2, 3), (3, 2)]

class OP(Game):

    def __init__(self, grid_size=50, knife_length=1):
        self.grid_size = grid_size
```

```python
        self.knife_length = knife_length
        self.reset()
        self.state_changed = True

    @property
    def name(self):
        return "OP"
    @property
    def nb_actions(self):
        return 5

    def play(self, action):
        assert action in range(5), "Invalid action."
        self.plus_scored = False
        self.minus_scored = False
        self.non_scored = False
        self.move_knife(action)
        tip = self.knife[0]

        if self.hit_artery():
            self.minus_scored = True
            self.art_border.remove(tip)
        elif self.hit_marker():
            self.plus_scored = True
            self.mark_border.remove(tip)
        elif self.hit_peritoneum():
            self.non_scored = True
            self.per_border.remove(tip)
        elif self.hit_border():
            self.game_over = True
        else:
            self.minus_scored = True


    def arrange(self):
        self.artery = (1,21),(1,22),(2,21),(2,22),(3,21),(3,22),(4,21),(4,22),
(5,21),(5,22),(6,21),(6,22),(7,21),(7,22),(8,21),(8,22),(1,27),(1,28),(2,
27),(2,28),(3,27),(3,28),(4,27),(4,28),(5,27),(5,28),(6,27),(6,28),(7,27),
(7,28),(8,27),(8,28),(11,21),(11,22),(11,23),(12,21),(12,22),(12,23),(13,
21),(13,22),(13,23),(14,21),(14,22),(14,23),(15,21),(15,22),(15,23),(16,21),
(16,22),(16,23),(11,26),(11,27),(11,28),(12,26),(12,27),(12,28),(13,26),(13,
27),(13,28),(14,26),(14,27),(14,28),(15,26),(15,27),(15,28),(16,26),(16,27),
(16,28),(21,1),(21,2),(22,1),(22,2),(23,1),(23,2),(24,1),(24,2),(25,1),(25,
2),(26,1),(26,2),(27,1),(27,2),(28,1),(28,2),(23,3),(23,4),(23,5),(24,3),
(24,4),(24,5),(25,3),(25,4),(25,5),(26,3),(26,4),(26,5),(24,6),(24,7),(24,
8),(24,9),(25,6),(25,7),(25,8),(25,9),(23,12),(23,13),(23,14),(24,12),(24,
13),(24,14),(25,12),(25,13),(25,14),(26,12),(26,13),(26,14),(21,15),(21,16),
(22,15),(22,16),(23,15),(23,16),(24,15),(24,16),(25,15),(25,16),(26,15),(26,
16),(27,15),(27,16),(28,15),(28,16),(21,33),(21,34),(22,33),(22,34),(23,33),
(23,34),(24,33),(24,34),(25,33),(25,34),(26,33),(26,34),(27,33),(27,34),(28,
33),(28,34),(23,35),(23,36),(23,37),(24,35),(24,36),(24,37),(25,35),(25,36),
```

```
(25,37),(26,35),(26,36),(26,37),(24,40),(24,41),(24,42),(24,43),(25,40),(25,
41),(25,42),(25,43),(23,44),(23,45),(23,46),(24,44),(24,45),(24,46),(25,44),
(25,45),(25,46),(26,44),(26,45),(26,46),(21,47),(21,48),(22,47),(22,48),(23,
47),(23,48),(24,47),(24,48),(25,47),(25,48),(26,47),(26,48),(27,47),(27,48),
(28,47),(28,48),(33,21),(33,22),(33,23),(34,21),(34,22),(34,23),(35,21),(35,
22),(35,23),(40,21),(40,22),(40,23),(41,21),(41,22),(41,23),(42,21),(42,22),
(42,23),(43,21),(43,22),(43,23),(44,21),(44,22),(44,23),(45,21),(45,22),(45,
23),(46,21),(46,22),(46,23),(47,21),(47,22),(47,23),(48,21),(48,22),(48,23),
(33,26),(33,27),(33,28),(34,26),(34,27),(34,28),(35,26),(35,27),(35,28),(40,
26),(40,27),(40,28),(41,26),(41,27),(41,28),(42,26),(42,27),(42,28),(43,26),
(43,27),(43,28),(44,26),(44,27),(44,28),(45,26),(45,27),(45,28),(46,26),(46,
27),(46,28),(47,26),(47,27),(47,28),(48,26),(48,27),(48,28)
        self.marker = (9,21),(9,22),(10,21),(10,22),(9,27),(9,28),(10,27),(10,
28),(24,10),(24,11),(25,10),(25,11),(24,38),(24,39),(25,38),(25,39),(36,21),
(36,22),(36,23),(37,21),(37,22),(37,23),(38,21),(38,22),(38,23),(39,21),(39,
22),(39,23),(36,26),(36,27),(36,28),(37,26),(37,27),(37,28),(38,26),(38,27),
(38,28),(39,26),(39,27),(39,28)
        self.uterus = (21,17),(22,17),(23,17),(24,17),(25,17),(26,17),(27,17),
(28,17),(20,18),(21,18),(22,18),(23,18),(24,18),(25,18),(26,18),(27,18),(28,
18),(29,18),(19,19),(20,19),(21,19),(22,19),(23,19),(24,19),(25,19),(26,19),
(27,19),(28,19),(29,19),(30,19),(18,20),(19,20),(20,20),(21,20),(22,20),(23,
20),(24,20),(25,20),(26,20),(27,20),(28,20),(29,20),(30,20),(31,20),(17,21),
(17,22),(17,23),(17,24),(17,25),(17,26),(17,27),(17,28),(18,21),(18,22),(18,
23),(18,24),(18,25),(18,26),(18,27),(18,28),(19,21),(19,22),(19,23),(19,24),
(19,25),(19,26),(19,27),(19,28),(20,21),(20,22),(20,23),(20,24),(20,25),(20,
26),(20,27),(20,28),(21,21),(21,22),(21,23),(21,24),(21,25),(21,26),(21,27),
(21,28),(22,21),(22,22),(22,23),(22,24),(22,25),(22,26),(22,27),(22,28),(23,
21),(23,22),(23,23),(23,24),(23,25),(23,26),(23,27),(23,28),(24,21),(24,22),
(24,23),(24,24),(24,25),(24,26),(24,27),(24,28),(25,21),(25,22),(25,23),(25,
24),(25,25),(25,26),(25,27),(25,28),(26,21),(26,22),(26,23),(26,24),(26,25),
(26,26),(26,27),(26,28),(27,21),(27,22),(27,23),(27,24),(27,25),(27,26),(27,
27),(27,28),(28,21),(28,22),(28,23),(28,24),(28,25),(28,26),(28,27),(28,28),
(29,21),(29,22),(29,23),(29,24),(29,25),(29,26),(29,27),(29,28),(30,21),(30,
22),(30,23),(30,24),(30,25),(30,26),(30,27),(30,28),(31,21),(31,22),(31,23),
(31,24),(31,25),(31,26),(31,27),(31,28),(32,21),(32,22),(32,23),(32,24),(32,
25),(32,26),(32,27),(32,28),(18,29),(19,29),(20,29),(21,29),(22,29),(23,29),
(24,29),(25,29),(26,29),(27,29),(28,29),(29,29),(30,29),(31,29),(19,30),(20,
30),(21,30),(22,30),(23,30),(24,30),(25,30),(26,30),(27,30),(28,30),(29,30),
(30,30),(20,31),(21,31),(22,31),(23,31),(24,31),(25,31),(26,31),(27,31),(28,
31),(29,31),(21,32),(22,32),(23,32),(24,32),(25,32),(26,32),(27,32),(28,32)
        self.peritoneum = (22,3),(22,4),(22,5),(27,3),(27,4),(27,5),(22,35),(22,
36),(22,37),(27,35),(27,36),(27,37),(22,44),(22,45),(22,46),(27,44),(27,45),
(27,46),(6,17),(6,18),(6,19),(6,20),(7,17),(7,18),(7,19),(7,20),(8,17),(8,
18),(8,19),(8,20),(9,17),(9,18),(9,19),(9,20),(10,17),(10,18),(10,19),(10,
20),(11,17),(11,18),(11,19),(11,20),(12,17),(12,18),(12,19),(12,20),(13,17),
(13,18),(13,19),(13,20),(14,17),(14,18),(14,19),(14,20),(15,17),(15,18),(15,
19),(15,20),(16,17),(16,18),(16,19),(16,20),(17,7),(17,8),(17,9),(17,10),
(17,11),(17,12),(17,13),(17,14),(17,15),(17,16),(18,7),(18,8),(18,9),(18,
10),(18,11),(18,12),(18,13),(18,14),(18,15),(18,16),(19,7),(19,8),(19,9),
```

```
(19,10),(19,11),(19,12),(19,13),(19,14),(19,15),(19,16),(20,7),(20,8),(20,
9),(20,10),(20,11),(20,12),(20,13),(20,14),(20,15),(20,16),(17,17),(18,17),
(19,17),(20,17),(17,18),(18,18),(19,18),(17,19),(18,19),(17,20),(21,3),(21,
4),(21,5),(21,6),(21,7),(21,8),(21,9),(21,10),(21,11),(21,12),(21,13),(21,
14),(22,6),(22,7),(22,8),(22,9),(22,10),(22,11),(22,12),(22,13),(22,14),(23,
6),(23,7),(23,8),(23,9),(23,10),(23,11),(26,6),(26,7),(26,8),(26,9),(26,10),
(26,11),(27,6),(27,7),(27,8),(27,9),(27,10),(27,11),(27,12),(27,13),(27,14),
(28,3),(28,4),(28,5),(28,6),(28,7),(28,8),(28,9),(28,10),(28,11),(28,12),
(28,13),(28,14),(29,7),(29,8),(29,9),(29,10),(29,11),(29,12),(29,13),(29,
14),(29,15),(29,16),(30,7),(30,8),(30,9),(30,10),(30,11),(30,12),(30,13),
(30,14),(30,15),(30,16),(31,7),(31,8),(31,9),(31,10),(31,11),(31,12),(31,
13),(31,14),(31,15),(31,16),(32,7),(32,8),(32,9),(32,10),(32,11),(32,12),
(32,13),(32,14),(32,15),(32,16),(33,17),(33,18),(33,19),(33,20),(34,17),(34,
18),(34,19),(34,20),(35,17),(35,18),(35,19),(35,20),(36,17),(36,18),(36,19),
(36,20),(37,17),(37,18),(37,19),(37,20),(38,17),(38,18),(38,19),(38,20),(39,
17),(39,18),(39,19),(39,20),(40,17),(40,18),(40,19),(40,20),(29,17),(30,17),
(31,17),(32,17),(30,18),(31,18),(32,18),(31,19),(32,19),(32,20),(6,23),(6,
24),(6,25),(6,26),(7,23),(7,24),(7,25),(7,26),(8,23),(8,24),(8,25),(8,26),
(9,23),(9,24),(9,25),(9,26),(10,23),(10,24),(10,25),(10,26),(11,24),(11,25),
(12,24),(12,25),(13,24),(13,25),(14,24),(14,25),(15,24),(15,25),(16,24),(16,
25),(33,24),(33,25),(34,24),(34,25),(35,24),(35,25),(36,24),(36,25),(37,24),
(37,25),(38,24),(38,25),(39,24),(39,25),(40,24),(40,25),(6,29),(6,30),(6,
31),(6,32),(7,29),(7,30),(7,31),(7,32),(8,29),(8,30),(8,31),(8,32),(9,29),
(9,30),(9,31),(9,32),(10,29),(10,30),(10,31),(10,32),(11,29),(11,30),(11,
31),(11,32),(12,29),(12,30),(12,31),(12,32),(13,29),(13,30),(13,31),(13,32),
(14,29),(14,30),(14,31),(14,32),(15,29),(15,30),(15,31),(15,32),(16,29),(16,
30),(16,31),(16,32),(33,29),(33,30),(33,31),(33,32),(34,29),(34,30),(34,31),
(34,32),(35,29),(35,30),(35,31),(35,32),(36,29),(36,30),(36,31),(36,32),(37,
29),(37,30),(37,31),(37,32),(38,29),(38,30),(38,31),(38,32),(39,29),(39,30),
(39,31),(39,32),(40,29),(40,30),(40,31),(40,32),(17,33),(17,34),(17,35),(17,
36),(17,37),(17,38),(17,39),(17,40),(17,41),(17,42),(18,33),(18,34),(18,35),
(18,36),(18,37),(18,38),(18,39),(18,40),(18,41),(18,42),(19,33),(19,34),(19,
35),(19,36),(19,37),(19,38),(19,39),(19,40),(19,41),(19,42),(20,33),(20,34),
(20,35),(20,36),(20,37),(20,38),(20,39),(20,40),(20,41),(20,42),(29,33),(29,
34),(29,35),(29,36),(29,37),(29,38),(29,39),(29,40),(29,41),(29,42),(30,33),
(30,34),(30,35),(30,36),(30,37),(30,38),(30,39),(30,40),(30,41),(30,42),(31,
33),(31,34),(31,35),(31,36),(31,37),(31,38),(31,39),(31,40),(31,41),(31,42),
(32,33),(32,34),(32,35),(32,36),(32,37),(32,38),(32,39),(32,40),(32,41),(32,
42),(21,35),(21,36),(21,37),(21,38),(21,39),(21,40),(21,41),(21,42),(21,43),
(21,44),(21,45),(21,46),(22,38),(22,39),(22,40),(22,41),(22,42),(22,43),(23,
38),(23,39),(23,40),(23,41),(23,42),(23,43),(26,38),(26,39),(26,40),(26,41),
(26,42),(26,43),(27,38),(27,39),(27,40),(27,41),(27,42),(27,43),(28,35),(28,
36),(28,37),(28,38),(28,39),(28,40),(28,41),(28,42),(28,43),(28,44),(28,45),
(28,46),(17,32),(18,32),(19,32),(20,32),(17,31),(18,31),(19,31),(17,30),(18,
30),(17,29),(29,32),(30,32),(31,32),(32,32),(30,31),(31,31),(32,31),(31,30),
(32,30),(32,29)
        self.preborder = (1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,
3),(2,4),(2,5),(2,6),(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),(4,1),(4,2),(4,3),
(4,4),(4,5),(4,6),(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),(6,1),(6,2),(6,3),(6,
```

4),(6,5),(6,6),(7,1),(7,2),(7,3),(7,4),(7,5),(7,6),(8,1),(8,2),(8,3),(8,4),
(8,5),(8,6),(9,1),(9,2),(9,3),(9,4),(9,5),(9,6),(10,1),(10,2),(10,3),(10,4),
(10,5),(10,6),(11,1),(11,2),(11,3),(11,4),(11,5),(11,6),(12,1),(12,2),(12,
3),(12,4),(12,5),(12,6),(13,1),(13,2),(13,3),(13,4),(13,5),(13,6),(14,1),
(14,2),(14,3),(14,4),(14,5),(14,6),(15,1),(15,2),(15,3),(15,4),(15,5),(15,
6),(16,1),(16,2),(16,3),(16,4),(16,5),(16,6),(17,1),(17,2),(17,3),(17,4),
(17,5),(17,6),(18,1),(18,2),(18,3),(18,4),(18,5),(18,6),(19,1),(19,2),(19,
3),(19,4),(19,5),(19,6),(20,1),(20,2),(20,3),(20,4),(20,5),(20,6),(1,7),(1,
8),(1,9),(1,10),(1,11),(1,12),(1,13),(1,14),(1,15),(1,16),(2,7),(2,8),(2,9),
(2,10),(2,11),(2,12),(2,13),(2,14),(2,15),(2,16),(3,7),(3,8),(3,9),(3,10),
(3,11),(3,12),(3,13),(3,14),(3,15),(3,16),(4,7),(4,8),(4,9),(4,10),(4,11),
(4,12),(4,13),(4,14),(4,15),(4,16),(5,7),(5,8),(5,9),(5,10),(5,11),(5,12),
(5,13),(5,14),(5,15),(5,16),(6,7),(6,8),(6,9),(6,10),(6,11),(6,12),(6,13),
(6,14),(6,15),(6,16),(7,7),(7,8),(7,9),(7,10),(7,11),(7,12),(7,13),(7,14),
(7,15),(7,16),(8,7),(8,8),(8,9),(8,10),(8,11),(8,12),(8,13),(8,14),(8,15),
(8,16),(9,7),(9,8),(9,9),(9,10),(9,11),(9,12),(9,13),(9,14),(9,15),(9,16),
(10,7),(10,8),(10,9),(10,10),(10,11),(10,12),(10,13),(10,14),(10,15),(10,
16),(11,7),(11,8),(11,9),(11,10),(11,11),(11,12),(11,13),(11,14),(11,15),
(11,16),(12,7),(12,8),(12,9),(12,10),(12,11),(12,12),(12,13),(12,14),(12,
15),(12,16),(13,7),(13,8),(13,9),(13,10),(13,11),(13,12),(13,13),(13,14),
(13,15),(13,16),(14,7),(14,8),(14,9),(14,10),(14,11),(14,12),(14,13),(14,
14),(14,15),(14,16),(15,7),(15,8),(15,9),(15,10),(15,11),(15,12),(15,13),
(15,14),(15,15),(15,16),(16,7),(16,8),(16,9),(16,10),(16,11),(16,12),(16,
13),(16,14),(16,15),(16,16),(1,17),(1,18),(1,19),(1,20)   ,(2,17),(2,18),(2,
19),(2,20),(3,17),(3,18),(3,19),(3,20),(4,17),(4,18),(4,19),(4,20),(5,17),
(5,18),(5,19),(5,20),(1,23),(1,24),(1,25),(1,26),(2,23),(2,24),(2,25),(2,
26),(3,23),(3,24),(3,25),(3,26),(4,23),(4,24),(4,25),(4,26),(5,23),(5,24),
(5,25),(5,26),(1,29),(1,30),(1,31),(1,32),(2,29),(2,30),(2,31),(2,32),(3,
29),(3,30),(3,31),(3,32),(4,29),(4,30),(4,31),(4,32),(5,29),(5,30),(5,31),
(5,32),(1,33),(1,34),(1,35),(1,36),(1,37),(1,38),(1,39),(1,40),(1,41),(1,
42),(1,43),(1,44),(1,45),(1,46),(1,47),(1,48),(2,33),(2,34),(2,35),(2,36),
(2,37),(2,38),(2,39),(2,40),(2,41),(2,42),(2,43),(2,44),(2,45),(2,46),(2,
47),(2,48),(3,33),(3,34),(3,35),(3,36),(3,37),(3,38),(3,39),(3,40),(3,41),
(3,42),(3,43),(3,44),(3,45),(3,46),(3,47),(3,48),(4,33),(4,34),(4,35),(4,
36),(4,37),(4,38),(4,39),(4,40),(4,41),(4,42),(4,43),(4,44),(4,45),(4,46),
(4,47),(4,48),(5,33),(5,34),(5,35),(5,36),(5,37),(5,38),(5,39),(5,40),(5,
41),(5,42),(5,43),(5,44),(5,45),(5,46),(5,47),(5,48),(6,33),(6,34),(6,35),
(6,36),(6,37),(6,38),(6,39),(6,40),(6,41),(6,42),(6,43),(6,44),(6,45),(6,
46),(6,47),(6,48),(7,33),(7,34),(7,35),(7,36),(7,37),(7,38),(7,39),(7,40),
(7,41),(7,42),(7,43),(7,44),(7,45),(7,46),(7,47),(7,48),(8,33),(8,34),(8,
35),(8,36),(8,37),(8,38),(8,39),(8,40),(8,41),(8,42),(8,43),(8,44),(8,45),
(8,46),(8,47),(8,48),(9,33),(9,34),(9,35),(9,36),(9,37),(9,38),(9,39),(9,
40),(9,41),(9,42),(9,43),(9,44),(9,45),(9,46),(9,47),(9,48),(10,33),(10,34),
(10,35),(10,36),(10,37),(10,38),(10,39),(10,40),(10,41),(10,42),(10,43),(10,
44),(10,45),(10,46),(10,47),(10,48),(11,33),(11,34),(11,35),(11,36),(11,37),
(11,38),(11,39),(11,40),(11,41),(11,42),(11,43),(11,44),(11,45),(11,46),(11,
47),(11,48),(12,33),(12,34),(12,35),(12,36),(12,37),(12,38),(12,39),(12,40),
(12,41),(12,42),(12,43),(12,44),(12,45),(12,46),(12,47),(12,48),(13,33),(13,
34),(13,35),(13,36),(13,37),(13,38),(13,39),(13,40),(13,41),(13,42),(13,43),

(13,44),(13,45),(13,46),(13,47),(13,48),(14,33),(14,34),(14,35),(14,36),(14,
37),(14,38),(14,39),(14,40),(14,41),(14,42),(14,43),(14,44),(14,45),(14,46),
(14,47),(14,48),(15,33),(15,34),(15,35),(15,36),(15,37),(15,38),(15,39),(15,
40),(15,41),(15,42),(15,43),(15,44),(15,45),(15,46),(15,47),(15,48),(16,33),
(16,34),(16,35),(16,36),(16,37),(16,38),(16,39),(16,40),(16,41),(16,42),(16,
43),(16,44),(16,45),(16,46),(16,47),(16,48),(17,43),(17,44),(17,45),(17,46),
(17,47),(17,48),(18,43),(18,44),(18,45),(18,46),(18,47),(18,48),(19,43),(19,
44),(19,45),(19,46),(19,47),(19,48),(20,43),(20,44),(20,45),(20,46),(20,47),
(20,48),(29,1),(29,2),(29,3),(29,4),(29,5),(29,6),(30,1),(30,2),(30,3),(30,
4),(30,5),(30,6),(31,1),(31,2),(31,3),(31,4),(31,5),(31,6),(32,1),(32,2),
(32,3),(32,4),(32,5),(32,6),(33,1),(33,2),(33,3),(33,4),(33,5),(33,6),(34,1)
,(34,2),(34,3),(34,4),(34,5),(34,6),(35,1),(35,2),(35,3),(35,4),(35,5),(35,
6),(36,1),(36,2),(36,3),(36,4),(36,5),(36,6),(37,1),(37,2),(37,3),(37,4),
(37,5),(37,6),(38,1),(38,2),(38,3),(38,4),(38,5),(38,6),(39,1),(39,2),(39,
3),(39,4),(39,5),(39,6),(40,1),(40,2),(40,3),(40,4),(40,5),(40,6),(41,1),
(41,2),(41,3),(41,4),(41,5),(41,6),(42,1),(42,2),(42,3),(42,4),(42,5),(42,
6),(43,1),(43,2),(43,3),(43,4),(43,5),(43,6),(44,1),(44,2),(44,3),(44,4),
(44,5),(44,6),(45,1),(45,2),(45,3),(45,4),(45,5),(45,6),(46,1),(46,2),(46,
3),(46,4),(46,5),(46,6),(47,1),(47,2),(47,3),(47,4),(47,5),(47,6),(48,1),
(48,2),(48,3),(48,4),(48,5),(48,6),(33,7),(33,8),(33,9),(33,10),(33,11),
(33,12),(33,13),(33,14),(33,15),(33,16),(34,7),(34,8),(34,9),(34,10),(34,
11),(34,12),(34,13),(34,14),(34,15),(34,16),(35,7),(35,8),(35,9),(35,10),
(35,11),(35,12),(35,13),(35,14),(35,15),(35,16),(36,7),(36,8),(36,9),(36,
10),(36,11),(36,12),(36,13),(36,14),(36,15),(36,16),(37,7),(37,8),(37,9),
(37,10),(37,11),(37,12),(37,13),(37,14),(37,15),(37,16),(38,7),(38,8),(38,
9),(38,10),(38,11),(38,12),(38,13),(38,14),(38,15),(38,16),(39,7),(39,8),
(39,9),(39,10),(39,11),(39,12),(39,13),(39,14),(39,15),(39,16),(40,7),(40,
8),(40,9),(40,10),(40,11),(40,12),(40,13),(40,14),(40,15),(40,16),(41,7),
(41,8),(41,9),(41,10),(41,11),(41,12),(41,13),(41,14),(41,15),(41,16),(42,
7),(42,8),(42,9),(42,10),(42,11),(42,12),(42,13),(42,14),(42,15),(42,16),
(43,7),(43,8),(43,9),(43,10),(43,11),(43,12),(43,13),(43,14),(43,15),(43,
16),(44,7),(44,8),(44,9),(44,10),(44,11),(44,12),(44,13),(44,14),(44,15),
(44,16),(45,7),(45,8),(45,9),(45,10),(45,11),(45,12),(45,13),(45,14),(45,
15),(45,16),(46,7),(46,8),(46,9),(46,10),(46,11),(46,12),(46,13),(46,14),
(46,15),(46,16),(47,7),(47,8),(47,9),(47,10),(47,11),(47,12),(47,13),(47,
14),(47,15),(47,16),(48,7),(48,8),(48,9),(48,10),(48,11),(48,12),(48,13),
(48,14),(48,15),(48,16),(41,17),(41,18),(41,19),(41,20),(42,17),(42,18),
(42,19),(42,20),(43,17),(43,18),(43,19),(43,20),(44,17),(44,18),(44,19),
(44,20),(45,17),(45,18),(45,19),(45,20),(46,17),(46,18),(46,19),(46,20),(47,
17),(47,18),(47,19),(47,20),(48,17),(48,18),(48,19),(48,20),(41,24),(41,25),
(42,24),(42,25),(43,24),(43,25),(44,24),(44,25),(45,24),(45,25),(46,24),(46,
25),(47,24),(47,25),(48,24),(48,25),(41,29),(41,30),(41,31),(41,32),(42,29),
(42,30),(42,31),(42,32),(43,29),(43,30),(43,31),(43,32),(44,29),(44,30),(44,
31),(44,32),(45,29),(45,30),(45,31),(45,32),(46,29),(46,30),(46,31),(46,32),
(47,29),(47,30),(47,31),(47,32),(48,29),(48,30),(48,31),(48,32),(33,33),(33,
34),(33,35),(33,36),(33,37),(33,38),(33,39),(33,40),(33,41),(33,42),(34,33),
(34,34),(34,35),(34,36),(34,37),(34,38),(34,39),(34,40),(34,41),(34,42),(35,
33),(35,34),(35,35),(35,36),(35,37),(35,38),(35,39),(35,40),(35,41),(35,42),
(36,33),(36,34),(36,35),(36,36),(36,37),(36,38),(36,39),(36,40),(36,41),(36,

42),(37,33),(37,34),(37,35),(37,36),(37,37),(37,38),(37,39),(37,40),(37,41),
(37,42),(38,33),(38,34),(38,35),(38,36),(38,37),(38,38),(38,39),(38,40),(38,
41),(38,42),(39,33),(39,34),(39,35),(39,36),(39,37),(39,38),(39,39),(39,40),
(39,41),(39,42),(40,33),(40,34),(40,35),(40,36),(40,37),(40,38),(40,39),(40,
40),(40,41),(40,42),(41,33),(41,34),(41,35),(41,36),(41,37),(41,38),(41,39),
(41,40),(41,41),(41,42),(42,33),(42,34),(42,35),(42,36),(42,37),(42,38),(42,
39),(42,40),(42,41),(42,42),(43,33),(43,34),(43,35),(43,36),(43,37),(43,38),
(43,39),(43,40),(43,41),(43,42),(44,33),(44,34),(44,35),(44,36),(44,37),(44,
38),(44,39),(44,40),(44,41),(44,42),(45,33),(45,34),(45,35),(45,36),(45,37),
(45,38),(45,39),(45,40),(45,41),(45,42),(46,33),(46,34),(46,35),(46,36),(46,
37),(46,38),(46,39),(46,40),(46,41),(46,42),(47,33),(47,34),(47,35),(47,36),
(47,37),(47,38),(47,39),(47,40),(47,41),(47,42),(48,33),(48,34),(48,35),(48,
36),(48,37),(48,38),(48,39),(48,40),(48,41),(48,42),(29,43),(29,44),(29,45),
(29,46),(29,47),(29,48),(30,43),(30,44),(30,45),(30,46),(30,47),(30,48),(31,
43),(31,44),(31,45),(31,46),(31,47),(31,48),(32,43),(32,44),(32,45),(32,46),
(32,47),(32,48),(33,43),(33,44),(33,45),(33,46),(33,47),(33,48),(34,43),(34,
44),(34,45),(34,46),(34,47),(34,48),(35,43),(35,44),(35,45),(35,46),(35,47),
(35,48),(36,43),(36,44),(36,45),(36,46),(36,47),(36,48),(37,43),(37,44),(37,
45),(37,46),(37,47),(37,48),(38,43),(38,44),(38,45),(38,46),(38,47),(38,48),
(39,43),(39,44),(39,45),(39,46),(39,47),(39,48),(40,43),(40,44),(40,45),(40,
46),(40,47),(40,48),(41,43),(41,44),(41,45),(41,46),(41,47),(41,48),(42,43),
(42,44),(42,45),(42,46),(42,47),(42,48),(43,43),(43,44),(43,45),(43,46),(43,
47),(43,48),(44,43),(44,44),(44,45),(44,46),(44,47),(44,48),(45,43),(45,44),
(45,45),(45,46),(45,47),(45,48),(46,43),(46,44),(46,45),(46,46),(46,47),(46,
48),(47,43),(47,44),(47,45),(47,46),(47,47),(47,48),(48,43),(48,44),(48,45),
(48,46),(48,47),(48,48)

```python
    def move_knife(self, action):
        if action == 4 or (action, self.previous_action) in forbidden_moves:
            action = self.previous_action
        else:
            self.previous_action = action
        head = self.knife[0]
        if action == 0:
            p = (head[0] - 1, head[1])
        elif action == 1:
            p = (head[0] - 1, head[1])
        elif action == 2:
            p = (head[0], head[1] - 1)
        elif action == 3:
            p = (head[0], head[1] + 1)
        self.knife.insert(0, p)
        self.knife.pop()

    def get_state(self):
        canvas = np.ones((self.grid_size, ) * 2)
        canvas[1:-1, 1:-1] = 0.0
        for seg in self.knife:
            canvas[seg[0], seg[1]] = 0.8
```

```python
        for seg in self.art_border:
            canvas[seg[0],seg[1]] = 0.9
        for seg in self.mark_border:
            canvas[seg[0], seg[1]] = 0.5
        for seg in self.ut_border:
            canvas[seg[0],seg[1]] = 0.7
        for seg in self.per_border:
            canvas[seg[0],seg[1]] = 0.1
        return canvas

    def get_score(self):
        if self.game_over:
            score = -1
        elif self.plus_scored:
            score = +1
        elif self.minus_scored:
            score = -1
        elif self.non_scored:
            score = 0
        else:
            score = -1
        return score

    def reset(self):
        grid_size = self.grid_size
        knife_length = self.knife_length
        self.knife = [(10,24)]
        self.game_over = False
        self.plus_scored = False
        self.minus_scored = False
        self.non_scored = False
        self.arrange()
        if np.random.randint(2) == 0:
            self.previous_action = 0
        else:
            self.previous_action = 1
        self.border    []
        for z in range(grid_size):
          self.border += [(z, 0), (z, grid_size - 1), (0, z), (grid_size - 1, z)]
        self.border += self.preborder
        self.ut_border = []
        self.ut_border += self.uterus
        self.art_border = []
        self.art_border += self.artery
        self.mark_border = []
        self.mark_border += self.marker
        self.per_border = []
        self.per_border += self.peritoneum
```

```
def left(self):
    self.play(0)
def right(self):
    self.play(1)
def up(self):
    self.play(2)
def down(self):
    self.play(3)
def idle(self):
    self.play(4)

def hit_border(self):
    return self.knife[0] in self.border or self.knife[0] in self.ut_border
def hit_artery(self):
    return self.knife[0] in self.art_border
def hit_marker(self):
    return self.knife[0] in self.mark_border
def hit_peritoneum(self):
    return self.knife[0] in self.per_border

def is_over(self):
    return self.hit_border()
def is_won(self):
    return len(self.mark_border) < 1
```

Author details

Masakazu Sato*, Kaori Koga, Tomoyuki Fujii and Yutaka Osuga

*Address all correspondence to: masakasatou-tky@umin.ac.jp

Department of Obstetrics and Gynecology, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku, Tokyo, Japan

References

[1] Ravi D, Wong C, Deligianni F, Berthelot M, Andreu Perez J, Lo B, et al. Deep learning for health informatics. IEEE Journal of Biomedical and Health Informatics. Jan 2017;**21**(1): 4-21. DOI: 10.1109/JBHI.2016.2636665. PubMed PMID: 28055930

[2] Scholkopf B. Artificial intelligence: Learning to see and act. Nature. 2015;**518**(7540):486-487. DOI: 10.1038/518486a. Epub 2015/02/27, PubMed PMID: 25719660

[3] Zhang YC, Kagen AC. Machine learning Interface for medical image analysis. Journal of Digital Imaging. Oct 2017;**30**(5):615-621. DOI: 10.1007/s10278-016-9910-0. PubMed PMID: 27730415

[4] Gibney E. DeepMind algorithm beats people at classic video games. Nature. 2015;**518**(7540): 465-466. DOI: 10.1038/518465a. Epub 2015/02/27, PubMed PMID: 25719643

[5] Gibney E. Google AI algorithm masters ancient game of Go. Nature. 2016;**529**(7587): 445-446. DOI: 10.1038/529445a. Epub 2016/01/29, PubMed PMID: 26819021

[6] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature. 2015;**518**(7540):529-533. DOI: 10.1038/nature14236. PubMed PMID: 25719670

[7] Littman ML. Reinforcement learning improves behaviour from evaluative feedback. Nature. 2015;**521**(7553):445-451. DOI: 10.1038/nature14540. PubMed PMID: 26017443

[8] Wallace SK, Fazzari MJ, Chen H, Cliby WA, Chalas E. Outcomes and postoperative complications after hysterectomies performed for Benign compared with malignant indications. Obstetrics and Gynecology. 2016;**128**(3):467-475. DOI: 10.1097/AOG.0000000000001591. PubMed PMID: 27500339

[9] Rampasek L, Goldenberg A. TensorFlow: Biology's gateway to deep learning? Cell Systems. 2016;**2**(1):12-14. DOI: 10.1016/j.cels.2016.01.009. PubMed PMID: 27136685

[10] Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks. 2015;**61**:85-117. DOI: 10.1016/j.neunet.2014.09.003. PubMed PMID: 25462637

[11] Beyer-Berjot L, Berdah S, Hashimoto DA, Darzi A, Aggarwal R. A virtual reality training curriculum for laparoscopic colorectal surgery. Journal of Surgical Education. 2016;**73**(6):932-941. DOI: 10.1016/j.jsurg.2016.05.012. Epub 2016/06/28; PubMed PMID: 27342755

[12] Khan ZA, Kamal N, Hameed A, Mahmood A, Zainab R, Sadia B, et al. SmartSIM—A virtual reality simulator for laparoscopy training using a generic physics engine. The International Journal of Medical Robotics + Computer Assisted Surgery: MRCAS. 2016;**16**:437. DOI: 10.1002/rcs.1771. Epub 2016/09/28; PubMed PMID: 27671920

[13] Li XL, Du DF, Jiang H. The learning curves of robotic and three-dimensional laparoscopic surgery in cervical cancer. Journal of Cancer. 2016;**7**(15):2304-2308. DOI: 10.7150/jca.16653. PubMed PMID: 27994668; PubMed Central PMCID: PMCPMC5166541

[14] Romero-Loera S, Cárdenas-Lailson LE, de la Concha-Bermejillo F, Crisanto-Campos BA, Valenzuela-Salazar C, Moreno-Portillo M. Skills comparison using a 2D vs. 3D laparoscopic simulator. Cirugia y Cirujanos. 2016;**84**(1):37-44. DOI: 10.1016/j.circen.2015.12.012. (English Edition)

[15] Kusy M, Zajdel R. Application of reinforcement learning algorithms for the adaptive computation of the smoothing parameter for probabilistic neural network. IEEE Transactions on Neural Networks and Learning Systems. 2015;**26**(9):2163-2175. DOI: 10.1109/TNNLS.2014.2376703. PubMed PMID: 25532211

[16] Senda K, Hattori S, Hishinuma T, Kohda T. Acceleration of reinforcement learning by policy evaluation using nonstationary iterative method. IEEE Transactions on Cybernetics. 2014;**44**(12):2696-2705. DOI: 10.1109/TCYB.2014.2313655. PubMed PMID: 24733037

[17] Xu B, Yang C, Shi Z. Reinforcement learning output feedback NN control using deterministic learning technique. IEEE Transactions on Neural Networks and Learning Systems. 2014;**25**(3):635-641. DOI: 10.1109/TNNLS.2013.2292704. PubMed PMID: 24807456

Application of AI in Modeling of Real System in Chemistry

M. H. Ahmadi Azqhandi and M. Shekari

Additional information is available at the end of the chapter

Abstract

In recent years, discharge of synthetic dye waste from different industries leading to aquatic and environmental pollution is a serious global problem of great concern. Hence, the removal of dye prediction plays an important role in wastewater management and conservation of nature. Artificial intelligence methods are popular owing due to its ease of use and high level of accuracy. This chapter proposes a detailed review of artificial intelligence-based removal dye prediction methods particularly multiple linear regression (MLR), artificial neural networks (ANNs), and least squares-support vector machine (LS-SVM). Furthermore, this chapter will focus on ensemble prediction models (EPMs) used for removal dye prediction. EPMs improve the prediction accuracy by integrating several prediction models. The principles, advantages, disadvantages, and applications of these artificial intelligence-based methods are explained in this chapter. Furthermore, future directions of the research on artificial intelligence-based removal dye prediction methods are discussed.

Keywords: multiple linear regression (MLR), artificial neural networks (ANNs), least squares-support vector regression (LS-SVM)

1. Introduction

Recently, pollution of water sources by various contaminants becomes a global environmental issue [1]. Among different types of water contaminants, dyes as part of human's life are a major contamination group [2]. Dyes are widely used as coloring agents in the textile, plastics, paper, leather, food, antiseptics, cosmetics, fungicides, and so forth and can be entered to the environment through colored wastewater from these industries [3]. However, during the coloration processes, a significant amount (20–50%) of these dyes is lost and released into

the environment as colored wastewater. However, due to toxicity, carcinogenicity, mutagenic and teratogenic properties, and a long-standing environmental pollution, contaminants have become a great environmental concern with potential adverse effects to human health [4]. Also, it can enter the body via the pulmonary system or the digestive system by ingesting contaminated water or food [5–7]. Even though they can be influenced on the photosynthesis process through reduce light penetration and result in reduction oxygen levels in water and, in severe case, consequential the suffocation of aquatic flora and fauna.

Therefore, it is necessary to be removed from wastewater before discharging into bodies of water. Though a number of processes such as ozonation, filtration, membrane, coagulation, precipitation, adsorption, electrochemical techniques, and biosorption have been applied to treat colored textile wastewater from aqueous media [8–11]. However, most of the mentioned methods have shown various restrictions including generation of huge amounts of sludge by the means of electrochemical and chemical coagulation processes or request to high technology with high cost in the membrane technology and advanced oxidation process [12]. Among different treatment methods, adsorption has found particular attention from the researchers worldwide due to the fact that the adsorption is an easy operating, effective, single, and cost-effective option for pollutant removal from aqueous environment.

Although the adsorption is an easy operating technique, it is known as a complicated process in the chemistry and dependent on several factors which have a direct impact on the process performance. Thus, it is vital to select an appropriate mathematical model for optimizing and predicting the removal process. The modeling and optimization of adsorption are still in the stage of research. Commonly used models for describing kinematic and/or equilibrium studies (e.g., second-order models or Langmuir, Freundlich) may be inadequate in determining the relationship between factors and evaluating their effect on the absorption process. Optimization is a way for determining the best solution in terms of certain quality criteria, such as process efficiency, and results in improving the performance of the process or designed system [13, 14].

The optimization of the adsorption process tries to find out the design and/or environmental parameters at which the adsorption process would give the best efficiency (**Figure 1**) [15].

Typically, experiments are carried out in a way that one factor is applied and then analyzed, while other factors remain unaffected. In the usual way, one-factor-at-a-time approach is generally time-consuming, and it is impossible to achieve optimal desirability due to the need for a large number of experiments and lack of interaction among factors. This approach is time-consuming; the researcher must screen all the variables independently and require a large number of tests, leading to high cost of study. In addition, a variable in time does not include interactions between selected parameters. This method is called one variable at time (OVAT) [16].

Multivariate statistics techniques (MST) can significantly reduce the number of experiments and explanations of the independent variable (in combination or individually) in the process. MST helps to develop and optimize the operating system, significantly reducing the cost of testing [17].
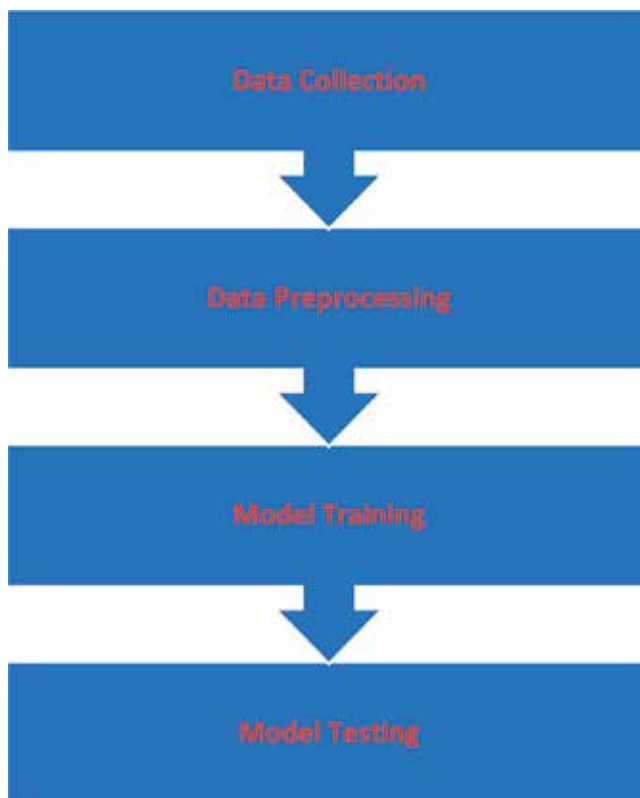
Figure 1. Framework for AI-based prediction model.

Adsorption process is a complex process; therefore, due to complexity of the relationship between output and input parameters, it is difficult to be modeled using statistical approaches. Computational intelligence models are often more flexible than statistical models when modeling complex datasets with possible nonlinearities or missing data [18]. Recently, powerful AI prediction method, such as random forest (RF), adaptive neuro-fuzzy inference system (ANFIS), least square-support vector regression (LS-SVR), radial basis function neural network (RBF-NN), boosted regression tree (BRT), and artificial neural network (ANN) in modeling adsorption process have been successfully used [18–24]. AI is the branch of computer science concerned with making computers behave like humans. The term was coined in 1956 by John McCarthy at the Massachusetts Institute of Technology; it can be employed to explain and model many complex chemical systems because of its reliability, robustness, simplicity, and nonlinearity. AI approach can be learned from experimental data to solve of the complex nonlinear, multidimensional functional relationships without any prior assumptions about their nature [25].

The previous reviews of adsorption procedure and engineering applications of AI confirmed that there is no specific review on the usage of AI for adsorption process.

The two main objectives of this study are (i) summarizing research on the absorption of dyes by AI models and (ii) providing more research needs for AIs for dye absorption.

2. AI definition

AI is a subcategory of computer science. Its goal is to enable the development of computers that are able to do things normally done by people. The Stanford researcher, John McCarthy, was named, in 1956, at the current Dartmouth conference, where the mission of the AI field was defined. If we start with this definition, each program can be considered as AI if it does something that we usually think as intelligent in humans.

The typical AI-based prediction method consists of four main steps. The first step is to acquire input and output data. Input data are those aspects that affect or relate to output data. These aspects include but are not limited to pH, sonication time, adsorbent dose, temperature, and initial concentration of contaminant. The output data is removal percentage. The next step is to preprocess the collected data in an appropriate format before using them to train the forecast model. Some data pre-processing techniques such as data normalization, data transfer, and data interpolation are applied at this stage to improve data quality and reduce negative impact. When the data is ready, the third step is to train the prediction model.

Since the crucial concept of empirical modeling is learning from historical data, a training process is essential for the development of the model. This step is achieved by selecting the appropriate parameters for the model. The type of parameters is determined by the algorithms selected by the researcher, while selecting the proper parameter can guarantee the performance of the model. The last step is testing the model. At this stage, the data test is examined to test the prediction of model performance. Performance indicators such as RMSE, R^2, MAE, and AAD are used to evaluate performance.

The AI can be more classified into four types (i.e., multiple linear regression (MLR), artificial neural network (ANN), least square-support vector machine (LS-SVM), and ensemble prediction models) based on the learning algorithms; the following part of this section describes main techniques used for AI-based prediction model.

2.1. Multiple linear regression

MLR is a statistical technique that uses several explanatory variables to predict a response variable. The goal of MLR is to model the relationship between the input and response variables [26, 27].

The model for MLR, given n observations, is

$$y_i \ = \ B_0 + B_1 \, x_{i1} + B_2 \, x_{i2} + \ldots + B_p \, x_{ip} + \mathrm{E} \quad \text{where i } \epsilon [1{:}n]. \tag{1}$$

The response surface methodology (RSM) is the most popular methods used in the absorption research literature. RSM determine the mathematical relation between parameters and responses. Modeling or model fitting in RSM consists of two steps: coding of experimental data and regression. For the first, input and output data was coded by using the general equation due to RSM operate on coded input values like +1, 0, and − 1 instead of actual values [28].

In the next step, coded experimental data are fitted to a selected model using multiple linear regression (MLR). In spite of the simplicity of MLR, it was used recently in the removal processes.

As the first reported work was in 2000, conducted by Annadurai. Annadurai [29] developed regression models to predict Direct Scarlet B. The inputs for the regression models include the pH, temperature, and the particle size. The proposed models showed promising features to be easy and efficient forecast tools for calculating removal percentage dye from aqueous solution. More recently, our group simplified their MLR model by introducing only four inputs, namely, pH, sonication time, adsorbent dose, and the initial dye concentration. Their results indicated that the proposed method cannot well predict the removal dye percentage [13, 30–35].

Ease of use is accounted as one of the main advantages of the MLR method because no parameter needs to be adjusted. Meanwhile, since no detailed physical information is required, this method is efficient and cost-effective. Nonetheless, the MLR is the main constraint due to the inability to deal with nonlinear problems, although the previous research has proven that the MLR can be used as an efficient tool for predicting parentage removal [10, 25, 36–48].

2.2. Artificial neural network (ANN)

ANNs are computing systems inspired by the biological neural networks that constitute animal brains [25, 49]. An ANN is based on a collection of connected units or nodes called artificial neurons (analogous to biological neurons in an animal brain). Each connection between artificial neurons can transmit a signal from one to another. The artificial neuron (AN) that receives the signal can process it and then signal artificial neurons connected to it. ANs typically have a weight that increases or decreases the strength of the signal at a connection. Signals travel from the first input, to the last output layer, possibly after traversing the layers multiple times. The original goal of the ANN approach was to solve problems in the same way that a human brain would. Over time, attention is focused on matching-specific mental abilities, leading to deviations from biology. ANNs have been used on a variety of tasks, including speech recognition, computer vision, social network filtering, machine translation, playing board and video games, and medical diagnosis.

In the past two decades, many studies have been carried out to predict various types of decolorization dye from aqueous solution, such as electrocoagulation process [49], Fenton process [50], and adsorption [51] by applying ANNs. M. Ahmadi and Kh. Naderi applied general regression neural network (GRNN) to predict the removal of methylene blue (MB) and Basic Yellow 28 (BY28) from aqueous solution. Their findings indicated that a well-designed GRNN is able to predict the removal of azo dye based on sonication time, initial dye concentration, and adsorbent mass. Ahmadi and J. Pooralhossini used backpropagation neural network (BPNN) to predict the decolorization of sunset yellow (SY) and disulfine blue (DB) [52]. The obtained results show that the BPNN model outperforms the classical statistical model in terms of R^2, RMSE, MAE, and AAD for both dyes. Ahmadi and team used BPNN to predict the efficiency of two carcinogenic dye (methylene blue (MB) and malachite green (MG)) adsorption onto Mn@ CuS/ZnS nanocomposite-loaded activated carbon (Mn@ CuS/ZnS-NC-AC) as a novel adsorbent to identify the model parameters in order to improve the prediction performance [35]. Ahmadi and Dastkhoon used neural network to predict Safranin-O (SO) and indigo carmine (IC) adsorption onto Ni:FeO(OH)-NWs-AC. In this work, the influence of process variables (initial dye concentration, adsorbent mass, and sonication time) on the removal of both dyes was investigated by central composite rotatable design (CCRD) of RSM, multilayer perceptron (MLP) neural network, and Doolittle factorization algorithm (DFA). The ANN model

was found to be more precise than the other models. We performed the sensitivity analysis (by using of weight neuron) and confirmed that sonication time has the essential factor affecting the removal of SO and IC [33]. Ahmadi et al. developed a BP neural network model and partial least squares (PLS) to predict the ultrasonic-assisted simultaneous removal of fast green (FG), eosin Y (EY), and quinine yellow (QY) from aqueous media following the use of MOF-5 as a metal organic framework and activated carbon hybrid (AC-MOF-5). The obtained results show that ANN and PLS model is a powerful tool for prediction of under-study dye adsorption by AC-MOF-5 [53].

The main advantage of ANN method is its ability to detect complex nonlinear relationship between the inputs and outputs that this characteristic makes it possible to be applied for real systems. However, ANN method fails to establish any interconnection relationship between building physical parameters and removal percentage, which limits the model's fitting ability.

2.3. Least square-support vector machine (LS-SVM)

SVM as a learning method was developed by Vapnik and is a powerful tool [34, 54]. This supervised learning method can be used for regression or classification in nonlinear models, and density estimation leads to complex optimization problems, typically quadratic programming. However, this method (SVM) is often time-consuming and difficult to adapt, suffering from the problem of a large memory requirement and CPU time when trained in batch mode. This limitation is overcome by LS-SVM as the modified version of SVM which solves the set of linear equations instead of the quadratic programming problem to minimize the complex nature of the optimization processes. The theory and more details of SVM and LS-SVM can be found in the literature.

Liang and team first applied LS-SVM in the area of water quality measurements in 2011. They showed that the model output made a relatively good training fitting effect and the predict effect was relatively satisfactory. The model has not only a good learning accuracy but a good generalization ability. The predictive fitting precision of the test data set was more than 90%, and the prediction error is minimum, and RMSE is 0.0028 [55].

Similarly, our group used LS-SVM for the optimization and/or modeling of pH, ZnS-NPs-AC mass, MB concentration, and sonication time to develop respective predictive equations for the simulation of the efficiency of MB adsorption. The obtained results using LS-SVM exhibit a nonlinear approach which shows better performances in comparison to central composite design (CCD) for the prediction of MB adsorption [34].

Niyaz Mohammad Mahmoodi and team used of least square-support vector machine (LS-SVM) to model the dye removal. The graphical plots and the values of statistical parameter showed LS-SVM as an intelligent model suitable for modeling of dye adsorption [56].

Our group compared LS-SVM with other AI-based prediction methods in removal dye prediction. We compared SVR with several ANN models for prediction of the adsorption of methylene blue (MB) from aqueous solutions by zinc sulfide nanoparticles with activated carbon (ZnS-NPs-AC). Also, a multiple linear regression (MLR) model and LS-SVM model with principal component analysis (PCA) were used for pre-processing to predict the efficiency of

methylene blue adsorption onto copper oxide nanoparticle loaded on activated carbon (CuO-NP-AC) by Ghaedi et al. Both studies indicate that SVM has a better performance in building model and prediction than other AI-based prediction methods [57].

2.3.1. Advantages and limitations

The main advantage of LS-SVM that was introduced by Foucquier et al. [58] is based on the structural risk minimization principle which aims to minimize the upper bound of the general error consisting of the sum of the training error. Also, SVM provides a better balance between prediction accuracy and computation speed comparing with ANNs and RSM [34]. The limitation of SVM method is the determination of kernel function. There is no uniform standard for determining which kernel will result in the most accurate SVM. Researchers have to determine the kernel function based on the characteristics of the data as well as their own experience.

2.3.2. Ensemble prediction models

Whereof each prediction method has its own limitations, currently, a trend has evolved to introduce new mathematical methods, called ensemble learning. These methods have been more extensive in analytical chemistry for complex data analysis. The adaptability of data mining methods makes them able to deal with typical problems: too many descriptors in the model, mixtures of different data types, complex data having missing values, multiple classes, or unbalanced data sets. For instance, some data mining methods such as support vector machine [34] and artificial neural networks (ANNs) [59] have been applied in different realms of science and engineering. The Ensemble prediction has become increasingly popular in chemistry [60, 61].

This method differs from each different prediction method because it creates a composite model which integrates different individual prediction models. Instead of a prediction algorithm, this method acts as a framework to reduce prediction errors by combining different reduction algorithms together. Ensemble prediction methods have been successfully applied in several areas of chemistry that they have large data volumes, including spectroscopy [62, 63], quantitative structure–activity relationship (QSAR) modeling [64, 65], and omics sciences [66]. Ensemble prediction is able to manage with many types of responses and predictors such as categorical or numeric and loss functions such as Laplace, Gaussian, Poisson, and Bernoulli [67]. De'ath [68] has shown that ensemble prediction methods, unlike many other regression methods, can be used for both prediction and explanation of the underlying relationships between response and predictors.

The two main outputs of an ensemble prediction model which are the partial dependence plots and the variable importance rankings can be used together for model interpretation. Friedman [69] has proposed that while these outputs might not offer a complete description, they can at least give an insight of the relation between the response and the predictors.

Some authors (e.g., Hastie et al.) [70] have shown that ensemble prediction is one of the most powerful machine/statistical learning ideas than have been presented during the 1990s, and

it has been suggested [71, 72] that the application of ensemble prediction to classification and regression trees results in individual classifiers (e.g., classification trees, regression trees) which generally are competitive with any other method.

In addition, Breiman [71] shows that applying ensemble prediction to classification and regression tree can actually be much quicker than fitting a neural net classifier.

Due to the high prediction accuracy, ensemble learning method has become a favorable topic in recent years and has already been applied to many fields successfully. For example, our group [73] developed boosted regression tree (BRT), an ensemble method for fitting statistical models that differ fundamentally from conventional techniques that aim to fit a single parsimonious model. In this study, response surface methodology (RSM), artificial neural network (ANN), and BRT have been used for the optimization and/or modeling of stirred time (min), pH, adsorbent mass (mg), and concentrations of MB and Cd^{2+} ions (mg L^{-1}) to develop respective predictive equations for simulation of the efficiency of MB and Cd^{2+} adsorption based on experimental data set achieved in batch study. All three models showed good predictions in this study. But the BRT model was more precise than the other models, and it showed that BRT is a powerful tool for modeling and optimizing removal of MB and Cd(II).

Similarly, our group [32] used adaptive network-based fuzzy inference system (ANFIS) ensemble models as a support tool for examining data and making prediction to recognize and predict the removal percentage in MB and SY dye solution of different concentrations. The predictive capabilities of MLR and ANFIS are compared in terms of square correlation coefficient (R^2), root mean square error (RMSE), mean absolute error (MAE), and absolute average deviation (AAD) against the empirical data. It is found that the ANFIS model shows the better prediction accuracy than the CCD model.

In another work by our group [13], random forest (RF) and response surface methodology (RSM) were used to model and predict the efficiency of malachite green removal from aqueous solution by ultrasound-assisted adsorption onto the silver hydroxide nanoparticles loaded on activated carbon (AgOH-NPs-AC). The parameters such as pH, initial MG concentration, sonication time, and adsorbent dosage involved in the adsorption process were set within the ranges 2.0–10, 4–20 mg L^{-1}, 2–6 min, and 0.005–0.025 g, respectively. The performance of the RF and CCD models for the description of experimental data was evaluated in terms of R^2, RMSE, MAE, and AAD. The obtained results showed that the RF model outperformed classical statistical model for modeling the process of dye adsorption.

Also, ensemble prediction approach (i.e., radial basis function neural network (RBF-NN) and random forest (RF)) was developed and evaluated against a quadratic response surface model to predict the maximum removal efficiency of brilliant green (BG) from aqueous media in relation to BG concentration (4–20 mg L^{-1}), sonication time (2–6 min), and ZnS-NP-AC mass (0.010–0.030 g) by ultrasound-assisted adsorption [31]. All three (i.e., RBF network, RF, and polynomial) models were compared against the experimental data using four statistical indices, namely, R^2, RMSE, MAE, and AAD. Graphical plots were also used for model comparison. The obtained results using RBF network and RF exhibit a better performance than MLR for both dyes.

The main advantage of the ensemble method is the improvement of accuracy. Also, these methods incorporate important advantages such as accommodating missing data and handling

different types of predictor variables. In addition, they have no need for prior data transformation or elimination of outliers, can fit complex nonlinear relationships, and can automatically handle interaction effects between predictors. However, compared with other predictive methods, ensemble models require more time to calculate and a high level of knowledge as a combination of different base models. Another disadvantage of the model is the fact that its predictive function depends on the selection of base models. In the previous study, the researchers selected the base model based on their previous knowledge. There is a lack of approach to determine which base model should be considered and included in the ensemble model.

3. Discussion

According to previous researches, each type of AI-based prediction method has its own disadvantages and advantages; thus scientists have to select suitable method to solve their problems. For example, MLR is more appropriate than other methods in predicting removal dye because of its ease of use and high calculation speed. While LS-SVM and ANNs are more suitable for real system with high nonlinearity because of their high level of prediction accuracy. On the other hand, some researchers have tried to compare AI-based prediction methods with other methods in removal dye prediction. For example, Tanzifi [74] compared ANNs with RSM for predicting removal Amido Black 10B. The comparison of the adsorption efficiencies obtained by the ANN model and the experimental data evidenced that the ANN model could estimate the behavior of the Amido Black 10B dye adsorption process under various conditions. Their study proposed ANN model as a simpler and more efficient building energy prediction tool when compared with energy simulation software. Based on these researches, the advantages and disadvantages of AI-based prediction methods are summarized below.

3.1. Advantages

The advantages of AI model are:

1. Comparing AI with other engineering approaches shows that the building and/or development of the AI-based prediction methods does not need any detailed physical information which in return saves both cost and time for leading the prediction.

2. Based on previous study, if model is well trained, AI methods give promising prediction accuracy.

3. The data gaining and data loading process is relatively simple, which means the prediction model can be easily obtained.

4. Conclusion

There are many published papers on the prediction of dye adsorption using OVAT. However, in this chapter, we review the important research studies on dye adsorption forecasting using

AI methods. The literature survey in this chapter showed that the AI approaches can be successfully used for the modeling and predication of dye adsorption process with acceptable accuracy compared to conventional linear models such as RSM. The future research proposed for AIs in the field of dye removal for carrying out extensive studies are as follows:

i.   The prediction capability of other AI models such as group method of data handling (GMDH), random forest (RF), neural gas network, regression tree, and radial basis function network (RBFN) for dye adsorption needs further research studies.

ii. The hybridization of the AIs together such as LS-SVM and GMDH forecasting methods, regarding the potential of predict dye adsorption, was proposed.

iii. A few studies have been reported about combining ANN approaches with optimization algorithms. However, it is necessary to extend the optimization of network configuration for modeling adsorption process using an evolutionary computation method such as ant colony, PSO, GA algorithm, tabu search, artificial bee colony, firefly algorithm, teaching-learning-based optimization, harmony search, shuffled frog-leaping algorithm, simulated annealing, and invasive weed optimization.

Based on the reported study and discussions presented in this chapter, it can be concluded that AI methods are excellent approaches for the adsorption of dyes. The information offered in this chapter would be highly useful to the scientists working in the field of dye adsorption and AIs in their investigations.

Author details

M. H. Ahmadi Azqhandi* and M. Shekari

*Address all correspondence to: mhahmadia58@gmail.com

Applied Chemistry Department, Faculty of Petroleum and Gas (Gachsaran), Yasouj University, Gachsaran, Iran

References

[1] Silva TL, Cazetta AL, Souza PS, Zhang T, Asefa T, Almeida VC. Mesoporous activated carbon fibers synthesized from denim fabric waste: Efficient adsorbents for removal of textile dye from aqueous solutions. Journal of Cleaner Production. 2018;**171**:482-490

[2] Sarvajith M, Reddy GKK, Nancharaiah Y. Textile dye biodecolourization and ammonium removal over nitrite in aerobic granular sludge sequencing batch reactors. Journal of Hazardous Materials. 2018;**342**:536-543

[3] Li L, Qi G, Wang B, Yue D, Wang Y, Sato T. Fulvic acid anchored layered double hydroxides: A multifunctional composite adsorbent for the removal of anionic dye and toxic metal. Journal of Hazardous Materials. 2018;**343**:19-28

[4] Liu C, Omer A. Adsorptive removal of cationic methylene blue dye using carboxymethyl cellulose/k-carrageenan/activated montmorillonite composite beads: Isotherm and kinetic studies. International Journal of Biological Macromolecules. 2018;**106**:823-833

[5] Gao H, Zhao S, Cheng X, Wang X, Zheng L. Removal of anionic azo dyes from aqueous solution using magnetic polymer multi-wall carbon nanotube nanocomposite as adsorbent. Chemical Engineering Journal. 2013;**223**:84-90

[6] Zhu H-Y, Fu Y-Q, Jiang R, Yao J, Xiao L, Zeng G-M. Novel magnetic chitosan/poly (vinyl alcohol) hydrogel beads: Preparation, characterization and application for adsorption of dye from aqueous solution. Bioresource Technology. 2012;**105**:24-30

[7] Madrakian T, Afkhami A, Ahmadi M, Bagheri H. Removal of some cationic dyes from aqueous solutions using magnetic-modified multi-walled carbon nanotubes. Journal of Hazardous Materials. 2011;**196**:109-114

[8] Verma AK, Dash RR, Bhunia P. A review on chemical coagulation/flocculation technologies for removal of colour from textile wastewaters. Journal of Environmental Management. 2012;**93**(1):154-168

[9] Asgher M, Bhatti HN. Evaluation of thermodynamics and effect of chemical treatments on sorption potential of *Citrus* waste biomass for removal of anionic dyes from aqueous solutions. Ecological Engineering. 2012;**38**(1):79-85

[10] Körbahti BK, Artut K, Geçgel C, Özer A. Electrochemical decolorization of textile dyes and removal of metal ions from textile dye and metal ion binary mixtures. Chemical Engineering Journal. 2011;**173**(3):677-688

[11] Gupta V. Application of low-cost adsorbents for dye removal—A review. Journal of Environmental Management. 2009;**90**(8):2313-2342

[12] Azqhandi MHA, Rajabi F, Keramati M. Synthesis of Cd doped ZnO/CNT nanocomposite by using microwave method: Photocatalytic behavior, adsorption and kinetic study. Results in Physics. 2017:1106-1114

[13] Solaymani E, Ghaedi M, Karimi H, Azqhandi A, Hossein M, Asfaram A. Intensified removal of malachite green by AgOH-AC nanoparticles combined with ultrasound: Modeling and optimization. Applied Organometallic Chemistry. 2017:1-12

[14] Amini M, Yazdanbakhsh A, Eslami A, Aghayani E. Optimization of coagulation-flocculation process for dye and COD removal from real dyeing wastewater and evaluation of effluent biodegradability in a carpet factory. Journal of Health in the Field. 2018;**13**(3):24-33

[15] Tuzen M, Sarı A, Saleh TA. Response surface optimization, kinetic and thermodynamic studies for effective removal of rhodamine B by magnetic AC/CeO$_2$ nanocomposite. Journal of Environmental Management. 2018;**206**:170-177

[16] Sleiman M, Vildozo D, Ferronato C, Chovelon J-M. Photocatalytic degradation of azo dye Metanil yellow: Optimization and kinetic modeling using a chemometric approach. Applied Catalysis B: Environmental. 2007;**77**(1):1-11

[17] Cestari AR, Vieira EF, Mota JA. The removal of an anionic red dye from aqueous solutions using chitosan beads—The role of experimental factors on adsorption using a full factorial design. Journal of Hazardous Materials. 2008;**160**(2):337-343

[18] Arabzadeh S, Ghaedi M, Ansari A, Taghizadeh F, Rajabi M. Comparison of nickel oxide and palladium nanoparticle loaded on activated carbon for efficient removal of methylene blue kinetic and isotherm studies of removal process. Human & Experimental Toxicology. 2015;**34**(2):153-169

[19] Jamshidi M, Ghaedi M, Dashtian K, Hajati S, Bazrafshan A. Ultrasound-assisted removal of Al^{3+} ions and alizarin red S by activated carbon engrafted with Ag nanoparticles: Central composite design and genetic algorithm optimization. RSC Advances. 2015;**5**(73):59522-59532

[20] Roosta M, Ghaedi M, Daneshfar A, Sahraei R. Ultrasound assisted microextraction-nano material solid phase dispersion for extraction and determination of thymol and carvacrol in pharmaceutical samples: Experimental design methodology. Journal of Chromatography B. 2015;**975**:34-39

[21] Azad FN, Ghaedi M, Dashtian K, Hajati S, Pezeshkpour V. Ultrasonically assisted hydrothermal synthesis of activated carbon–HKUST-1-MOF hybrid for efficient simultaneous ultrasound-assisted removal of ternary organic dyes and antibacterial investigation: Taguchi optimization. Ultrasonics Sonochemistry. 2016;**31**:383-393

[22] Ghaedi M, Negintaji G, Marahel F. Solid phase extraction and removal of brilliant green dye on zinc oxide nanoparticles loaded on activated carbon: New kinetic model and thermodynamic evaluation. Journal of Industrial and Engineering Chemistry. 2014;**20**(4):1444-1452

[23] Ghaedi M, Khafri HZ, Asfaram A, Goudarzi A. Response surface methodology approach for optimization of adsorption of Janus green B from aqueous solution onto ZnO/Zn(OH)$_2$-NP-AC: Kinetic and isotherm study. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. 2016;**152**:233-240

[24] Azad FN, Ghaedi M, Dashtian K, Hajati S, Goudarzi A, Jamshidi M. Enhanced simultaneous removal of malachite green and safranin O by ZnO nanorod-loaded activated carbon: Modeling, optimization and adsorption isotherms. New Journal of Chemistry. 2015;**39**(10):7998-8005

[25] Khataee AR, Zarei M, Moradkhannejhad L. Application of response surface methodology for optimization of azo dye removal by oxalate catalyzed photoelectro-Fenton process using carbon nanotube-PTFE cathode. Desalination. 2010;**258**(1):112-119

[26] Khanmohammadi M, Azghandi MA. Application of Doolittle algorithm as a multivariate calibration method for infrared spectrometric determination of some ingredients in detergent washing powder. Analytical Chemistry Letters. 2011;**1**(3):202-206

[27] Khanmohammadi M, Azqhandi MA. Introducing an orthogonal-triangular decomposition algorithm and its application in multivariate calibration. Analytical Methods. 2011;**3**(12):2721-2725

[28] Daneshi A, Younesi H, Ghasempouri SM, Sharifzadeh M. Production of poly-3-hydroxy-butyrate by *Cupriavidus necator* from corn syrup: Statistical modeling and optimization of biomass yield and volumetric productivity. Journal of Chemical Technology and Biotechnology. 2010;**85**(11):1528-1539

[29] Annadurai G, Sheeja R. Use of Box-Behnken design of experiments for the adsorption of verofix red using biopolymer. Bioprocess and Biosystems Engineering. 1998;**18**(6):463-466

[30] Ghaedi M, Azqhandi MHA, Asfaram A. Application of machine/statistical learning, artificial intelligence and statistical experimental design for modeling and optimization of methylene blue and Cd (II) removal from binary aqueous solution by natural walnut carbon. Physical Chemistry Chemical Physics. 2017;**19**:11299-11317

[31] Azqhandi MA, Ghaedi M, Yousefi F, Jamshidi M. Application of random forest, radial basis function neural networks and central composite design for modeling and/or optimization of the ultrasonic assisted adsorption of brilliant green on ZnS-NP-AC. Journal of Colloid and Interface Science. 2017;**505**:278-292

[32] Porhemmat S, Rezvani AR, Ghaedi M, Azqhandi MHA, Bazrafshan AA. Nanocomposites: Synthesis, characterization and its application to removal azo dyes using ultrasonic assisted method: Modeling and optimization. Ultrasonics Sonochemistry. 2017;**38**:530-543

[33] Dastkhoon M, Ghaedi M, Asfaram A, Azqhandi MHA, Purkait MK. Simultaneous removal of dyes onto nanowires adsorbent use of ultrasound assisted adsorption to clean waste water: Chemometrics for modeling and optimization, multicomponent adsorption and kinetic study. Chemical Engineering Research and Design. 2017;**124**:222-237

[34] Asfaram A, Ghaedi M, Azqhandi MA, Goudarzi A, Dastkhoon M. Statistical experimental design, least squares-support vector machine (LS-SVM) and artificial neural network (ANN) methods for modeling the facilitated adsorption of methylene blue dye. RSC Advances. 2016;**6**(46):40502-40516

[35] Asfaram A, Ghaedi M, Azqhandi MHA, Goudarzi A, Hajati S. Ultrasound-assisted binary adsorption of dyes onto Mn@ CuS/ZnS-NC-AC as a novel adsorbent: Application of chemometrics for optimization and modeling. Journal of Industrial and Engineering Chemistry. 2017;**54**:377-388

[36] Aleboyeh A, Daneshvar N, Kasiri M. Optimization of CI acid red 14 azo dye removal by electrocoagulation batch process with response surface methodology. Chemical Engineering and Processing: Process Intensification. 2008;**47**(5):827-832

[37] Moghaddam SS, Moghaddam MA, Arami M. Coagulation/flocculation process for dye removal using sludge from water treatment plant: Optimization through response surface methodology. Journal of Hazardous Materials. 2010;**175**(1):651-657

[38] Cho I-H, Zoh K-D. Photocatalytic degradation of azo dye (reactive red 120) in TiO_2/UV system: Optimization and modeling using a response surface methodology (RSM) based on the central composite design. Dyes and Pigments. 2007;**75**(3):533-543

[39] Ravikumar K, Krishnan S, Ramalingam S, Balu K. Optimization of process variables by the application of response surface methodology for dye removal using a novel adsorbent. Dyes and Pigments. 2007;**72**(1):66-74

[40] Ravikumar K, Ramalingam S, Krishnan S, Balu K. Application of response surface methodology to optimize the process variables for reactive red and acid brown dye removal using a novel adsorbent. Dyes and Pigments. 2006;**70**(1):18-26

[41] Sharma S, Malik A, Satya S. Application of response surface methodology (RSM) for optimization of nutrient supplementation for Cr (VI) removal by Aspergillus lentulus AML05. Journal of Hazardous Materials. 2009;**164**(2):1198-1204

[42] Gönen F, Aksu Z. Single and binary dye and heavy metal bioaccumulation properties of Candida tropicalis: Use of response surface methodology (RSM) for the estimation of removal yields. Journal of Hazardous Materials. 2009;**172**(2):1512-1519

[43] Khataee A, Fathinia M, Aber S, Zarei M. Optimization of photocatalytic treatment of dye solution on supported TiO_2 nanoparticles by central composite design: Intermediates identification. Journal of Hazardous Materials. 2010;**181**(1):886-897

[44] Daneshvar N, Salari D, Khataee A. Photocatalytic degradation of azo dye acid red 14 in water on ZnO as an alternative catalyst to TiO_2. Journal of Photochemistry and Photobiology A: Chemistry. 2004;**162**(2):317-322

[45] Chatterjee S, Kumar A, Basu S, Dutta S. Application of response surface methodology for methylene blue dye removal from aqueous solution using low cost adsorbent. Chemical Engineering Journal. 2012;**181**:289-299

[46] Körbahti BK. Response surface optimization of electrochemical treatment of textile dye wastewater. Journal of Hazardous Materials. 2007;**145**(1):277-286

[47] Zuorro A, Fidaleo M, Lavecchia R. Response surface methodology (RSM) analysis of photodegradation of sulfonated diazo dye reactive green 19 by UV/H_2O_2 process. Journal of Environmental Management. 2013;**127**:28-35

[48] Sadeghi-Kiakhani M, Arami M, Gharanjig K. Preparation of chitosan-ethyl acrylate as a biopolymer adsorbent for basic dyes removal from colored solutions. Journal of Environmental Chemical Engineering. 2013;**1**(3):406-415

[49] Daneshvar N, Khataee A, Djafarzadeh N. The use of artificial neural networks (ANN) for modeling of decolorization of textile dye solution containing CI basic yellow 28 by electrocoagulation process. Journal of Hazardous Materials. 2006;**137**(3):1788-1795

[50] Elmolla ES, Chaudhuri M, Eltoukhy MM. The use of artificial neural network (ANN) for modeling of COD removal from antibiotic aqueous solution by the Fenton process. Journal of Hazardous Materials. 2010;**179**(1):127-134

[51] Balci B, Keskinkan O, Avci M. Use of BDST and an ANN model for prediction of dye adsorption efficiency of Eucalyptus camaldulensis barks in fixed-bed system. Expert Systems with Applications. 2011;**38**(1):949-956

[52] Pooralhossini J, Zanjanchi MA, Ghaedi M, Asfaram A, Azqhandi MHA. Statistical optimization and modeling approach for azo dye decolorization: Combined effects of ultrasound waves and nanomaterial-based adsorbent. Applied Organometallic Chemistry. 2018;**32**(3):1-14

[53] Askari H, Ghaedi M, Dashtian K, Azghandi MHA. Rapid and high-capacity ultrasonic assisted adsorption of ternary toxic anionic dyes onto MOF-5-activated carbon: Artificial neural networks, partial least squares, desirability function and isotherm and kinetic study. Ultrasonics Sonochemistry. 2017;**37**:71-82

[54] Yovel Y, Franz MO, Stilz P, Schnitzler HU. Plant classification from bat-like echolocation signals. PLoS Computational Biology. 2008;**4**(3):e1000032

[55] Liang L, Wang Q, Chen Y. In application of support vector machine in online monitoring of wastewater treatment based on combined kernel functions. In: 2011 International Conference on Electrical and Control Engineering (ICECE); IEEE. 2011. pp. 3840-3843

[56] Mahmoodi NM, Hosseinabadi-Farahani Z, Chamani H. Nanostructured adsorbent (MnO_2): Synthesis and least square support vector machine modeling of dye removal. Desalination and Water Treatment. 2016;**57**(45):21524-21533

[57] Ghaedi M, Ghaedi A, Hossainpour M, Ansari A, Habibi M, Asghari A. Least square-support vector (LS-SVM) method for modeling of methylene blue dye adsorption using copper oxide loaded on activated carbon: Kinetic and isotherm study. Journal of Industrial and Engineering Chemistry. 2014;**20**(4):1641-1649

[58] Foucquier A, Robert S, Suard F, Stéphan L, Jay A. State of the art in building modelling and energy performances prediction: A review. Renewable and Sustainable Energy Reviews. 2013;**23**:272-288

[59] Azad FN, Ghaedi M, Asfaram A, Jamshidi A, Hassani G, Goudarzi A, Azqhandi MHA, Ghaedi A. Optimization of the process parameters for the adsorption of ternary dyes by Ni doped FeO (OH)-NWs–AC using response surface methodology and an artificial neural network. RSC Advances. 2016;**6**(24):19768-19779

[60] Sayegh A, Tate JE, Ropkins K. Understanding how roadside concentrations of NOx are influenced by the background levels, traffic density, and meteorological conditions using boosted regression trees. Atmospheric Environment. 2016;**127**:163-175

[61] Zhang W, Du Z, Zhang D, Yu S, Hao Y. Boosted regression tree model-based assessment of the impacts of meteorological drivers of hand, foot and mouth disease in Guangdong, China. Science of the Total Environment. 2016;**553**:366-371

[62] Ghasemi JB, Tavakoli H. Application of random forest regression to spectral multivariate calibration. Analytical Methods. 2013;**5**(7):1863-1871

[63] Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, Hamprecht FA. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics. 2009;**10**(1):213

[64] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: A classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information and Computer Sciences. 2003;**43**(6):1947-1958

[65] Polishchuk PG, Muratov EN, Artemenko AG, Kolumbin OG, Muratov NN, Kuz'min VE. Application of random forest approach to QSAR prediction of aquatic toxicity. Journal of Chemical Information and Modeling. 2009;**49**(11):2481-2488

[66] Brereton RG, Lloyd GR. Support vector machines for classification and regression. Analyst. 2010;**135**(2):230-267

[67] Ridgeway G. Package gbm: Generalized Boosted Regression Models. Version 1.6. 3.2. Austria: R Foundation for Statistical Computing Vienna; 2012

[68] De'Ath G. Boosted trees for ecological modeling and prediction. Ecology. 2007;**88**(1):243-251

[69] Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics. 2001:1189-1232

[70] Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: Data mining, inference and prediction. The Mathematical Intelligencer. 2005;**27**(2):83-85

[71] Breiman L. Arcing classifier (with discussion and a rejoinder by the author). The Annals of Statistics. 1998;**26**(3):801-849

[72] Breiman L. Random forests. Machine Learning. 2001;**45**(1):5-32

[73] Mazaheri H, Ghaedi M, Azqhandi MA, Asfaram A. Application of machine/statistical learning, artificial intelligence and statistical experimental design for the modeling and optimization of methylene blue and Cd (ii) removal from a binary aqueous solution by natural walnut carbon. Physical Chemistry Chemical Physics. 2017;**19**(18):11299-11317

[74] Tanzifi M, Yaraki MT, Kiadehi AD, Hosseini SH, Olazar M, Bhati AK, Agarwal S, Gupta VK, Kazemi A. Adsorption of Amido black 10B from aqueous solution using polyaniline/SiO_2 nanocomposite: Experimental investigation and artificial neural network modeling. Journal of Colloid and Interface Science. 2018;**510**:246-261

Application of AI in Chemical Engineering

Zeinab Hajjar, Shokoufe Tayyebi and
Mohammad Hosein Eghbal Ahmadi

Additional information is available at the end of the chapter

Abstract

A major shortcoming of traditional strategies is the fact that solving chemical engineering problems due to the highly nonlinear behavior of chemical processes is often impossible or very difficult. Today, artificial intelligence (AI) techniques are becoming useful due to simple implementation, easy designing, generality, robustness and flexibility. The AI includes various branches, namely, artificial neural network, fuzzy logic, genetic algorithm, expert systems and hybrid systems. They have been widely used in various applications of the chemical engineering field including modeling, process control, classification, fault detection and diagnosis. In this chapter, the capabilities of AI are investigated in various chemical engineering fields.

Keywords: chemical engineering, AI algorithms, classification, process control, modeling, optimization, fault detection and diagnosis

1. Introduction

Artificial intelligence (AI) applications in chemical engineering have increased dramatically recently. This chapter deals with various applications of artificial intelligence (AI) in the chemical engineering field including process such as modeling, optimization, process control, fault detection and diagnosis. The aim of the chapter is to provide an overview of the field by presenting the capabilities and limitations of using the AI approach, focusing on artificial neural network (ANN) and fuzzy logic methods.

It is shown that complexities of conventional approaches when dealing with chemical processes which are inherently highly nonlinear can be tackled through the application of AI methods. Four illustrative relevant examples are also presented.

After reading this chapter, the reader is expected to have a basic grounding in the application of AI methods in chemical engineering and understand their implementation issues.

## 2. Application of AI in chemical process modeling

Chemical process models which present the system behavior are useful in all phases of chemical engineering, from research and design to optimization and control and even plant operations [1].

Generally, There are two major types of modeling approaches in chemical engineering, namely, mechanistic (white box, first principle) and AI-based approach like ANN and fuzzy logic methods. In the mechanistic approach, fundamental physical and chemical laws, such as conservation laws, construct the basis of the model. This approach contains algebraic and differential equations which involve mass, energy and momentum balances. Due to the large number of variables affecting the process behavior and complex mathematical equations governing the system, many chemical processes are nonlinear and complicated. Consequently, it is hard and sometimes even impossible to present them by mechanistic models. Even if such a model has been developed, it might be impractical to solve or identify its parameters. Moreover, a mechanistic model needs detailed knowledge and a lot of skill and ingenuity to incorporate the basic phenomena of the process in the model. Difficulties can arise from poor knowledge [2]. In some cases, considering some assumptions such as physical properties' constancy, ideality of gas phase and linearization of the nonlinear equations of the model is inevitable, which all impose limitations on the model leading to the reduction of the model's robustness [3].

On the contrary, AI-based techniques have demonstrated their superb ability and have received much attention for chemical process modeling. These techniques, for which developing detailed knowledge of the process is of less concern, may overcome the drawbacks of the mechanistic approach when dealing with complex and nonlinear systems. Using AI-based methods, inherently qualitative variables in chemical processes like catalyst deactivation in reactors can also be considered in the model, while these types of variables are not possible to implement in mechanistic models.

The most common methods of AI for modeling purposes in chemical engineering are ANN and fuzzy logic, which sometimes are hybridized with evolutionary algorithms [4–7]. In addition to ANN and fuzzy logic methods, their hybrid scheme named adaptive-network-based fuzzy inference system (ANFIS) which is actually a fuzzy inference system implemented in the framework of adaptive networks has also been applied for modeling purposes in chemical engineering.

The first step of developing an AI-based model is defining the input/output variables of the system which is to be modeled (**Figure 1**). Afterward, according to the experimental data or the knowledge of the governing phenomena, the model is developed. The parameters that characterize the AI-based model like the number of fuzzy sets (when using fuzzy logic), the number and the transfer functions of hidden layers (when using the ANN method) depend on the complexity and nonlinearity of the system and the types of variables affecting the process.
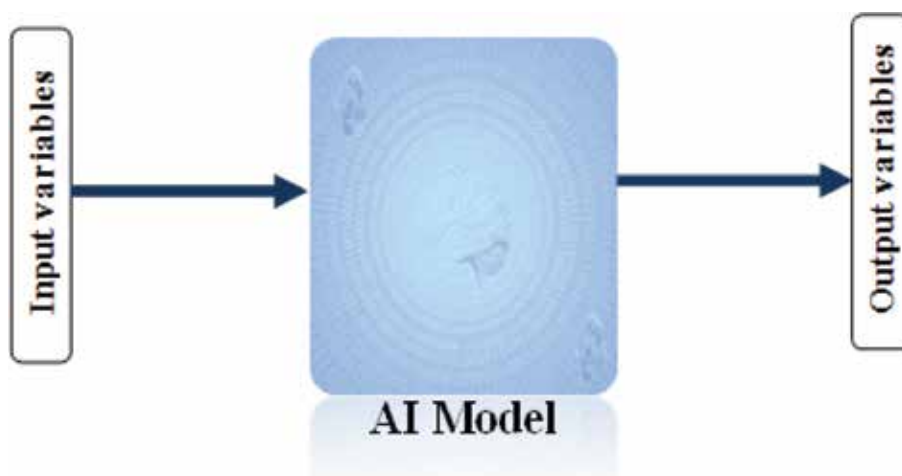
Figure 1. Schematic of AI model.

Among the types of ANN structures, the multi-layer perceptron (MLP) neural network which has a feed-forward scheme is believed as the most useful topology for system modeling [8]. Moreover, the recurrent ANN model which is a mapping of past inputs and outputs to the future outputs can be used for dynamic processes.

In the fuzzy model approach where the two types of which are Mamdani [9] and Takagi-Sugeno (TS) [10], all the uncertainties and model complications are treated in linguistic expressions in the form of "If-Then" rules based on the theory of fuzzy logic [11]. TS like ANN is commonly considered as a data-driven AI-based modeling approach [12].

Mamdani Fuzzy that differs in the way the information and rules are presented has several superiorities over the TS approach for the modeling of chemical processes. First, qualitative experience and knowledge of the experts who are dealing with the process are incorporated in the development of the model [12]. In addition, there is no need for data in order to build the Mamdani fuzzy model. Consequently, a Mamdani fuzzy model is more intuitive, transparent and interpretable [13]. In contrast, each TS-type model is a local approximator and the predictability of the model is valid for the specific operating condition of the process under which the model was developed and tested [14]. Accordingly, it can hardly be applied for analyzing the process behavior and cannot be scaled up or down and therefore is less useful for industrial practice. Despite the capabilities of Mamdani method, it is worth underlying that a Mamdani fuzzy model suffers from the large number of rules when dealing with the processes with large number of variables.

Genetic algorithm (GA) can be used to optimize the performance of a fuzzy model. The role of GA is recognized as optimal parameters' estimation such as the parameters of scaling functions and the universes of discourse [15, 16] or the membership functions (MFs) [17, 18]. GA is also applied as a method for rule reduction/selection by removing some rules like redundant, unnecessary or misleading ones [17] when dealing with high-dimensional problems in which the number of rules is so large that it cannot be managed efficiently.
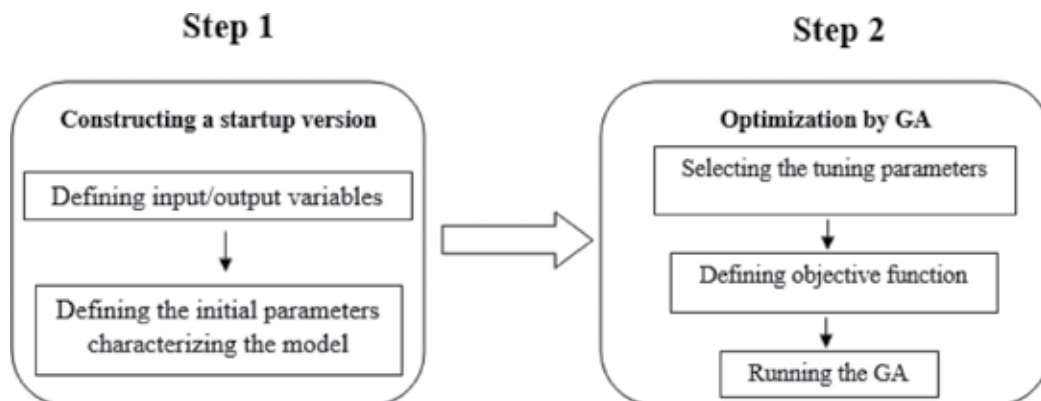
Figure 2. Schematic of hybrid Mamdani fuzzy and GA modeling method.

The hybrid Mamdani fuzzy and GA modeling methods commonly consist of two main steps: (i) constructing a start-up version of the model using only the heuristic knowledge and (ii) tuning procedure using the GA. The schematic of this algorithm is shown in **Figure 2**. In the first step, the output variables determining the behavior of the system are defined, given that, the input variables which affect the selected output variables are determined. Afterward, a base fuzzy model is defined, characterized by the number and types of fuzzy sets of variables and the production rules presenting the behavior of the process based on the knowledge and expertise of the experts who have been working with the system. This model is used as the start-up version of the model which has to be tuned. In the second step, GA is formulated for optimization of parameters that characterize model, such as membership function parameters, membership function types and so on.

3. Application of AI in optimization of chemical processes

Chemical process optimization has its origins in linear programming at the beginning of the 1960s [19]. This problem is finding the best solution from a variety of efficient alternatives of design or operating variables in order to minimize or maximize a desired objective function. In a general way, the objective function can be the minimization of the operating costs and the undesired material production or the maximization of energy efficiency, the yields and operation productivity, the profitability, safety and the reliability of the plant.

Most chemical processes are nonlinear and complex, so there are many solutions (in some cases becoming endless) in the optimization problems. Such problems are often too complex to be solved through gradient-based optimization approaches. Evolutionary algorithms (EAs) like GA [20], harmony search [21], particle swarm optimization [22] and so on categorized in the AI-based method that is a generic population-based metaheuristic optimization algorithm are capable of efficiently finding an optimal solution in complex problems, such as optimization of chemical processes.

4. Application of neural networks in chemical process control

The process control strategies have been developed to improve the performance of the process, reduce energy consumption and ensure high safety and environmental goals. The conventional controllers cannot show satisfactory responses in many industrial chemical processes with high nonlinear dynamics and parameter uncertainties, whereas AI approaches can be effectively controlled for a number of complex and nonlinear processes [23].

Because of their high potential for handling nonlinear relationships and self-learning capabilities, there has been considerable interest in the use of neural networks for the control in different fields of chemical processes such as thermal processes [24], reaction processes [25] and separation and purification [26, 27].

One of the algorithms based on neural network control is the inverse model control. In this approach, it is assumed that the input vector for neural network is the required future or reference output together with the past inputs and the past output variables; the approach can help to make for better performance the controlled variables when the unmeasured disturbance is present. The manipulated variable of the controlled plant is the output of the neural network controller [23]. In the system with time delay (τ), if the orders of dynamic model for the output and the input are n and m, the inverse model can be expressed as the function of the input and output as shown below:

$$M(k) = \varnothing\left(y_{sp}, y(k-1), \ldots, y(k-n), M(k-1-\tau), \ldots, M(k-m-\tau)\right) \tag{1}$$

where $\varnothing$ represents the function of the inverse model, k is the discrete time and M, y and y_{sp} are the control action, the output of the plant and the set point controller, respectively. Therefore, the controller predicts the control action, as shown in **Figure 3**, by having current and past values of the process model state variables and the past control action.

Fuzzy systems have been used in different applications for controlling chemical processes [28–30]. Researchers also used fuzzy logic controller coupled with an optimal control in an exothermic chemical reaction [31], a batch polymerization reactor [32] and polymerization processes [33]. In addition, since time delay can be often seen in many industrial chemical processes, a possible alternative is the fuzzy model predictive control (FMPC) which has been proposed [34, 35]. In systems with uncertainties of the system model, the choice of type-1 fuzzy may not always be the appropriate solution for a control problem [36]. In these cases, the type-2 fuzzy logic control has been represented in many fields of chemical processes [37, 38].

Hybrid controller based on AI strategies combines two or more AI techniques in order to improve control performance of the chemical process. One of the most popular strategies is the adaptive neuro-fuzzy inference system (ANFIS) controller. This approach is a hybrid intelligent system which uses the learning ability of the neural network with the knowledge representation of the fuzzy logic [39]. The schematic of ANFIS model with two inputs (x_1 and x_2) and one output (ϕ) is shown in **Figure 4**. As shown in **Figure 4**, the ANFIS architecture contains five layers of feed-forward neural network which are explained as follows:
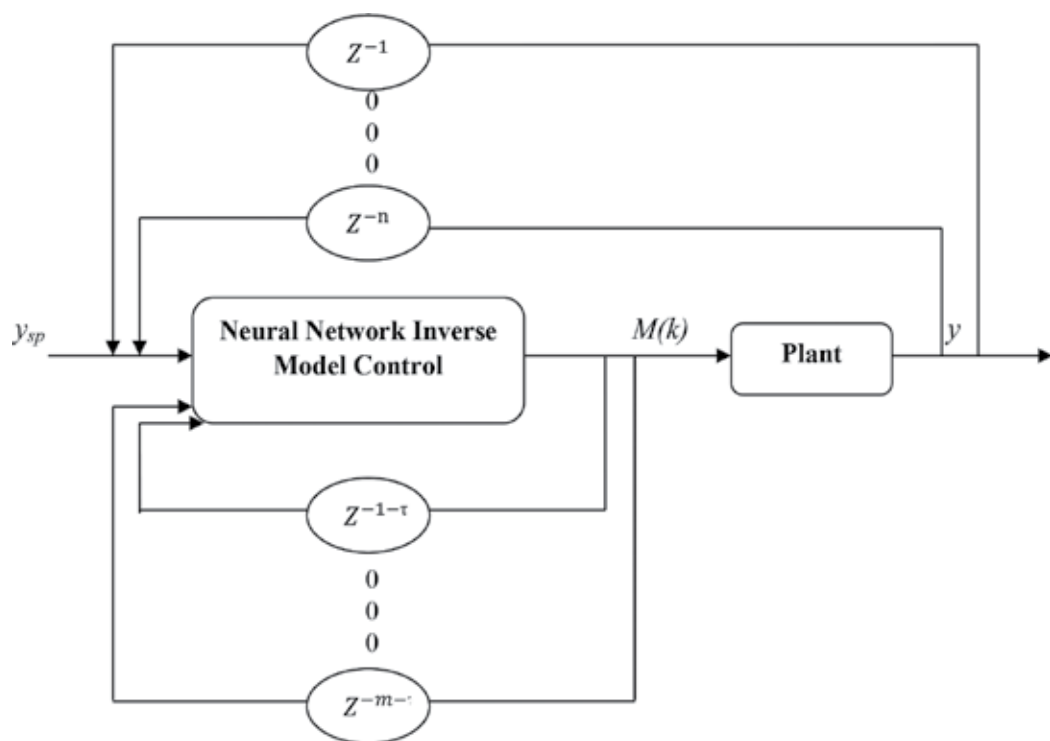
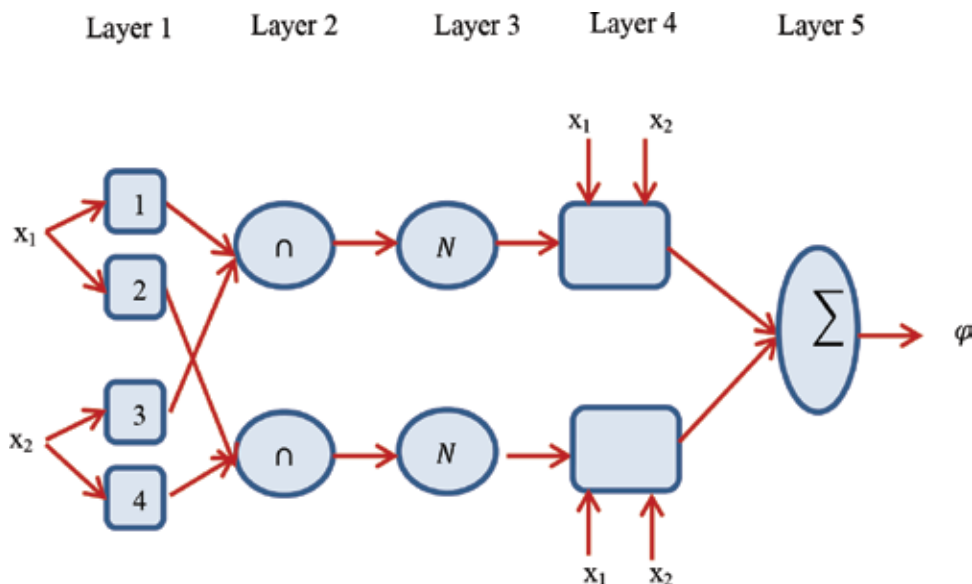**Figure 3.** Schematic of neural network inverse model control.

Figure 4. Schematic architecture of ANFIS model with two fuzzy rules for two inputs and one output.

4.1. First layer

This layer is named as an input layer. Each neuron in this layer saves the parameters of the membership function and crisp inputs are converted to membership degree values which change between 0 and 1.

4.2. Second layer

Each neuron of this layer performs a connective operation (i.e., "AND") to calculate the firing strength of a rule.

4.3. Third layer

A normalization process is performed by the neurons of this layer.

4.4. Fourth layer

The normalized firing strength is multiplied by a linear combination of the inputs (i.e., Takagi-Sugeno fuzzy rule) in order to obtain a rules layer.

4.5. Fifth layer

The last layer of the network is the weighted average of the outputs of the fourth layer.

The application of ANFIS in the process control of chemical plants was seen in the distillation column [40], biodiesel reactor [41].

5. Application of AI techniques in fault detection and diagnosis of chemical engineering

A fault is defined as a deviation from an acceptable range of an observable variable or calculated parameter that is referred to as a failure. A failure can be described as diversity of malfunctions in the real plant which can be caused due to instruments' failures, disturbances and plant parameter uncertainties. The abnormal conditions in a plant can result in financial losses. Therefore, in the chemical processes, fault detection and diagnosis have been the focal point of many researches and various fault detection and diagnosis strategies have been presented in the literature. The fault diagnostic systems should possess desirable characteristics such as quick detection, isolability, robustness and multiple fault identifiability [42].

One of the intelligent fault diagnosing techniques is neural network systems. Because of their high potential for capturing nonlinear relationships, neural networks represent a powerful tool for fault diagnosis [43–47]. In fault detection based on neural networks, the number of neurons in the input and output layers are equal to the number of measured variables and the number of potential faults in the process, respectively. The outputs of the neural diagnoser

are binary variables representing the occurrence of a fault (if the corresponding value is 1) or the lack of fault occurrence (if the corresponding value is 0) [47].Another fault diagnosis approach of AI techniques is fuzzy logic, which is applied in chemical processes [48–51]. In fault diagnosis based on fuzzy logic, the fuzzy relations between faults and symptoms are assumed to be from one to many (i.e., one fault may cause several symptoms). For example:

$$\text{If } Sym_1 \text{ is } S_{1,n} \text{ AND } Sym_2 \text{ is } S_{2,n} \text{ AND } \ldots \text{ AND } Sym_m \text{ is } S_{m,n} \text{ Then } F_1 \text{ is } H_n$$

where Sym_i ($i = 1\ldots m$) is the vector of fuzzy input variables (symptoms) and F_j ($j = 1 \ldots n$) is the fuzzy output variables (Faults). $S_{i,j}$ is the input linguistic value relevant to j^{th} output and H_j is the output linguistic value. The schematic of a fuzzy fault detection system is shown in **Figure 5** [51].

Although the neural network is a powerful tool for fault diagnosis due to its ability in capturing the nonlinear relationship with no heuristic reasoning about process, it requires a large number of data corresponding to various operating conditions in which the effects of various faults exist. On the other hand, the fuzzy diagnoser system expresses the heuristic knowledge between the symptoms and their corresponding faults of the process such as linguistic rules and does not require any quantitative data sets corresponding to history and trends of the system under any operating conditions [52, 53]. The disadvantage of the fuzzy diagnoser is that managing heuristic and knowledge-based rules is more difficult and time demanding and sometimes even impossible for plant-wide integrated processes [54, 55]. Therefore, neuro-fuzzy diagnoser applications in chemical plants are proposed in the literature [56–58].

In the following, four illustration studies of AI techniques are presented for various purposes (i.e. modeling, optimization, process control, fault detection and diagnosis) of chemical processes.
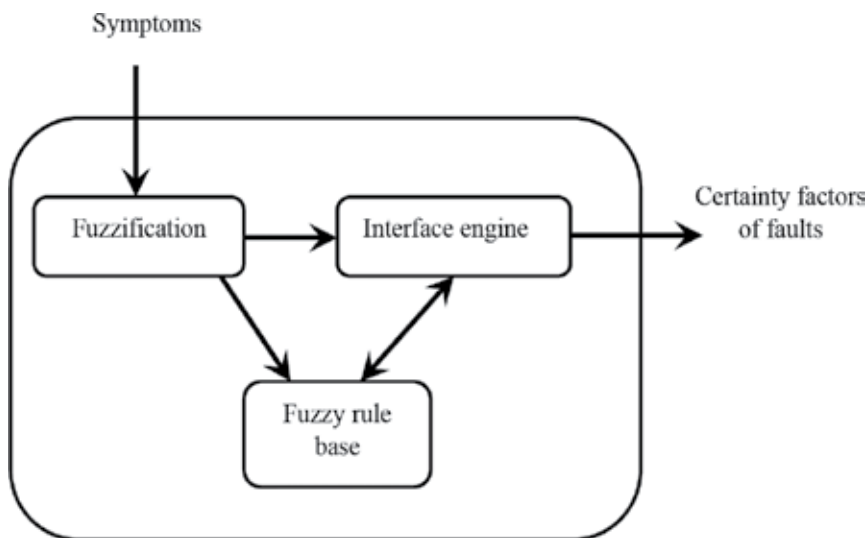
Figure 5. Schematic diagram of a fuzzy fault detection scheme.

Illustration case study one: Prediction of virus removal from water using microfiltration membrane based on hybrid Mamdani fuzzy and GA.

One of the separation technologies for virus removal from water for municipal effluent reuse is the application of membranes. Conventional modeling approaches for predicting membrane performance suffer from various limitations such as the lack of predictive fouling models or complexities of estimating the properties of the membrane surface and membrane interactions. In this case study, an optimum Mamdani fuzzy model is developed for removal prediction of two types of viruses from water [2]. The GA is employed for optimal estimation of parameters characterizing the membership functions of the input/output variables of the model. The first step is defining input/output variables of the model. The amount of virus rejection (R%) which is considered as an output variable of the model is determined as follows:

$$R\% = 100\left(1 - \frac{C_p}{C_f}\right) \tag{2}$$

where C_p and C_f are virus concentrations in permeate and feed, respectively. The input variables of the model are concentration of FMD virus (CFMD), concentration of IBR virus (CIBR), operating pressure (P), volume (V) and rpm (stirring speed). The experimental data can found in the works of Madaeni and Kurdian [2]. All the variables that exist in the system are discretized by Gaussian-type membership functions.

The next steps are setting a fuzzy inference system using initial fuzzy sets with parameters and defining the fitness function. Two parameters of the Gaussian membership function including $\bar{x}$ and σ (Eq. (3)) are obtained via GA.

$$f(x) = \exp\left(-\left(\frac{x - \bar{x}}{\sigma}\right)^2\right) \tag{3}$$

The mean square error (MSE) is selected as the fitness function as follows:

$$MSE = (y_m - y_e)^2 \tag{4}$$

where y_m and y_e denote the vector of fuzzy model and data set, respectively.

There are two methods for the genetically tuning procedure. For cases with a low number of rules, the parameters of membership functions in all rules (for both input and output variables) can be considered as decision variables through optimization formulation. In this way, a variable at each rule can have different optimized shape membership functions. This approach increases the prediction capability yet at the cost of reducing the interpretability of the model. The second method is used when there are a large number of rules in the model. In this method, each variable of the model in all rules has the same optimized shape of the membership function. This method has a lower number of decision variables in optimization formulation for the same case when compared with the first method.

According to the possible combination of input variables, 10 rules can be defined for this model, and due to the low number of rules, the first-mentioned method for parameter optimization is applied. The population size and maximum generation number of GA are set as 500 and 100, respectively. Having passed the optimization procedure, the fuzzy model is developed with optimized parameters. This model is developed based on qualitative rules, bypassing the complexities and drawbacks of the white-box modeling method.

The comparison between model data and experimental data shows an accuracy of near 90% for the developed fuzzy model [2].

Illustrative case study two: Optimization of fluidized bed reactor of oxidative coupling of methane based on GA.

In this case study, optimization of C_2 (ethane + ethylene) yield in the oxidative coupling of methane (OCM) over the $Mn/Na_2WO_4/SiO_2$ catalyst in a fluidized bed reactor is carried out [20].

OCM is a series of chemical reactions, first presented in the 1980s by Keller and Bhasin [59] for the direct conversion of natural gas into the desired product of ethylene and other value-added chemicals. One of the barriers to the commercialization of this process is the low yield of the reactions. Various solutions have been proposed for yield improvement in the literature [60, 61]. One possibility to improve the C_2 yield is stage-wise feeding configuration along the reactor as shown in **Figure 6** [62].

In this scheme, it is assumed that the injected gas in each stage has only oxygen in content and all the methane is introduced to the reactor at the entrance of the bed.
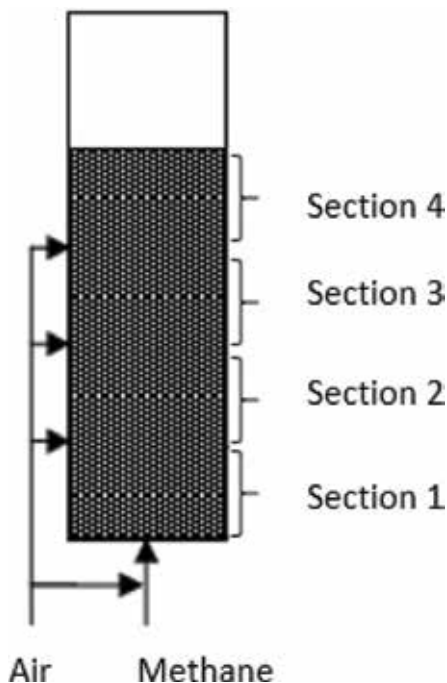
Figure 6. Stage-wise feeding configuration in OCM reactor.

In this case study, the main process variables which are optimized to enhance C_2 yield are listed in **Table 1**.

The kinetic model presented by Daneshpayeh et al. [63] is used as the reaction sub-model. The reactor is firstly modeled and then solved [62]. Afterward, using GA, the C_2 yield is optimized for one, two and three secondary injections of oxygen.

C_2 yield considered as the fitness function is defined as follows:

$$Y_{C_2} = \frac{2 \times N_{C_2}}{N_{CH_4}} \times 100 \qquad (5)$$

The main GA parameters are presented in **Table 2**.

The best results are achieved for three injections of oxygen along the reactor. The optimum values of decision variables are presented in **Table 3**. The maximum C_2 yield of 22.87% is achieved for three secondary oxygen injections at the operating temperature of 746.05°C. This optimized C_2 yield which is achieved by an AI-based method is approximately 4% higher than the original model [61].

Illustrative case study three: Online genetic-ANFIS control for advanced microwave biodiesel reactor.

The microchem reactor based on the microwave process technology is used to produce biodiesel which is good for the environment. The reactor temperature should be controlled to ensure an optimal yield of biodiesel and to minimize the generation of unwanted byproducts. For this aim, Wali et al. implemented an artificial intelligent controller design based on the online genetic-ANFIS temperature control for advanced biodiesel microwave reactor [41]. The microwave power supply as the manipulated variable, the reactor temperature as control variable and the feed-flow rate as the disturbance variable have been considered in this process.

The online genetic-ANFIS controller has been evaluated at different operation conditions (set-point tracking and disturbance rejection). The genetic-ANFIS controller successfully tracks the demands of reactor temperature set-point faster than adaptive control without any oscillations [41].

Illustrative case study four: Neuromorphic multiple-fault diagnosing the plant-wide system.

Decision variables	Constraints
Methane flow rate (U_c)	0.03–0.5 m³/s
Oxygen flow rate at each injection part	0.03–0.5 m³/s (×0.21)[*]
Operating temperature (T)	700–850°C
Length of each section of reactor (L_i)	0.5–4 m

*Note that actually it is the air which entered the bed.

Table 1. Decision variables and their constraints.

Parameter	Value
Population size	50
Generations	10,000
Survival probability	0.5
Linear crossover probability	0.5
Mutation probability	0.167

Table 2. GA parameters.

Decision variable	Optimum value
Oxygen flow at the beginning of the reactor (m^3/s)	0.0833
Oxygen flow at second section of the reactor (m^3/s)	0.0500
Oxygen flow at third section of the reactor (m^3/s)	0.0536
Oxygen flow at fourth section of the reactor (m^3/s)	0.0755
Methane flow at the beginning of the reactor (m^3/s)	0.3998
Length of first section (m)	0.6345
Length of second section (m)	3.9389
Length of third section (m)	0.5882
Length of fourth section (m)	1.0116
Temperature (°C)	746.05
Yield	22.87

Table 3. Optimum values of decision variables at maximum yield.

In plant-wide systems, because of system complexity and overlapping symptoms, conventional neural networks operating based on steady-state characteristic data are not usually capable of diagnosing multiple concurrent faults. To overcome this problem, Tayyebi et al. proposed a new neuromorphic diagnosis framework based on augmented input containing steady-state characteristic data along with newly defined dynamic characteristic data [47]. In this approach, the input vector of the neural network diagnosis has been selected in such a way that different faults cause distinctive symptoms. Therefore, information related to both the history of the process and the steady state has been utilized to achieve distinctive symptoms. Accordingly, one can use characteristic points of the dynamic trend of each measured variable to uniquely distinguish and detect various faults. To evaluate the proposed approach based on the hybrid parameter, the Tennessee Eastman process (TE) that contains large numbers of measurements and manipulated variables and overlapping faults was used as the plant-wide benchmark. The performance of the neuromorphic diagnoser based on the augmented inputs has been compared with that of the conventional neuromorphic diagnostic system whose inputs are steady-state characteristic data. The comparison showed that the proposed method

outperformed the conventional neuromorphic diagnoser for the detection of multiple concurrent faults. It was also shown that the proposed scheme can correctly diagnose various combinations of six concurrent faults of the TE process (from two to six simultaneous faults). This achievement reflects the major advantage of the proposed approach, which is its ability to perform fault diagnosis in situations where multiple concurrent faults with overlapping symptoms have occurred [47].

6. Conclusion

AI techniques provide tools to tackle complex problems. Challenging and useful applications of AI techniques have been introduced in the chemical engineering processes. Four illustrative case studies are investigated in fields of process modeling, optimization, process control and fault detection and diagnosis. From the description of the various applications, the ability of AI techniques has been revealed in a wide range of fields in chemical processes.

Author details

Zeinab Hajjar*, Shokoufe Tayyebi and Mohammad Hosein Eghbal Ahmadi

*Address all correspondence to: hajjarz1@gmail.com

Research Institute of Petroleum Industry (RIPI), Tehran, Iran

References

[1] Luyben W. Process modeling, simulation and control for chemical engineers. Petroleum Refinery Engineering. 1996;**2**:289-290

[2] Madaeni SS, Kurdian AR. Fuzzy modeling and hybrid genetic algorithm optimization of virus removal from water using microfiltration membrane. Chemical Engineering Research and Design. 2011;**89**:456-470

[3] Araromi DO, Sonibare JA, Emuoyibofarhe JO. Fuzzy identification of reactive distillation for acetic acid recovery from waste water. Journal of Environmental Chemical Engineering. 2014;**2**:1394-1403

[4] Hajjar Z, Kazemeini M, Rashidi A, Tayyebi S. Artificial intelligence techniques for modeling and optimization of the HDS process over a new graphene based catalyst. Phosphorus, Sulfur, and Silicon and the Related Elements. 2016;**191**:1256-1261

[5] Soltanali S, Halladj R, Tayyebi S, Rashidi A. Neural network and genetic algorithm for modeling and optimization of effective parameters on synthesized ZSM-5 particle size. Materials Letters. 2014;**136**:138-140

[6] Soltanali S, Halladj R, Tayyebi S, Rashidi A. Application of genetic-fuzzy approach for estimation of nano ZSM-5 crystallinity. Materials Letters. 2015;**150**:39-43

[7] Hajjar Z, Khodadadi A, Mortazavi Y, Tayyebi S, Soltanali S. Artificial intelligence modeling of DME conversion to gasoline and light olefins over modified nano ZSM-5 catalysts. Fuel. 2016;**179**:79-86

[8] Gibbs MS, Morgan N, Maier HR, Dandy GC, Nixon JB, Holmes M. Investigation into the relationship between chlorine decay and water distribution parameters using data driven methods. Mathematical and Computer Modelling. 2006;**44**:485-498

[9] Mamdani E, Assilian S. An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies. 1975;**7**:1-13

[10] Takagi T, Sugeno M. Fuzzy identification of systems and its applications to modeling and control. IEEE Transactions on Systems, Man, and Cybernetics. 1985;**15**:116-132

[11] Zadeh LA. Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems. 1978;**1**:3-28

[12] Adoko AC, Gokceoglu C, Wu L, Zuo QJ. Knowledge-based and data-driven fuzzy modeling for rockburst prediction. International Journal of Rock Mechanics and Mining Sciences. 2013;**61**:86-95

[13] Sala A, Guerra TM, Babuška R. Perspectives of fuzzy systems and control. Fuzzy Sets and Systems. 2005;**156**:432-444

[14] Habbi H, Zelmat M, Bouamama BO. A dynamic fuzzy model for a drum-boiler-turbine system. Automatica. 2003;**39**:1213-1219

[15] Gudwin R, Gomide F, Pedrycz W. Context adaptation in fuzzy processing and genetic algorithms. International Journal of Intelligence Systems. 1998;**13**:929-948

[16] Cordon O, Herrera F, del Jesus MJ, Villar P. A multiobjective genetic algorithm for feature selection and granularity learning in fuzzy-rule based classification systems. In: IFSA World Congr. 20th NAFIPS Int. Conf. 2001. Jt. 9th. 3 2001. pp. 1253-1258

[17] Cordon O, del Jesus MJ, Herrera F. Genetic learning of fuzzy rule-based classification systems cooperating with fuzzy reasoning methods, International Journal of Intelligence Systems. 1998;**13**:1025-1053

[18] Pulkkinen P, Koivisto H. A dynamically constrained multiobjective genetic fuzzy system for regression problems. IEEE Transactions on Fuzzy Systems. 2010;**18**:161-177

[19] Chaves I, López J, Zapata J, Robayo A. Process Analysis and Simulation in Chemical Engineering. 2016

[20] Eghbal-Ahmadi M-H, Zaerpour M, Daneshpayeh M, Mostoufi N. Optimization of fluidized bed reactor of oxidative coupling of methane. International Journal of Chemical Reactor Engineering. 2012;**10**:1-21

[21] Yousefi M, Enayatifar R, Darus AN, Abdullah AH. Optimization of plate-fin heat exchangers by an improved harmony search algorithm. Applied Thermal Engineering. 2013;**50**:877-885

[22] Mahmood HA, Adam N, Sahari BB, Masuri SU. Development of a particle swarm optimisation model for estimating the homogeneity of a mixture inside a newly designed CNG-H2-AIR mixer for a dual fuel engine: An experimental and theoretic study. Fuel. 2017;**218**:131-150

[23] Tayyebi S, Alishiri M. The control of MSF desalination plants based on inverse model control by neural network. Desalination. 2014;**333**:92-100

[24] Nowak G, Rusin A. Using the artificial neural network to control the steam turbine heating process. Applied Thermal Engineering. 2016;**108**:204-210

[25] Li S, Li Y. Neural network based nonlinear model predictive control for an intensified continuous reactor. Chemical Engineering and Processing. 2015;**96**:14-27

[26] Fernandez de Canete J, Gonzalez S, del Saz-Orozco P, Garcia I. A harmonic balance approach to robust neural control of MIMO nonlinear processes applied to a distillation column, Journal of Process Control. 2010;**20**:1270-1277

[27] Damour C, Benne M, Grondin-Perez B, Chabriat J. Nonlinear predictive control based on artificial neural network model for industrial crystallization. Journal of Food Engineering. 2010;**99**:225-231

[28] Hojjati H, Sheikhzadeh M, Rohani S. Control of supersaturation in a semibatch antisolvent crystallization process using a fuzzy logic controller. Industrial and Engineering Chemistry Research. 2007;**46**:1232-1240

[29] Underwood CP. Fuzzy multivariable control of domestic heat pumps. Applied Thermal Engineering. 2015;**90**:957-969

[30] Baroud Z, Benmiloud M, Benalia A, Ocampo-Martinez C. Novel hybrid fuzzy-PID control scheme for air supply in PEM fuel-cell-based systems. International Journal of Hydrogen Energy. 2017;**42**:10435-10447

[31] Karr CL, Sharma SK, Hatcher WJ, Harper TR. Fuzzy control of an exothermic chemical reaction using genetic algorithms. Engineering Applications of Artificial Intelligence. 1993;**6**:575-582

[32] Etinkaya SC, Zeybek Z, Hapoglu H, Alpbaz M. Optimal temperature control in a batch polymerization reactor using fuzzy-relational models-dynamics matrix control. Computers and Chemical Engineering. 2006;**30**:1315-1323

[33] Lima NMN, Linan LZ, Filho RM, Wolf Maciel MR, Embiruc M, Grácio u, F. Modeling and predictive control using fuzzy logic: Application for a polymerization system. AIChE Journal. 2010;**56**:965-978

[34] Chang X-H, Yang G-H. Fuzzy robust constrained model predictive control for nonlinear systems. Asian Journal of Control. 2011;**13**(6):947-955

[35] Teng L, Wang Y, Cai W, Li H. Robust model predictive control of discrete nonlinear systems with time delays and disturbances via T–S fuzzy approach. Journal of Process Control. 2017;**53**:70-79

[36] Miccio M, Cosenza B. Control of a distillation column by type-2 and type-1 fuzzy logic PID controllers. Journal of Process Control. 2014;**24**:475-484

[37] Galluzzo M, Cosenza B. Nonlinear fuzzy control of fed-batch reactor for the penicillin production. Computers and Chemical Engineering. 2012;**36**:273-281

[38] Galluzzo M, Cosenza B. Control of a non-isothermal continuous stirred tank reactor by a feedback-feed forward structure using type-2 fuzzy logic controllers. Information Sciences. 2011;**181**:3535-3550

[39] Perendeci A, Arslan S, Celebi SS, Tanyolac A. Prediction of effluent quality of an anaerobic treatment plant under unsteady state through ANFIS modeling with on-line input variables. Chemical Engineering Journal. 2008;**145**:78-85

[40] Fernandez de Canete J, Garcia-Cerezo A, Garcia-Moral I, Del Saz P, Ochoa E. Object-oriented approach applied to ANFIS modeling and control of a distillation column. Expert Systems with Applications. 2013;**40**:5648-5660

[41] Wali WA, Al-Shamma AI, Hassan KH, Cullen JD. Online genetic-ANFIS temperature control for advanced microwave biodiesel reactor. Journal of Process Control. 2012;**22**:1256-1272

[42] Venkatasubramanian V, Rengaswamy R, Yin K, Kavuri SN. A review of process fault detection and diagnosis part I: Quantitative model-based methods. Computers and Chemical Engineering. 2003;**27**:293-311

[43] Tayarani-Bathaie SS, Khorasani K. Fault detection and isolation of gas turbine engines using a bank of neural networks. Journal of Process Control. 2015;**36**:22-41

[44] Tan WL, Nor NM, Abu Bakar MZ, Ahmed Z, Sata SA. Optimum parameters for fault detection and diagnosis system of batch reaction using multiple neural networks. Journal of Loss Prevention in the Process Industries. 2012;**25**:138-141

[45] Behbahani RM, Jazayeri-Rad H, Hajmirzaee S. Fault detection and diagnosis in a sour gas absorption column using neural networks. Chemical Engineering and Technology. 2009;**32**(5):840-845

[46] Zhang Z, Zhao J. A deep belief network based fault diagnosis model for complex chemical processes. Computers & Chemical Engineering. 2017;**107**:395-407

[47] Tayyebi S, Boozarjomehry RB, Shahrokhi M. Neuromorphic multiple-fault diagnosing system based on plant dynamic characteristics. Industrial and Engineering Chemistry Research. 2013;**52**:12927-12936

[48] Musulin E, Yélamos I, Puigjaner L. Integration of principal component analysis and fuzzy logic systems for comprehensive process fault detection and diagnosis. Industrial and Engineering Chemistry Research. 2006;**45**:1739-1750

[49] Tarifa EE, Scenna NJ. Fault diagnosis, direct graphs, and fuzzy logic. Computers & Chemical Engineering. 1997;**21**:S649-S654

[50] Tarifa EE, Scenna NJ. Fault diagnosis for MSF dynamic states using a SDG and fuzzy logic, Fault diagnosis for MSF dynamic states using a SDG and fuzzy logic. Desalination. 2004;**166**:93-101

[51] Tayyebi S, Shahrokhi M, Boozarjomehry RB. Fault diagnosis in a yeast fermentation bioreactor by genetic fuzzy system. Iranian journal of chemistry and chemical engineering. 2010;**29**:61-72

[52] Hang J, Zhang J, Cheng M. Application of multi-class fuzzy support vector machine classifier for fault diagnosis of wind turbine. Fuzzy Sets and Systems. 2016;**297**:128-140

[53] Jahromi AT, Er MJ, Li X, Lim BS. Sequential fuzzy clustering based dynamic fuzzy neural network for fault diagnosis and prognosis. Neurocomputing. 2016;**196**:31-41

[54] Vachtsevanos G, Lewis F, Roemer M. Intelligent Fault Diagnosis and Prognosis for Engineering Systems. New York: John Wiley & Sons, Inc., Hoboken, New Jersey; 2006

[55] Dou D, Zhou S. Comparison of four direct classification methods for intelligent fault diagnosis of rotating machinery. Applied Soft Computing. 2016;**46**:459-468

[56] Lau CK, Heng YS, Hussain MA, Mohamad Nor MI. Fault diagnosis of the polypropylene production process (UNIPOL PP) using ANFIS. ISA Transactions. 2010;**49**:559-566

[57] Bonsignore L, Davarifar M, Rabhi A, Tina GM, Elhajjaji A. Neuro-fuzzy fault detection method for photovoltaic systems. Energy Procedia. 2014;**62**:431-441

[58] Shabanian M, Montazeri M. A neuro-fuzzy online fault detection and diagnosis algorithm for nonlinear and dynamic systems. International Journal of Control, Automation and Systems. 2011;**9**:665-670

[59] Keller GE, Bhasin MM. Synthesis of ethylene via oxidative coupling of methane. I. Determination of active catalysts. Journal of Catalysis. 1982;**73**:9-19

[60] Kao YK, Lei L, Lin YS. A comparative simulation study on oxidative coupling of methane in fixed-bed and membrane reactors. Industrial and Engineering Chemistry Research. 1997;**36**:3583-3593

[61] Lu Y, Dixon AG, Moser WR, Ma YH, Balachandran U. Oxygen-permeable dense membrane reactor for the oxidative coupling of methane. Journal of Membrane Science. 2000;**170**:27-34

[62] Daneshpayeh M, Mostoufi N, Khodadadi A, Sotudeh-Gharebagh R, Mortazavi Y. Modeling of stagewise feeding in fluidized bed reactor of oxidative coupling of methane. Energy and Fuels. 2009;**23**:3745-3752

[63] Daneshpayeh M, Khodadadi A, Mostoufi N, Mortazavi Y, Sotudeh-Gharebagh R, Talebizadeh A. Kinetic modeling of oxidative coupling of methane over $Mn/Na_2WO_4/SiO_2$ catalyst. Fuel Processing Technology. 2009;**90**:403-410

Application of Biomedical Text Mining

Lejun Gong

Additional information is available at the end of the chapter

Abstract

With the enormous volume of biological literature, increasing growth phenomenon due to the high rate of new publications is one of the most common motivations for the biomedical text mining. Aiming at this massive literature to process, it could extract more biological information for mining biomedical knowledge. Using the information will help understand the mechanism of disease generation, promote the development of disease diagnosis technology, and promote the development of new drugs in the field of biomedical research. Based on the background, this chapter introduces the rise of biomedical text mining. Then, it describes the biomedical text-mining technology, namely natural language processing, including the several components. This chapter emphasizes the two aspects in biomedical text mining involving static biomedical information recognition and dynamic biomedical information extraction using instance analysis from our previous works. The aim is to provide a way to quickly understand biomedical text mining for some researchers.

Keywords: bioinformatics, text mining, natural language processing, information extraction

1. Introduction

With the rapid growth of the high-throughput biological technology, the study of biomedical science is entering omics era. It brings several omics data including genomics and transcriptomics; the vast amounts of biological data continue to emerge out of the life science research. The new phenomenon, new discovery, and new experimental data in biomedical research are mostly published in science journals by electronic text form. A large number of biological information is scattered in all kinds of studies. Handling these biomedical literatures could extract more biological information and discover new biomedical knowledge. Manual processing is like looking for a needle in a haystack. Biomedical literature can be seen

---

as a large unstructured data repository, which makes text mining come into play. Text mining has emerged as a potential solution to achieve knowledge for bridging between the free text and structured representation of biomedical information using artificial intelligence technology including natural language processing (NLP), machine learning (ML), and data mining to process large text collections. Therefore, text-mining technology is a powerful tool for mining valuable information from biomedical literature. There are broader definitions of biomedical text mining. Namely any work that extracts information from text could be considered as text mining, which would include a range of static information recognition, from dynamic information extraction to application of biomedical text mining. In the following sections, we describe the details according to the abovementioned aspects.

2. Natural language processing

Natural language processing is a field of artificial intelligence in computer science with interaction between computers and natural languages. With the mushroom growth of machine learning, much natural language processing research has a great relationship with machine learning. Many machine-learning algorithms have been applied to natural language-processing tasks. Extracting structured data from the complexity of heterogeneous-narrated medical reports is significantly challenged, and the work [1] obtained the results with F1 scores greater than 95% using machine learning, for example, HMM model. Weng et al. [2] proposed the machine learning to classify clinical notes to the medical subdomain. They reported that the classifier of the convolutional recurrent neural network with word embeddings yielded the best performance on iDASH and MGH datasets with F1 scores of 0.845 and 0.870, respectively. Basaldella et al. [3] proposed a hybrid approach using a two-stage pipeline with a machine-learning classifier combining a dictionary approach, and they achieved an overall precision of 86% at a recall of 60% on the named entity recognition task. For helping to obtain biomedical knowledge, a flourishing of ontologies attempted to represent the complexity of the biomedical concepts in text-mining area. The ontologies describe a wide variety of biological concepts spanning from biology to medicine. Moreover, they not only attempt to capture the meaning of a particular domain based on biomedical community but also are key element for knowledge management, and data integration [4]. Ontologies and controlled vocabularies could improve the efficiency and consistency of biomedical data curation, which has a great increasing interest in developing ontologies. For example, gene ontology (GO) [5] is a community-based bioinformatics resource related to gene function used to represent biological knowledge involving three aspects: molecular function, cellular component, and biological process. Molecular function describes the molecular activities of gene products, such as binding or catalysis. Cellular component indicates the place where gene products are active, namely a component of a cell such as an anatomical structure. While biological process represents pathways, larger processes are made up of the activities of multiple gene products. Another is GENIA [6] corpus developed to provide reference materials to allow natural language-processing techniques work for extracting information. GENIA corpus is a semantically annotated corpus of biological literature, which is being compiled and annotated in the scope of GENIA project, aiming at providing high-quality reference materials for bioinformatics. Natale et al. [7] proposed protein ontology (PRO) for enhancing and scaling

up the representation of protein entities represented in OWL and SPARQL format. Ontologies provide machine-readable descriptions for biomedical concepts linking domain-specific vocabulary. Ontology-based mining systems attempt to map a terminology in text to a concept in an ontology. Kim et al. [8] applied syntactic parsing on sentences with annotated GO concepts to exploit similarities of sentential syntactic dependencies for mapping to concepts.

The natural language processing (NLP) components include general tasks including tokenization, morphological analysis, POS tagging, and syntactic parsing.

Tokenization describes the processing that the text is broken into sentences and words. When the text is put into text-mining system, for example, a paper could be viewed as a continuous word stream; they are first broken up into chapters and paragraphs, and then the broken paragraphs continue to be pieced as sentences, words, and phonemes for a more sophisticated processing. For the tokenization task, the tokenizer could extract token features which are types of capitalization, digits, punctuation, special characters, and so on.

POS tagging is aiming at the words to annotate tags based on the context in the text. POS tags divide words into categories based on the role in the sentence. POS tags provide information about the word's semantic content. Nouns usually denote the entities, whereas prepositions express relationships between entities.

Syntactical parsing performs a full syntactical analysis of sentences according to a certain grammar theory including constituency and dependency grammars. Constituency grammars describe the syntactical structure of sentences in terms of phrases, namely element sequences. Most constituency grammars generally contain noun phrases, verb phrases, prepositional phrases, adjective phrases, and so on. Each phrase may consist of smaller phrases or words according to the rules of the grammar. The role of different phrases is contained in the syntactical structure of sentences. For example, a noun phrase may be marked as the subject of sentence, object. Dependency grammars focus on the direct relations between words, not considering the constituents. Dependency analysis uses Direct Acyclic Graph (DAG) to denote the relations between words using nodes and dependencies for edges. For example, a subject depends on the predicate verb, while an adjective depends on the noun and so on.

3. Biomedical text mining

With the enormous volume of biological literature, increasing growth phenomenon due to the high rate of new publications is one of the most common motivations for the biomedical text mining. It is reported that the growth in PubMed/Medline literature is exponential at this rate of publication. Thus, it is very difficult for researchers to keep up with the relevant publications in their own discipline, let alone related other disciplines.

Such a large scale and the rapid growth of biomedical literature data, carrying a lot of biological information, some new phenomena, biomedical discoveries, and new experimental data are often published in recent biomedical literature. Aiming at this massive literature to process, it could extract more biological information for mining hidden biomedical knowledge. These

vast amounts of biomedical literature, even in the field of expert, could not rely on the manual way from fully grasp the status quo and development trend of the research to obtain the information of interest for extracting biomedical knowledge. It is the urgent needs of application of text mining and information extraction from biomedical literature in the field of molecular biology and biomedical knowledge extraction. Biomedical text mining [9] is the frontier research field containing the collection combined computational linguistics, bioinformatics, medical information science, research fields, and so on. The development of biomedical text mining is less than 25 years [10], which belongs to a branch of bioinformatics. Bioinformatics is defined as application information science and technology to understand, organize, and manage biomolecular data. It aims to provide some tools and resources for biological researchers, facilitate them to get biological data, and analyze data, so as to discover new knowledge [11] of the biological world. Biomedical text mining is a sub-field of bioinformatics. It refers to the use of text-mining technology to process biomedical literature object, acquire biological information, organize and manage the acquired bioinformation, and provide it to researchers. Therefore, biomedical text mining can extract various biological information [12], such as gene and protein information, gene expression regulation information, gene polymorphism and epigenetic information, gene, and disease relationship. The biological information can help people to understand life phenomena and understand the rules of life activities. Using the information will help understand the mechanism of disease generation, promote the development of disease diagnosis technology, and promote the development of new drugs in the field of biomedical research. A large number of text-mining methods have been established to assist in the extraction of biological information. These methods could be proposed for extracting information that vary in their degree of reliance on dictionaries, statistical and knowledge-based approaches, automatic rule generation applying part-of-speech (POS) tagger, and some machine-learning algorithms, for example, Hidden Markov Models (HMMs) and decision trees. Cronin et al. [13] classified patient portal messages by a comparison of rule-based and machine-learning approach using a bag of words and natural language-processing (NLP) approaches. The best performance of classifier for individual communication subtypes was random forests for logistical-contact information with 0.963 receiver-operator curve.

In addition, there are some social institutes to focus on the development of biomedical text-mining technology. Based on the rapid development of omics era, the BioCreative called Critical Assessment of Information Extraction system in Biology is a community-wide effort for the evaluation of text mining and information extraction systems applied to the biomedicine domain using natural language processing [14, 15].The researches of biomedical text mining are presented at several conferences including Pacific Symposium on Biocomputing, BioNLP, and Practical Applications of Computational Biology and Bioinformatics [10].

4. Static biomedical information recognition

In the era of system biology, from the system perspective-related information on molecular biology research includes both biomedical entities, some genes, proteins, gene products, drugs, diseases such as basic, static entities to reflect its existence form, called static biomedical information.

There are some biological terminologies that describe domain objects in the medical literature, which is called entity. Such as gene that is the essence of life information, protein information that is the executor of gene function in biomedicine, identification of these entities in the life sciences plays an important role in revealing the phenomenon of life, which is the only way which must be passed to further explore these important biological entities, but also an important task in biomedical text mining. Biological entity representation in biomedical literature is extremely complex. The complexity of performance in both single entity in the form of a word entity, variable word length, and uppercase and lowercase mixed together, for example, urokinase, Cactus, IkapaBalpha, and so on. There are multi-words to form phrases, such as bradykinin B (1) receptor, protein phosphatase 2A, which brought a great deal of difficulty to establish biological entity boundaries; some of the same words or phrases that can be expressed in different categories of biomedical entities, such as c-myc, IL-2 protein can be expressed, can be said to detect gene, through the context, some biomedical entities have different forms of writing, such as protein phosphatase 2A, protein phosphatase, and 2A protein phosphatase 2A and refers to as the same biological entity in biomedical text. The complexity and diversity of the entity recognition of biomedical named entities has become a challenging study. Traditional recognition approaches have three methods containing dictionary-based, rules-based, and statistical machine learning.

Dictionary-based approach needs a detailed term dictionary for entity recognition from the document. Generally, the pre-given term dictionary is edited by the specific biological molecular database. Accordingly, the approach is limited by the term dictionary coverage but it has the relatively higher accuracy. Thus, the approach is widely applied in the actual development in the system. For example, Whatizit [16] and FRENDA [17].

Rule-based approach requires studying the entities' named features and laws for formulating rules to identify the entity, which make the staff to develop biomedical domain background knowledge. For example, biomedical entities are often noun phrases. The first character of a human gene nomenclature is a letter, and the rest is a combination of letters and numbers. Based on the successful rule system, AbGene [18], Tanabe and Wilbur used it to produce a large and high-quality gene protein dictionary from biomedical texts, and it is also used as a component of relationship extraction [19].

Statistical machine-learning approach is the rapid development along with the construction of biomedical corpora. The annotated corpus could be used as the training corpus of statistical machine-learning method. With the help of the biomedical corpora, entities could be identified from the biomedical text. There are some researches in the aspects. ABNER developed by Settles [20] used the Conditional Random Fields (CRF) as the statistical model to identify biomedical entities with the average 72.0% recall, the 69.1% precise, and 70.5% F-score in JNLPBA using morphological and semantic features. Mitsumori et al. [21] proposed an approach to process entity using Support Vector Machine (SVM) as a statistical model with internal and external resource features, which show that the performance of identification is improved by using the external biological dictionary features. Saha et al. [22] used maximum entropy model combined with word-clustering features and feature selection techniques to identify biomedical entities. The approach achieves better performance without using domain knowledge. Li et al. [23] use two-step CRFs to identify biomedical entities, the first CRF model

is used to identify named entity and the second CRF model is used to the types of named entities, which obtained 74.3% F-score in JNLPBA corpus. The mentioned three approaches have their own advantages, respectively. There is also a hybrid approach to be used for identifying biomedical entities. Our research group does some works related to identify biomedical entities involving the abovementioned three methods. We give several examples to illustrate the related static biomedical information recognition.

4.1. Dictionary-based approach to identify biomedical entities

In this section, we introduce our previous work [24–26] which is published in the International Journal of Pattern Recognition and Artificial Intelligence. We first achieve experimental literature and build a concept dictionary based on the authoritative corpus. Using part-of-speech (POS) tagging, phrase block's formulation and designed VWIA algorithm to identify entities for matching biomedical concepts. **Figure 1** [24] describes the pipeline.

4.1.1. Obtain experimental data

The experimental data of the study are from PubMed/Medline using the e-utilities API tool which is also used in the works [27–29] for automatically downloading literatures from the website (http://www.ncbi.nlm.nih.gov/). It looks like the web spider to catch a series of hyperlinks with a Uniform Resource Locator (URL). By obtaining URL using e-utilities related to
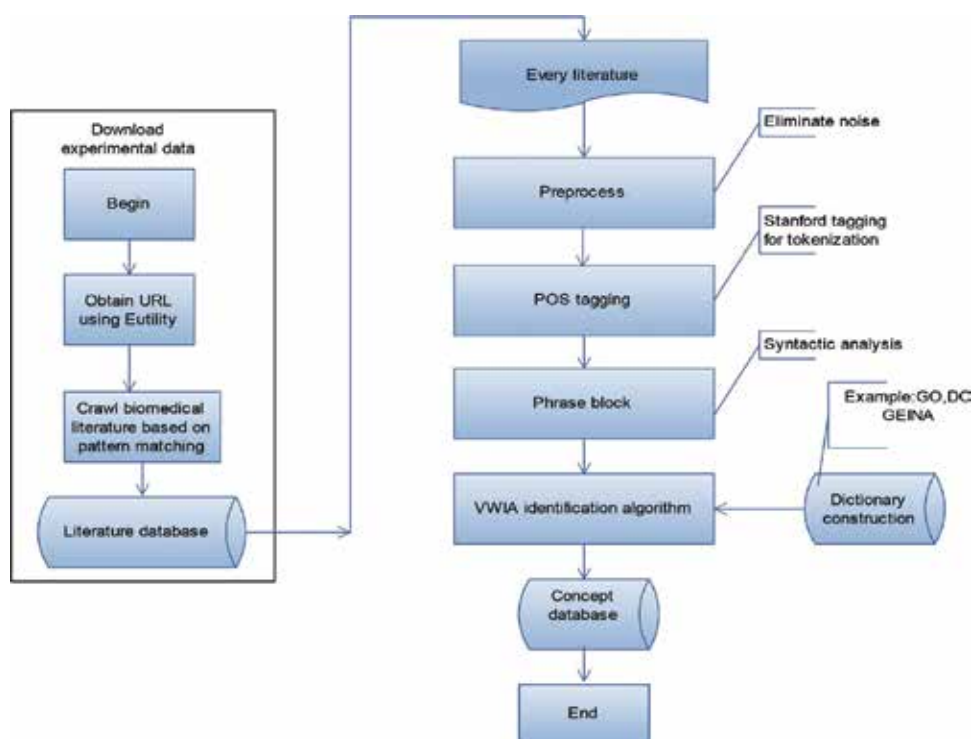

Figure 1. Pipeline of our approach to recognize biomedical concepts.

biomedical literature, the contents are achieved based on pattern mapping, and the crawled contents are stored in the local database to be further processed.

4.1.2. Preprocess, POS tagging, and phrase block

Aiming at the collected literature, to every biomedical literature, there may be several paragraphs, and each paragraph may include one or several sentences. The basic textual units identified by the tokenization as constituent tokens in each sentence are tagged by the Standford POS tagger [30]. It is a tool that appoints parts of speech to every word (for instance, verbs, adjectives, nouns, and so on) by reading text in some language with the implementation of log-linear-part-of-speech taggers. The results are shown in **Figure 2** aiming at an example sentence. Aiming at the features of biomedical concept, the POS-word pairs are important for the biomedical concepts. For example, nouns, adjectives, participles, and so on, could be components of some biomedical concepts, while some words, such as indefinite articles and verbs, would be omitted to the identification of the biomedical concepts. Thus, we extract phrase blocks combined by nouns, adjectives, and participles. These phrase blocks are preprocessed as a unified base form. For example, the word of tagging POS labels "NN" (single noun) and "NNS" (plural noun) transforms the word of single noun for the normalization. **Figure 2** shows an example about the POS tagging [25].

4.1.3. Biomedical dictionaries' construction

Considering the high precision of dictionary-based approach, the approach used it for the identification of the biomedical concepts. Thus, concept dictionary is built by some authorized biomedical ontologies, for example, Disease Ontology (DO), Gene Ontology (GO), and GENIA ontology. GENIA corpus is used to provide reference material for biotext mining. GENIA corpus is also semantically annotated corpus containing 2000 MEDLINE literature with almost 100,000 annotation information and more than 507,325 words including 18,546 sentences for biomedical terms in 3.02 version. Each article is encoded in an XML-based mark-up scheme with ID, title, and abstract in the corpus. Moreover, both abstracts and titles have been marked up for meaningful annotated terms, biologically and semantically. The corpus provides semantically annotated biological terms identifiable with any terminal concepts for GENIA ontology, For instances, "<cons lex = "IL-2_gene" sem = "G#DNA_domain_or_region" > IL-2 gene</cons>", the label "lex" describes the concepts, and the label "sem" represent the type of concept. Due to the structural scheme of GENIA ontology, it could be extracted by the regular expression. These extracted biologically meaningful terms build concept dictionary for the

IL-2/NN   gene/NN   expression/NN   and/CC   NF-
kappa/NN   B/NN   activation/NN   through/IN
CD28/NNP   requires/VBZ   reactive/JJ   oxygen/NN

Figure 2. Sentence with part-of-speech tags generated by Stanford maximum entropy part-of-speech tagger. Tags are NN: normal noun; IN: preposition or conjunction; VBZ: verb in present/past tense; JJ: adjective; CC: conjunction; NNP: proper noun.

biomedical text mining. For example, the concept taxonomy of GENIA 3.02 version contains some biologically relevant nominal categories as shown in **Table 1** [24].

4.1.4. Results

Biomedical texts are divided into a few sentences in biomedical literature, and for every sentence, some phrase blocks are parsed. The algorithm named Variable-step Window Identification Algorithm (VWIA) is developed for identifying biomedical concepts. The approach obtained the overall 95.0% F-measure aiming at the GENIA corpus. The implementation of the approach is shown in **Figure 3** [24].

4.2. Machine-learning approach

In this work [26, 27], we introduce machine learning to help recognize biomedical named entity. The pipeline architecture of our approach is shown in **Figure 4** [26, 27].

The pipeline of our system mainly contains four modules: preprocessing module, training module, tagging module, and testing module.

Categories	Numbers	Categories	Numbers
protein_molecule	21,632	peptide	524
protein_family_or_group	8372	body_part	449
DNA_domain_or_region	8054	atom	341
cell_type	7233	RNA_family_or_group	334
other_organic_compound	4096	polynucleotide	260
cell_line	3974	inorganic	256
protein_complex	2417	nucleotide	239
lipid	2359	mono_cell	222
virus	2126	other_artificial_source	209
multi_cell	1766	protein_substructure	129
DNA_family_or_group	1558	DNA_substructure	107
protein_domain_or_region	1017	protein_N/A	99
protein_subunit	920	carbohydrate	98
amino_acid_monomer	785	DNA_N/A	48
tissue	692	RNA_domain_or_region	39
cell_component	669	RNA_N/A	15
RNA_molecule	602	RNA_substructure	2
DNA_molecule	542	Total	72,185

Table 1. Biomedical concept categories and numbers in GENIA.

Figure 3. Visualization of the dictionary-based approach.
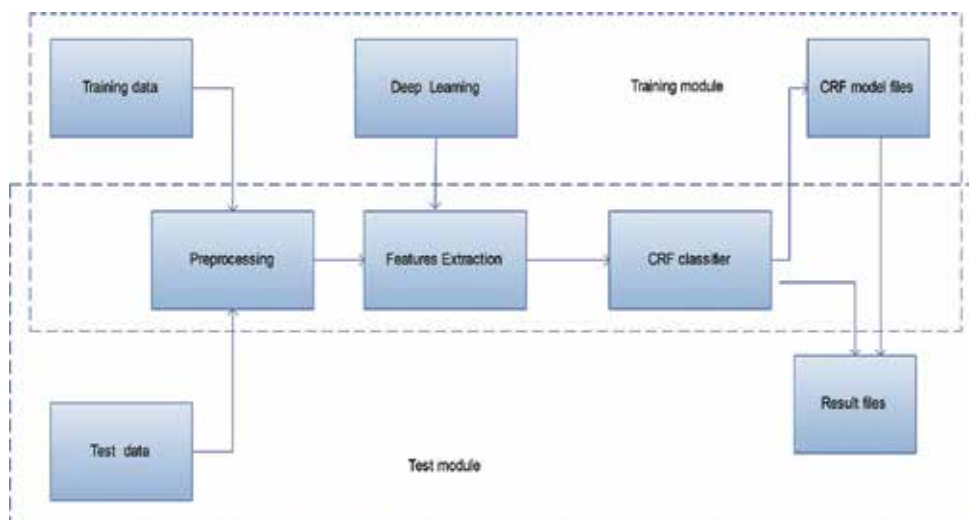
Figure 4. The pipeline architecture of machine-learning approach.

4.2.1. Preprocessing module

Applying machine-learning methods, the original biomedical texts could not directly be processed by the CRF model; they would be preprocessed to form the related formats. The original training, testing, and predicted texts would be processed to the specified format files with feature dataset.

4.2.2. Training module

Trained texts are put into the preprocessing module, aiming at the selected and extracted features to form the specified trained files that are trained with some parameters' set, which could obtain trained model files. The input of training module is the result of the preprocessed training texts, while the output of it is the model files including feature function set and weight parameter.

4.2.3. Tagging module

After the preprocessing module processes the test texts, the formed specified test files are put into a tagging module together with the model files, which could obtain the tagged files. The input of a tagging module is the result of preprocessed testing texts, while the output of it is the tagged result files.

4.2.4. Testing module

The testing module is to measure the performance of our system. After processing the test texts in the preprocessing module, the test standard files are put into a test module together with the tagging result files. The input of a tagging module is the results of preprocessed test files, while the output of it is the identified performance of our system including precision, recall, and F-measure.

4.2.5. Results

The approach considered features including POS features, word surface clue feature (uppercase, lowercase, numbers, specific char, initial) using the Genia corpus 3.02 version to train and test the system's performance with a 10-cross validation. The system's performance is shown in **Table 2** related to the six classes. According to the above method, using Java programming based on Linux OS to develop Biomedical entity recognition Miner system called (BerMiner). **Figure 5** shows the results of the identified biomedical entity.

Project	Precision (%)	Recall (%)	F-measure (%)
DNA	61.31	45.79	52.4
RNA	56.48	44.63	49.69
Protein	76.48	72.64	74.50
Cell-type	71.11	58.51	64.16
Cell-line	71.80	53.58	61.29
Virus	78.56	66.37	71.83

Table 2. Identified biomedical entity's performance based on machine learning [26, 27].
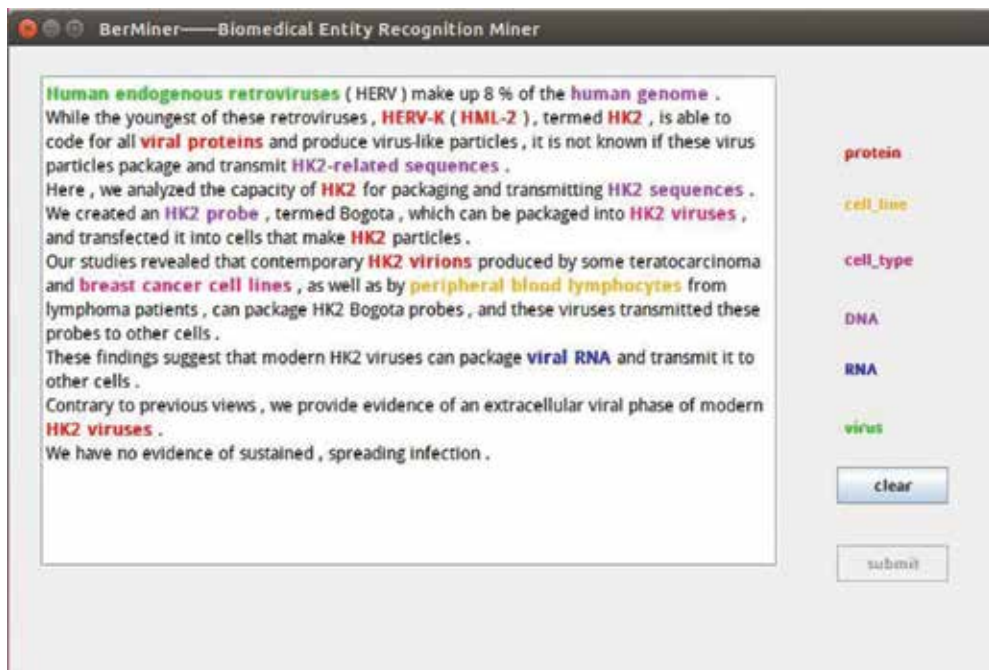
Figure 5. BerMiner extracted results with six categories using machine learning [26, 27].

5. Extraction of dynamic biomedical information

Biomedical entities produce a series of information interaction in the process of genetic information transfer and expression, such as genes and gene interaction relations, the relationship between genes and disease, the relationship between genes and gene product, molecular signal conduction pathways, and so on. The information is represented as the dynamic form. Dynamic biomedical information represents the process of activities of biomedical entities. Dynamic biomedical information is extracted, namely association between biomedical entities which is often extracted based on entity co-occurrence analysis with statistics theory. Glenisson et al. [28] proposed an approach to extract the relationship between genes using vector space method and k-medoids algorithm for gene clustering. Wren [29] explored the method to measure biological entities relationship using mutual information model. Wu et al. [30] researched the interactions between genes and drugs based on text-mining technology, which was divided into three steps. First, the approach identified genes and drugs entities from Medline abstract; the second step is to extract different levels of gene-2-drug pairs. The third step is to rank the gene-2-drug pairs based on mathematical statistical model. Our research group has done the similar works [31–33]. Here, we example the works [31, 33] for describing the process related to the extraction of dynamic biomedical information in detail.

5.1. Relationship extraction based on statistic model

Dynamic information represents the process of activities of biomedical entities. In this study, we focus on the dynamic process of biomedical entities, and extract dynamic biomedical information, namely association between biomedical entities, based on entity co-occurrence analysis which is attached to the statistics theory with more precision. Entity co-occurrence analysis considers that if any two entities occurred in a certain level of paper (e.g., a full text, a paragraph, a sentence, and a phrase), then the two entities could have be related. Different levels have different strengths of entity association. Through syntactic analysis of biomedical texts, the weight of association in phrase level is the highest than in other two levels, and the sentence level is higher than the full text. Due to the multiclass biomedical entities used in this study, extracted associations are also multiple types. We build a data modeling of entity association based on entity co-occurrence analysis with statistics. The data modeling could formally be represented as a three tuples as shown in Eq. (1)

$$D = (E, C, R) \tag{1}$$

Let E be a set of biomedical entities, and C be a set of types of association. Let R be a set of correlation of associations between biomedical entities.

Supposing $\varphi(e)$ represents entity category where $e \in E$; W is a set of weights of association levels where $w_k \in W$. To the two entities ($e_i, e_j \in E$) of the kth level, their co-occurrence frequency is represented as $f_k(e_i, e_j)$, and the correlation between two biomedical entities is defined as Eq. (2)

$$T = \left(e_i, e_j, C\left(\varphi(e_i), \varphi(e_j)\right), R(e_i, e_j)\right) \tag{2}$$

Let $C_{(\varphi(e_i), \varphi(e_j))}$ be association category related to the entity category $\varphi(e_i)$ and $\varphi(e_j)$. For instance, $C_{(\varphi(e_i), \varphi(e_j))}$ could represent association between gene and disease, or association between gene and microRNA, and so on. Let $R_{(e_i, e_j)}$ be correlated factor between entity e_i and e_j as shown in Eq. (3)

$$R(e_i, e_j) = \sum_k w_k f_k(e_i, e_j) \tag{3}$$

After building data modeling of entity association, we further consider extracting these dynamic biomedical information from biomedical literature. Aiming at the entities identified, we design an algorithm of Mining Multiclass Entity Association, named (MMEA), under the data modeling based on co-occurrence statistical analysis as shown in **Figure 6**.

The input of MMEA algorithm is entities identified building on the step of entity recognition. The MMEA algorithm first gets the category of each entity (lines 4–6). Aiming at an entity e_i, the algorithm decides the types of associations between it and other entities (line 7) and computes the correlation factor $R_{(e_i, e_j)}$ by Eq. (3) (line 8). The above-achieved results are stored in the four tuples $T = \left(e_i, e_j, C\left(\varphi(e_i), \varphi(e_j)\right), R(e_i, e_j)\right)$ for further processing (line 9). The process proceeds with the increment of i entity until all entity associations are obtained (lines 3–11).

1. Input: $E = \{e_1, e_2, ... e_m\}$:
2. Output: T ;
3. for each i
4. $e_i \rightarrow \varphi(e_i)$;
5.   for $j=i+1$ to m
6. $e_j \rightarrow \varphi(e_j)$;
7. $(\varphi(e_i), \varphi(e_j)) \rightarrow C(\varphi(e_i), \varphi(e_j))$;
8.   Compute $R(e_i, e_j)$ by $\sum_k w_k f_k(e_i, e_j)$;
9. $T = (e_i, e_j, C(\varphi(e_i), \varphi(e_j)), R(e_i, e_j))$;
10.   end;
11. end.

Figure 6. MMEA algorithm.

5.2. Relationship visualization

Biomedical text mining focuses on the centrality of user interactivity, and it needs to provide users for interacting with data results. The text-mining visualization facilitates user interactivity with graphical approaches. For example, we developed a circle network graph to allow
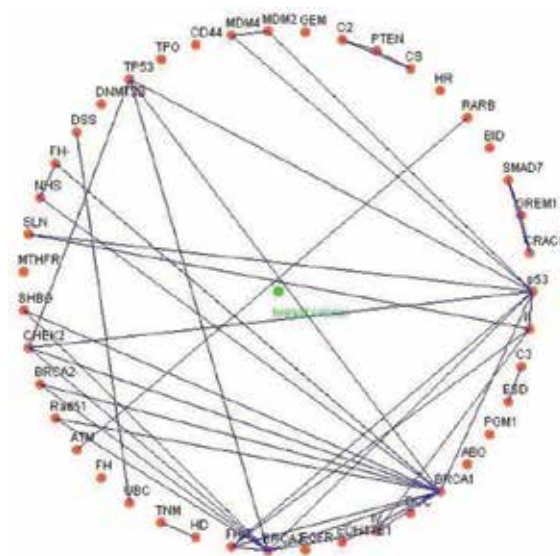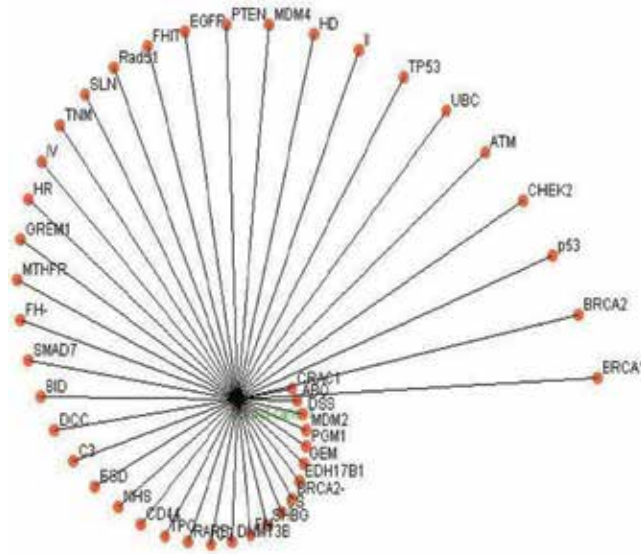

Figure 7. Relationship visualization between genes.

Figure 8. Relationship visualization between disease and gene.

disease researchers to explore the relationship between genes related to breast cancer. To capture the susceptive genes related to breast cancer, a fan-like network visualization is designed by our research group. The node represents the biomedical entities, and the link lines indicate the associations between entities (as shown in **Figures 7** and **8**).

6. Conclusions

This chapter first introduces the rise of biomedical text mining. Then, it describes the biomedical text-mining technology, namely natural language processing, including the several components. In the following sections, it emphasize the two aspects in biomedical text mining involving static biomedical information recognition and dynamic biomedical information extraction using instance analysis which our previous works.

Acknowledgements

Conflict of interest

In this chapter, the instances are from our previous works [24–27, 31–33].

Author details

Lejun Gong

Address all correspondence to: glj98226@163.com

Jiangsu Key Laboratory of Big Data Security and Intelligent Processing, School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China

References

[1] Zheng S, Lu JJ, Ghasemzadeh N, Hayek SS, Quyyumi AA, Wang F. Effective information extraction framework for heterogeneous clinical reports using online machine learning and controlled vocabularies. JMIR Medical Informatics. 2017;**5**(2):e12. DOI: 10.2196/medinform.7235

[2] Weng WH, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Medical Informatics and Decision Making. 2017;**17**(1):155. DOI: 10.1186/s12911-017-0556-8

[3] Basaldella M, Furrer L, Tasso C, Rinaldi F. Entity recognition in the biomedical domain using a hybrid approach. Journal of Biomedical Semantics. 2017;**8**(1):51. DOI: 10.1186/s13326-017-0157-6

[4] Krallinger M, Leitner F, Vazquez M, Salgado D, Marcelle C, Tyers M, Valencia A, Chatr-aryamontri A. How to link ontologies and protein-protein interactions to literature: Text-mining approaches and the BioCreative experience. Database: The Journal of Biological Databases and Curation. 2012;**2012**:bas017

[5] Gene Ontology Consortium. Gene ontology consortium: Going forward. Nucleic Acids Research. 2015;**43**(Database issue):D1049-D1056

[6] Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus – semantically annotated corpus for bio-textmining. Bioinformatics. 2003;**19**(Suppl 1):i180-i182

[7] Natale DA, Arighi CN, Blake JA, et al. Protein ontology (PRO): Enhancing and scaling up the representation of protein entities. Nucleic Acids Research. 2017;**45**(D1):D339-D346

[8] Kim JJ, Park JC. Bioie: Retargetable information extraction and ontological annotation of biological interactions from the literature. Journal of Bioinformatics and Computational Biology. 2004;**2**(3):551-568

[9] Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Briefings in Bioinformatics. 2005;**6**(1):57-71. Review

[10] Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: Current progress. Briefings in Bioinformatics. 2007;**8**(5):358-375

[11] Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. Methods of Information in Medicine. 2001;**40**(4):346-358

[12] Krallinger M, Erhardt RA, Valencia A. Text-mining approaches in molecular biology and biomedicine. Drug Discovery Today. 2005;**10**(6):439-445 Review

[13] Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. International Journal of Medical Informatics. 2017 Sep;**105**:110-120

[14] Leitner F, Mardis SA, Krallinger M, Cesareni G, Hirschman LA, Valencia A. An overview of BioCreative II.5. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2010;**7**(3):385-399

[15] Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: Critical assessment of information extraction for biology. BMC Bioinformatics. 2005;**6**(Suppl 1):S1

[16] Rebholz-Schuhmann D, Kirsch H, Couto F. Facts from text–is text mining ready to deliver? PLoS Biology. 2005;**3**(2):e65

[17] Barthelmes J, Ebeling C, Chang A, et al. BRENDA, AMENDA and FRENDA: The enzyme information system in 2007. Nucleic Acids Research. 2007;**35**(Database issue):D511-D514

[18] Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics. 2002;**18**(8):1124-1132

[19] Tanabe L, Wilbur WJ. Generation of a large gene/protein lexicon by morphological pattern analysis. Journal of Bioinformatics and Computational Biology. 2004;**1**(4):611-626

[20] Settles B. ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. 2005;**21**(14):3191-3192

[21] Mitsumori T, Fation S, Murata M, et al. Gene/protein name recognition based on support vector machine using dictionary as features. BMC Bioinformatics. 2005;**6**(Suppl 1):S8

[22] Saha SK, Sarkar S, Mitra P. Feature selection techniques for maximum entropy based biomedical named entity recognition. Journal of Biomedical Informatics. 2009;**42**(5):905-911

[23] Li L, Zhou R, Huang D. Two-phase biomedical named entity recognition using CRFs. Computational Biology and Chemistry. 2009;**33**(4):334-338

[24] Gong L, Yang R, Liu Q, Dong Z, Chen H, Yang G. A dictionary-based approach for identifying biomedical concepts. International Journal of Pattern Recognition and Artificial Intelligence. 2017;**31**(9):1757004. http://dx.doi.org/10.1142/S021800141757004X

[25] Gong L, Sun X. ATRMiner: A System for Automatic Biomedical Named Entities Recognition. Yantai: ICNC 2010; 2010. pp. 3842-3845

[26] Gong LJ, Yang RG, Yang HY, Jiang KY, Yang G. BerMiner: A machine learning system for identifying bio-entity. 2015 International Conference on Software Engineering and Information System (SEIS 2015); 2015:447-450

[27] Yang RG, Wu ZX, Yang Z, Yang G, Gong LJ. Identifying biomedical entity based on deep learning. 2015 International Conference on Software Engineering and Information System (SEIS 2015); 2015:713-718

[28] Glenisson P, Coessens B, Van Vooren S, et al. TXTGate: Profiling gene groups with text-based information. Genome Biology. 2004;**5**(6):R43

[29] Wren JD. Extending the mutual information measure to rank inferred literature relationships. BMC Bioinformatics. 2004;**5**:145

[30] Wu Y, Liu M, Zheng WJ, et al. Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. Pacific Symposium on Biocomputing. 2012:422-433

[31] Gong LJ, Yang RG, Dong ZJ, Chen H, Yang G. Extraction of disease-centred dynamic biomedical information from literature. Journal of Computational and Theoretical Nanoscience. 2015;**13**(1):722-727 12

[32] Gong LJ, Yang RG, Sun X. Prioritization of disease susceptibility genes using LSM/SVD. IEEE Transactions on Biomedical Engineering. 2013;**60**(12):3410-3417

[33] Gong LJ, Wei YB, Xie JM, Yuan ZD, Sun X. Text mining approach for relationships between genes and diseases. Dongnan Daxue Xuebao (Ziran Kexue Ban)/Journal of Southeast University (Natural Science Edition). 2010;**40**(3):486-490

Static/Dynamic Zoometry Concept to Design Cattle Facilities Using Back Propagation Neural Network (BPNN)

Sugiono Sugiono, Rudy Soenoko and
Rio Prasetyo Lukodono

Additional information is available at the end of the chapter

Abstract

The dairy cattle productivity is largely dependent on the facility quality and environmental condition. Various researchers had conducted a study in this field, but it is not developing the knowledge of animal dimensions and behaviors correlated with their facility design. Complexities of dynamics zoometry depend on cow behaviors that they are forced to use neural network (NN) approach. Hence, the purpose of this chapter is to create the concept of static and dynamic zoometry to guide the ergonomics facilities design. The research started with study literature on anthropometry, dairy cattle, facility design, and neural network. The following step is collecting the static zoometry data in 16 dimensions and dynamics zoometry in 7 dimensions. On the one hand, static data is utilized as an input factor. On the other hand, dynamic data is utilized as desire factor of back propagation neural network (BPNN) model. The result of BPNN training is utilized to design the dairy cattle facilities, e.g., cage with minimal length = 357.67 cm, width = 132.03 cm (per tail), and height = 205.28 cm. The chapter successfully developed the concept of zoometry approach and BPNN model as a pioneer of implementing comfort knowledge.

Keywords: animal comfort, cattle cage, facility design, neural network, zoometry, ergonomics, milk productivity

1. Introduction

Milk is an important food commodity in the world as it provides calcium, phosphorous, magnesium, and protein which are all essential for human health. Adequate consumption of milk

from early childhood and throughout life strengthens bones and protects against diseases. Consuming dairy product, especially milk, can increase children bone health and development, develop a positive brain concentrations of glutathione (GSH), and increase the body's resistance against many infectious diseases and diseases caused by malnutrition and reduce cancer risk [1–4]. Milk can directly be consumed or manufactured into other dairy products such as cheese, ice cream, butter, ghee, cream, yogurt, etc. Milk provides the following beneficial nutrients in varying quantities [5–8]:

1. Calcium—for healthy bones and teeth

2. Phosphorous—for energy release

3. Magnesium—for muscle function

4. Protein—for growth and repair

5. Vitamin B12—for production of healthy cells

6. Vitamin A—for good eyesight and immune function

7. Zinc—for immune function

8. Riboflavin—for healthy skin

9. Folate—for production of healthy cells

10. Vitamin C—for the formation of healthy connective tissues

11. Iodine—for regulation of the body's rate of metabolism (how quickly the body burns energy and the rate of growth

As community realizes the importance of milk product in their life, it boosted the demand for milk production. As consequence, the number of dairy cattle farmers and productivity should be increased. The government needs an approach to increase fresh milk production. One of the methods to increase the milk productivity is maintaining cow facilities and developing them to be more comfortable for the cattle (cage, free stall, floor, etc.). Several researchers investigated the dairy cattle comfort including cow house and respective facilities. Cook et al. reported that physical accommodation of dairy cattle should provide a relatively dry area for the dairy cattle to lie down and be comfortable [9]. Dairy cow overcrowds could reduce rest time, increase idle standing in alleys, alter feeding behavior, and, in general, reduce cow comfort [10]. A study conducted on 47 farms in Northeastern Spain explained the significant effect of a stall on the productivity of dairy cow and impact on milk production and cow health [11]. The cow facilities should be constructed to minimize time to reach food and water. Free stall facility is usually selected to minimize the effect of weather changes and improve cleanliness and cow comfort. Dairy facilities should be designed to keep the cows and calves comfortable in order to maintain dry matter intake (DMI) and thus maximize economic production. Other studies have reported similar results. Based on Jim Reynolds research, the cattle comfort can be identified into five factors for heat stress, sanitation, free stall design, walking surface, and

walking distance [12]. In short, dairy cattle facilities are an important factor of key success in milk business, and it should be designed to keep the cow and calves comfortable in order to maintain milk quality and thus maximize economic production.

In order to provide a solution to comfort design for dairy cow facilities, the physical ergonomics knowledge is needed. The important knowledge of physical ergonomics area for a human body is defined as anthropometry. It studies the variation of the human body dimensions in static and dynamics condition [13]. Using the same analogy, this study converts the concept of anthropometry into the concept of zoometry to study the impact of variation of animal body dimensions. This analogy imagines that animal (cow) requires comfort in order to gain optimal milk production, similar to human. A human requires comfort at work in order to reach high productivity. Zoometry is derived from Greek ζώo (zóo) which means animal and μέτρον (metríste) that means a measurement. Zoometry concept is defined as static and dynamics animal measurement dimension in order to determine the physical variation of the specific animal population. This research utilized dairy cattle zoometry concept to design cattle facility (house, free stall, floor, etc.) in order to increase the dairy cattle comfort and milk production. This study considers static and dynamic measurement because it constitutes cattle main activity. There are three main activities that are affected by the static and dynamic condition such as feeding, lying, and standing or transition events between lying and standing. Zoometry concept can deter cattle from annoyances or uncomfortable facility. The research promotes pioneer strategy to improve the cow comfort using zoometry concept.

Directly and manually measuring dairy cattle dimensions based on zoometry concept is time-consuming and costly. Analysis data of dairy cattle dimensions can be employed to eliminate the problems. Unfortunately, relationship complexity of each dimension on dairy cattle is hard to describe in mathematics formula or in the regression model. Artificial intelligence back propagation neural network (BPNN) I was used in determining the best solution. It can eliminate some cost and improve efficiency. The BPNN model can be utilized to predict pattern making-related body dimensions by inputting few key dairy cattle body dimensions.

2. Zoometry concept

2.1. Anthropometry

The first step to understand the zoometry concept is knowing the anthropometry concept. Cattle zoometry concept is adapting the principles of anthropometry rules for human dimensions. Anthropometry is human body measurement method in a population and analyzes that measurement for various purposes. This measurement could be utilized in architecture, product design, clothing design, child nutrients, design workplace, etc. The ISO 7250-1: 2008 intended to explain as a guide for an engineer who is required to determine the size of human physic for their job [14]. The knowledge on human body, technical measurement, and statistic are very important to obtain a higher quality of anthropometry implementation. Anthropometry will improve work facilities capable to facilitate a person in working physically through ease of reach and access and through the process of cognition of a job [15, 16].

The anthropometry will set up an economical design work facility which means from a marketing point of view, these facilities could be built cheaply. On the other hand, although the design is made with cheaply but can support the performance of human according to the technical needs. Some considerations that can be utilized in anthropometry include data utilized to indicate the characteristics of the target population with limitations contained in the field conditions. The criteria utilized for the design are appropriate for user facilities. The data collected are expected to represent broad conditions that can be utilized for a wider population [17]. The posture utilized by every individual work will be influenced by the body and tool dimension as well as facilities used. The level of relationship between work facility and posture is influenced by the characteristics and frequency of interaction between the two. Generally, anthropometry is divided in into two branches: static anthropometry and dynamic anthropometry [18, 19]. The static anthropometry deals with the measurement process when the human body is in a stable position or in a static condition. On the other hand, dynamic anthropometry deals with measurement process that relates to the measurement of range of human body movement, for example, of arm movement, walking position, and head movement to reach an object.

2.2. Cattle psychology and physiology

Similar to humans, cows are also able to perform a process of cognitive response to the symptoms of the surrounding environment. Cows perform various activities inside an environment which is described in **Table 1** [20].

Dairy cattle cognitive process is exhibited in performing activities by responding to surrounding environment, which is described as follows [21]:

1. ability to distinguish objects around;

2. have a certain emotional level that can be formed because of the interaction with the surrounding environment;

3. ability to show an emotional reaction which is a reflection of the process of cognition;

4. have a different personality from one to another; and

5. ability to conduct social learning.

No	Activities	Daily allocation (hour)
1	Eating	2–5
2	Lying	12–14
3	Interaction	2–3
4	Ruminating	7–10
5	Drinking	0.5
6	Outside pen	2.5–3.5

Table 1. Cow activity.

Dairy cattle psychology respects to some conditions, e.g., conditions of the cage, SUI as the comfort index, cow's desire to always be in a lying position, cattle density in the room, comfortable and clean sitting position, high-quality feed and drinks, minimum competition to obtain food and drink, cow gets enough room to conduct activity, anti-slip floor, comfortable air cycle, comfortable lighting, and a shady spot [22, 23].

Dairy cattle behaviors are well developed in feeding, environment/microclimate condition, facility design, house, and social communication. The dairy cattle are able to distinguish red, yellow, green, and blue colors. However capability in differentiating between green and blue is poor [24]. Moreover, cattle are able to distinguish simple shapes such as triangles, circles, and line. Color information is important in cow facility design to increase dairy cattle comfort. In the free area, the cow automatically moves from the dark to light area. They tend to avoid strong contrast between sun and shadow. Comparing to human hearing capability, cow possesses almost similar frequency range and are able to listen to high tones that human cannot hear. Cattle hearing is important in inter- and intra-species communication [25]. Cattle sense of touch is important in determining which herbage is rejected or accepted. The secondary/special olfactory system can detect pheromones, volatile chemicals that are important in reproduction and feed selection [26].

Cattle communicate by sending out a different signal such as poses, sound, and smells [24]. A high density of cattle inside the house limits the freedom movement and can increase social stress. Cattle possess a distinct circadian rhythm, in which the main rest, feed, and rumination activities vary according to a fixed pattern. Grazing occupies a large amount of time for dairy cows about 8 hours/day. Grazing behavior is affected by many factors, including environmental conditions and plant species. In a dairy herd of Friesian cows, it was found that there was a consistent order for lying down and standing up [27]. The natural lying down behavior begins when the animal sniffs at the ground while it slowly moves forward. The head and body of the fully developed cow are thrust 0.60–0.70 m forward during the lying down process. When a cow wants to get up in a natural way, it firstly rises to its knees, and afterward the hind part of its body is swung up via the knees, which function as a rocking point.

There are many kinds of dairy cattle, e.g., fries Holland (the Netherlands), Shorthorn (UK), Holstein Friesian (the Netherlands), Jersey (the UK, France), Brown Swiss (Switzerland), Red Danish (Denmark), Drought Master (Australia), etc. The dairy cattle in Indonesia is dominated by Holstein Friesian possessing white and black spot or red spot. The female cow has average weight = 560 Kg to 725 Kg, and the male cow has average weight = 820 Kg to 1000 Kg. The dairy cattle have to grow up from calf, adult cow, mature cow, and old cow. A new-born calf weighs from 90 to 100 pound and has height from 32 to 36 inches. The 6-month-old heifer starts to graze (eat grass) in the pasture. Heifer usually has weight about 400 pounds with height from 38 to 42 inch. Yearling cow has to weight about 700 pounds and height from 47 to 52 inch and still has quite a bit of growing to do before it joins milking herd in another year. Two-year-old dairy cattle start to produce milk and keep on growing for next few years to be a mature cow. It weighs 1200 pound and has height from 53.5 to 55 inch. The last stage is mature cow which has more than 1500 pound and produces optimal milk. Holstein Friesian can produce milk around 57.000 Kg per year with low fat content at approximately 3.5 to 3.7% [28].

Cow body dimension will influence the horizontal movement of the cow when it gets up or lies down. It uses space around 3 m. The moving forward motion is 0.6 m, and minimum distance to the bedding from the head or neck of the cow is approximately 0.2 m [24]. The reach of dairy cattle during feed intake depends on the type of tether and feed alley height. The body length of the cow, from the shoulder area to the tail head and spine, is not flexible which make it difficult for the cow to make sharp changes of direction while it is walking. Therefore much space is required when a cow turns. The range of vision of cattle covers 330–360^0, and the field of vision covered by both eyes at the same time is 25–30^0.

Similar to a human dimension, there are several things that can affect dairy cattle dimension as defined in measurements as follows:

• *Species*

Any different taxonomic levels will have a tendency of different dimensions. The higher differences level of the taxonomy will have higher different dimensions as well. Zoometry measurement should be taken in the similar to animal species (types).

• *Phase development*

Animals that undergo metamorphosis or change in phase of development will have different dimensions in every phase of its path.

• *Age*

Animal body size will vary in each period of growth.

• *Gender*

Male generally has larger body dimensions than females.

• *Clumps*

Diversity clumps in the animals lead to a tendency of difference in size in any dimension zoometry.

2.3. Zoometry concept

Similar to anthropometry for human body dimensions, the zoometry concept concerned with the comparative measurement of the animal body and its part as well as the variables which impact these measurements. The main goal of zoometry is to increase the cattle comfortability which can influence physiological and psychological condition. The good physiological and psychological conditions will increase the amount of daily milk production. Zoometry can be utilized to define the best size of cattle facilities, such as cage size, stall, floor, etc.

The first step to create the zoometry concept is developing a database of cattle dimensions. The zoometry data of dairy cattle are collected from dairy farmers in Indonesia with an average age between 3.5 to 6 years old (optimal daily milk production). Total dairy cattle taken as the sample was around 500 samples. Generally, the measurement is divided into two sections for static dimension (mentioned later as static zoometry) data and for cow movement

(dynamic zoometry) data. Equipment used in this research are a paper sheet, pen, ruler, stopwatch, and handy cam. The static measurement focused on the cow body dimension, e.g., length of the body, length of the leg, body width, neck length, etc. The dynamic zoometry is taken when the cow is engaged in physical activity. The measurement focused on cow movement (walking), moving the tail, moving head during drink or eating, and standing up to lying down or vice versa. To obtain detailed measurement, some videos have been taken during the measurement process. Briefly, statistic test for normality and validity data are presented in zoometry data analysis.

Figure 1 exhibits one example of measuring the static zoometry for a dairy cow. All measurers have received training before they start working on the dairy farm. Training included how to measure, understanding zoometry concept, knowing the dairy cattle behaviors, and implementing the animal research ethics. Measurements are made under tight control supervision. The quality control procedures during and after survey are explained afterward.

The research investigated zoometry for dairy cattle for both of static and dynamic zoometry measurements. Zoometry concept is important to understand the dairy cattle lives. Zoometry dimension is defined in both static and dynamic measurement of cattle body dimensions and cattle behaviors. There are 16 dimensions of static zoometry for dairy cattle. **Figure 2a** and **b** defined all the dimension (D_1, D_2, D_3,.., D16) of static zoometry in the 2D Picture. The dimensions D_1 to D_{10} explain the position in the front view of dairy cattle (lengthwise direction), and the dimension of D_{11} to D_{13} describes in the lateral direction of dairy cattle. The D1 is for the height of the head, D2 is the height of the body, D3 is the length of neck + head, D14 is head width, D15 is the length of tail, and D16 is the length of horns.

Table 2 shows some of the results of static zoometry measurement in 16 dimensions as mentioned before with 25 data, respectively. Homogeneity tests are used to ensure when the data collection of cow body size is in a uniform condition without any specific arrangement. According to the data, the average of data and the deviation standard are calculated for all

Figure 1. Dairy cattle zoometry measurement for static and dynamic data.
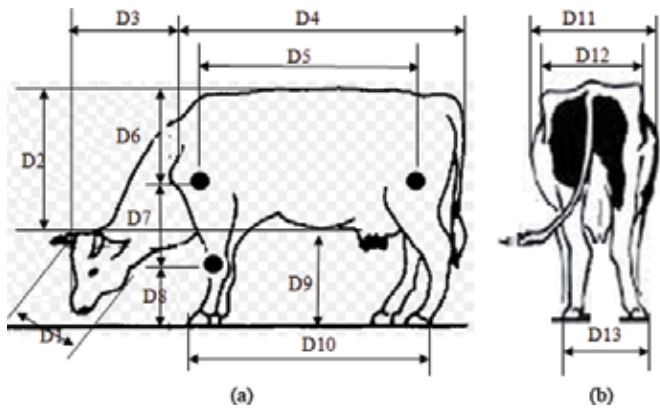
Figure 2. Static zoometry dimensions: (a) front view of cow dimensions (D_1 to D_{10}) and (b) the backside of cow dimensions (D_{11} to D_{13}).

No	Explanation	Dairy cattle dimension				
		1	2	—	499	500
D1	Height of the head	36	38	—	38	40
D2	Height of cow body	77	78	—	78	79
D3	Length of the head + neck	104	106	—	104	105
—	—	—	—	—	—	—
D15	Length of the tail	114	113	—	115	120
D16	Length of the corn	9	14	—	10	8

Table 2. Data measurement results from 500 number of source data of cattle for static dairy cattle dimension.

dimensions (D1–D16), as example, D1 has average $\bar{D} = 39.64$ and deviation standard (σ_1) = 2.68. Moreover, D1 has lower control limit (LCL) = 34.17, and upper control limit (UCL) = 44.47 is categorized in homogeneity data. The same way, the other static cattle dimensions (D2–D16) are also categorized in homogeneity data. As a result the data measurement is ready to use for the next step of sufficient data test.

Sufficient data test for static cattle dimensions is determined based on formula 1. To calculate the number of data requirement (N′), the research select confidence level of collecting data 95% (k = 2), and error = 5% (s = 0.05) which has N′ = 7. As a result the data D1 can be categorized in sufficient data (N′ < N). Using the same way, the data of the other dimension D2–D16 all are categorized in sufficient data:

$$N^{'} = \left[\frac{\left(k/s \sqrt{N \sum_{i=1}^{N} x_i - \left(\sum_{i=1}^{N} x_i \right)^2} \right)}{\sum_{i=1}^{N} x_i} \right]^2 \tag{1}$$

where N′ = data should be taken.

N = data have been collected

k = level of confidence

s = level of error

x_i = observation data

$$N' = \left[\frac{(2/0.05\ \sqrt{25(966289) - (983)^2})}{983}\right]^2 = 7$$

Human dynamic anthropometry is concerned with the measurement of human work or human motion, e.g., hand movement, sitting down, turning, etc. In analogy with human data, the dynamic zoometry is defined with animal measurement (dairy cattle) on movement and cattle behaviors, e.g., vertical head movement to reach food, the vertical movement to lie down/get up, tail movement, etc. The data is very important in designing the comfortable cattle facilities, e.g., free stall, watering system, floor, house, and feeding rack. **Figure 3** explains the dimensions of dairy cattle dynamic zoometry. There are seven dimensions for D_{17} to D_{23}. Following statements explained dynamic dimension:

- D_{17} dimension is angle scope for vertical movement of a cow head.

- D_{18} dimension is angle scope for horizontal movement of a cow head.

- D_{19} dimension is leg reach on walking movement.

- D_{20} dimension is angle scope for horizontal movement of cow tail.

- D_{21} dimension is length for laying down movement.

- D_{22} dimension is length for raising movement.

- D_{23} dimension is width for lying down or getting up movement.
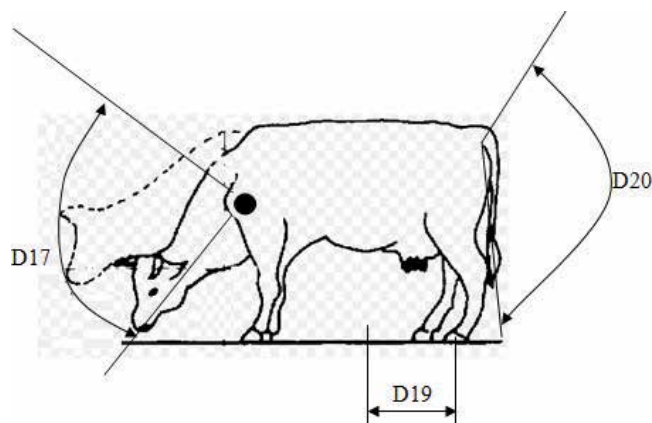
Figure 3. Dairy cattle dynamic zoometry dimensions (D_{17}, D_{19}, D_{20}).

To increase comfort during rising or lie down movement, the resting area must provide cattle with the easy movement for vertical, forward, and lateral movement without obstruction, injury, or fear. A rising motion includes the freedom to lunge forward, bob the head up or down, and stride forward. Resting motion also includes the freedom to lunge forward and bob the head. Each time the cow lies down, a cow puts about two-thirds of body weight on its front knees. Then the knees drop freely to the floor from a height of 20 to 30 centimeter. Therefore it is very important to provide best-quality bedding; as consequence, the cow can painlessly lie down at any time. The easy method to know the comfort level is to look at and check how fast a cow lies down in a cubicle.

3. Back propagation neural network (BPNN)

A neural network can be described as a *black box* that knows how to process input system to create useful outputs. Neural network is defined as "interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on the animal neuron [29]. The processing ability of the network is stored in the inter-unit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns." The NN calculation is very complex and difficult to understand by using a mathematical model. Neural network copied the working system of the biological nervous system as an example for the brain to process the information. Other experts define the neural network (NN) as a powerful data modeling tool capable to capture and represent complex input/output relationships involving many factors [30].

The most common neural network model is the multilayer perceptron (MLP) which contains three layers: input layer, hidden layer, and an output layer. This type of neural network is known as a supervised network because it requires the desired output in order to learn. The goal of this type of network is to create a model that correctly maps the input to the output using historical data; therefore, the model can then be utilized to produce an output when the desired output is unknown. The MLP and many other neural networks learn using an algorithm called "back propagation" [31]. The goal of a back propagation neural network (BPNN) is to minimize the error which in the project is shown as a mean square error (MSE). With each presentation, the output of the neural network is compared to the desired output, and an error is computed. This error is then fed back (back propagated) to the neural network and utilized to adjust the weights such that the error decreases with each iteration and the neural model gets increasingly closer to producing the desired output. This process is known as "training." **Figure 4** describes how the BPNN system works to minimize the gap between target and output by adjusting network weight.

BPNN works in three parts of a database called the set of training data, set of cross validation data, and set of testing data. *Training data* is utilized in the neural network to learn the databases' correlation and its function as an input data. *Cross validation data* is utilized to evaluate the performance of the learning process to avoid over-training. *Testing data* is utilized to evaluate the performance of the training when it is complete. Production input data was fed into the trained neural network to produce an output. There is no evidence of references
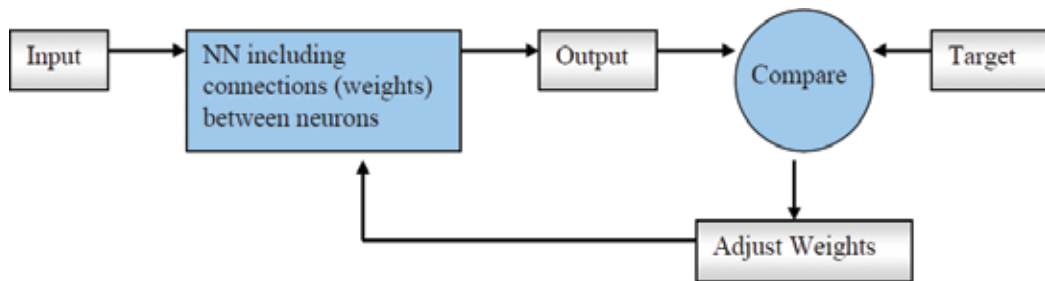
Figure 4. Description of BPNN training the data sets [29].

which explained how to divide the composition of training data, for cross validation and testing. The training data is required for a neural network to predict aerodynamic coefficient [32]. The paper shows manual NN training comparisons based on different transfer functions and training datasets. It is noted that dataset is an important part to obtain a better MSE performance. Commonly, the training data is >60%, the cross validation is ≈15%, and testing data is ≈10%.

Structure of BPNN model is a key performance to accelerate for reducing the gap between NN output and target during training the system. The BPNN structure contains a number of neurons, transfer function, and a number of hidden layers. The common method to create the BPNN structure is by the trial-and-error method. The method will be time-consuming during the training procedure. The optimal design of neural network using the Taguchi Method [33]. Moreover, we can use taguchi method to optimize the structure of BPNN for a limited amount of data [34]. Genetic algorithm (GA) is utilized to optimize the adjustment of the weight values during the training process.

Figure 5 chronologically explains the interconnection between BPNN and GA applications to adjust weight parameter in the quick propagation learning rule. The activities in **Figure 5** contain collection and preparation of data, define robust (optimum) NN architectures, initialize population (connection weights and thresholds), assign input and output values to NN, compute hidden layer values, compute output values, and compute fitness using MSE formula. To find the best NN weight, at the start of the genetic algorithm, an initial population of chromosomes is created. The gene values are assigned to the initial weights of the network, and the network is trained based on the back propagation neural network (BPNN) algorithm. The next step of the algorithm is the fitness values of all the chromosomes of population evaluated; the inverse of MSE is regarded as the fitness function of GA. The individual's genes are modified by crossover and mutation process. These operations result in a new-generation population of chromosomes. The generational process is repeated until convergence condition has been achieved. The weights of the BPNN network are created via a global optimization using GA, which increases the quality and the performance of the BPNN model. In the end, the neural network was trained with selected weight connection. According to the reference [36], the best BPNN structure is described as follows: one hidden layer, initial neuron = 17, tanh transfer function between input and hidden layer, the linear sigmoid transfer function between the hidden layer and output layer, quick propagation learning algorithm,
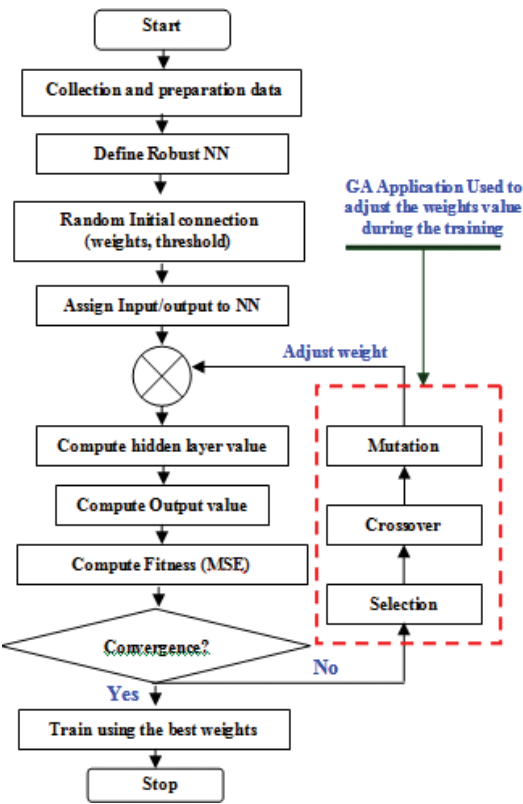
Figure 5. Neuro-genetic algorithm in BPNN development [35].

and 5000 epochs. GA parameters are selected as follows: the Roulette rule is employed to select the best chromosome based on proportionality to its rank, the initial values for learning rate and momentum are 0.5000 and 0.0166, number of population is 50 chromosomes and epoch number was 100 at maximum, initial network weight factor is 0.1074, mutation probability is 0.01, and heuristic crossover was utilized.

4. Construction of prediction model for cattle facilities using BPNN

The dairy cattle facilities should support cattle activities such as resting, drinking, eating, and milking. The facilities must guarantee the cattle will averse being stuck, injuries, and stress behaviors. The best facilities are indicated by cattle comfort level and increasing milk production. Moreover, the floor should not be warm and humid to reduce possible skin injuries and added thermal comfort. The other facility is a watering system; a watering cup should have an opening of at least 0.06 m^2, approximately 30 cm in diameters or similar opening size [13]. It is recommended that the main water supply is ring connected and that the water is under a constant pressure. The best method of watering is supplying via service pipe; it will make sure fresh water is always supplied with a minimum amount of dirt. The free stall is very

important for cattle to provide comfortable space for rest. Nigel B. Cook [9] research on free stall design for maximum cow comfort reported that 11.3 hours is needed for lying down in the stall and 2.9 hours for standing in the stall a day, in total 14.2 hours per day contact with free stall (around = 59.17%). The cow must have free stall design correctly sized as it correlates with milk production. A lot of farmers reduce stall length and width in order to save construction cost. It will reduce the level of cow comfort and milk production. Free stall should be designed correctly and maintained and should be sloped from front to back and provide a comfortable surface.

To control the steps of the research which are logically right, flowchart of developing and implementing the zoometry concept is presented as can be seen in **Figure 6**. According to the graph, the next step after collecting the static and dynamic data is doing the statistic test for checking the data quality. Manual measuring for dynamics data is time-consuming and costly and requires more energy, e.g., measuring the length for raising movement. The database construction is developed based on **Tables 2** and **3**, **Table 2** as input data and **Table 3** as desired data. As a result, BPNN can predict easily the dynamics zoometry dimension from any inputs of static zoometry dimension. To look for the best neural network structure, at the same time, the GA method is employed during NN training. BPNN module is ready to use to be part of designing the cattle comfort facilities, e.g., free stall, cattle house, etc.

Figure 7 describes BPNN training result using genetic algorithm (GA) optimization in one replication. According to the graph, the training process will stop in 49 generations with mean square error (MSE) = 0.0287. BPNN model is ready to predict any input data to determine output data (dynamics data). The BPNN model is very useful for the user conduct test to determine cow behavior correlated with cow dimensions in design process. The user can put any "normal value" of static cattle dimensions (D_1 to D_{16}) to predict dynamics cattle dimension (D_{17} to D_{23}). As example, the input values are $D_1 = 38$ cm, $D_2 = 78$ cm, $D_3 = 109$ cm, $D_4 = 157$ cm, $D_5 = 137$ cm, $D_6 = 51$ cm, $D_7 = 58$ cm, $D_8 = 41$ cm, $D_9 = 61$ cm, $D_{10} = 118$ cm, $D_{11} = 63$ cm, $D_{12} = 47$ cm, $D_{13} = 31$ cm, $D_{14} = 19$ cm, $D_{15} = 122$ cm, and $D_{16} = 14$ cm, and produced output values are $D_{17} = 52$ cm, $D_{18} = 214$ cm, $D_{19} = 54$ cm, $D_{20} = 121$ cm, $D_{21} = 312$ cm, $D_{22} = 301$ cm, and

No	Explanation	Dairy cattle dimension (cm)				
		1	2	—	499	500
D17	Vertical head movement (^{0})	50	55	—	43	45
D18	Horizontal head movement (^{0})	200	220	—	225	240
D19	Step walking (cm)	53	72	—	77	64
D20	Cow tail movement (^{0})	120	110	—	112	103
D21	Space for lying down (cm)	313	309	—	312	318
D22	Space for getting up (cm)	300	296	—	299	205
D23	Width space (cm)	132	130	—	120	118

Table 3. Data measurement results from 500 dairy cattle for dynamic dairy cattle dimension.
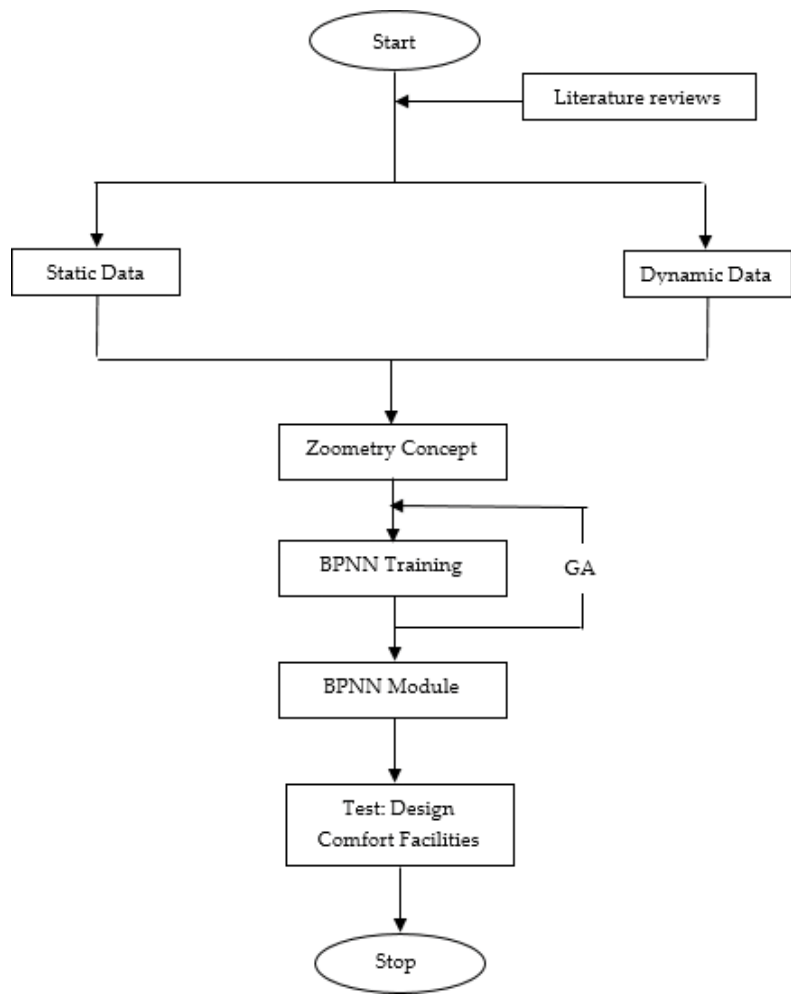
Figure 6. Flowchart of developing and implementing the zoometry concept using BPNN module.

D_{23} = 130 cm. The following stage involved implementing the zoometry concept and BPNN model to evaluate and redesign the cattle facilities. The first step of designing the cattle house is defined the house parameters which are described as follows:

- Length of cattle house (L)

Length of cattle house is defined as the total summation of length for lying down and length for getting up minus cattle length or $L = \bar{D}_{21} + \bar{D}_{22} - (\bar{D}_3 + \bar{D}_4)$. Based on data in **Table 1** for $\bar{D}_3 + \bar{D}_4$ and BPNN test for $\bar{D}_{21} + \bar{D}_{22}$, L has average 346.67 cm and deviation standard σ = 6.71 cm then by using percentile 95th will produce $L_{zoometry}$ = 346.67 cm + 1.64 × 6.71 cm = 357.67 cm.

- Width of cattle house (W)

The width of cattle house is defined as the space for lying down or getting up easily or defined in D_{23}. Based on the data in BPNN test, D_{23} has value = 125.96 cm and deviation
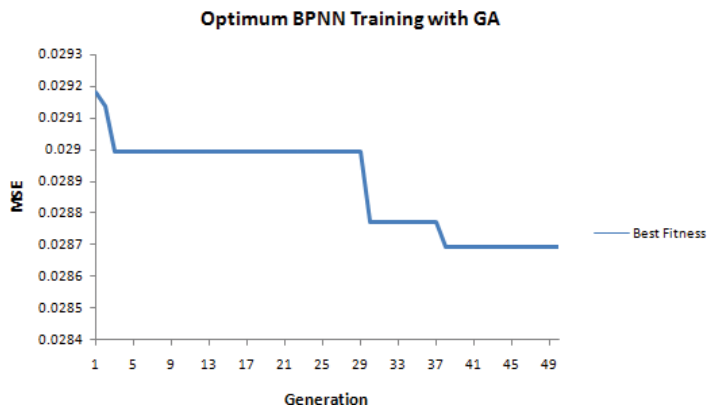
Figure 7. Static and dynamic zoometry training in BPNN-GA application.

standard $\sigma = 3.70$ cm and then using percentile 95th will produce $W_{zoometry} = 125.96$ cm + 1.64×3.70 cm = 132.03 cm

- Height of cattle house (H)

Height of cattle house is defined as the summation of $\bar{D}_2 + \bar{D}_9 + (\bar{D}_3 \times \tan(0.5 \times \bar{D}_{17}))$. Based on the data in **Table 1** and BPNN test, H has average 194.64 cm and deviation standard $\sigma = 6.49$ cm and then by using percentile 95th will produce $H_{zoometry} = 194.64$ cm + 1.64×6.49 cm = 205.28 cm.

Cattle house design is recommended based on the results of zoometry calculation of length, width, and height. The cattle house design should have minimum value of length = 357.67 cm, width = 132.03 cm, and height = 205.28 cm. The height dimension of cattle house must consider the other factors such as air circulation, lighting, and the other facilities especially in a tropical climate with higher level of temperature and relative humidity. It can increase the heat stress index, which finally reduces milk production. In the United States, climate change that makes higher temperature and humidity than normal is likely to affect milk production because dairy cows are sensitive to excessive temperature and humidity.

5. Conclusion(s)

The paper has successfully developed the concept of zoometry to describe the dimensions of dairy cattle to design facilities. There are two zoometry for static and dynamics condition with a total number of dimensions at 16 and 7. The chapter successfully presented the BPNN training as the complexity of dynamic data (cattle motion behavior) correlated with cattle dimension. The paper also describes how to implement the zoometry concept in order to develop cattle house design. This method could be used to design other facilities such as free stall, watering system, designing floor, and feeding rack.

Using 500 cattle data source, the zoometry concept still fluctuates despite success in homogeny test and data-sufficient test. As consequence, a huge number of dimension data is required to

obtain steady zoometry. The zoometry concept will be an important topic of research in the future correlated with cattle comfort and cattle productivity.

Author details

Sugiono Sugiono[1]*, Rudy Soenoko[2] and Rio Prasetyo Lukodono[1]

*Address all correspondence to: Sugiono_ub@ub.ac.id

1 Industrial Engineering, Universitas Brawijaya, Malang, Indonesia

2 Mechanical Engineering, Universitas Brawijaya, Malang, Indonesia

References

[1] Rozenberg S, Body JJ, Bruyère O, Bergmann P, Brandi ML, Cooper C, Reginster JY, et al. Effects of dairy products consumption on health: Benefits and beliefs—A Commentary from the Belgian Bone Club and the European Society for Clinical and Economic aspects of osteoporosis, osteoarthritis and musculoskeletal diseases. Calcified Tissue International. 2016. https://doi.org/10.1007/s00223-015-0062-x

[2] Choi IY, Lee P, Denney DR, Spaeth K, Nast O, Ptomey L, Sullivan DK, et al. Dairy intake is associated with brain glutathione concentration in older adults. American Journal of Clinical Nutrition. 2015;**101**(2):287-293. https://doi.org/10.3945/ajcn.114.096701

[3] Faghih A, Anoosheh M, Ahmadi F, Ghofranipoor F. The effect of boy students' participation on consumption of milk and dairy. Hormozgan Medical Journal. 2007;**10**(4):349-356

[4] Davoodi H, Esmaeili S, Mortazavian AM. Effects of milk and milk products consumption on cancer: A review. Comprehensive Reviews in Food Science and Food Safety. 2013;**12**(3):249-264. https://doi.org/10.1111/1541-4337.12011

[5] Flynn A, Cashman K. Nutritional aspects of minerals in bovine and human milks. In: Fox PF, editor. Advanced Dairy Chemistry, Vol. 3 Lactose, Water, Salts and Vitamins. 2nd ed. London: Chapman & Hall; 1997

[6] Fox PF, Mcsweeney PLH. Dairy Chemistry and Biochemistry. 1998;**1542**(9):478. https://doi.org/10.1007/978-3-319-14892-2

[7] Holt C. Effect of heating and cooling on the milk salts and their interaction with casein. In: Fox PF, editor. Heat Induced Changes in Milk. 2nd ed. Brussels: International Dairy Federation; 1995. pp. 159-178

[8] Öste R, Jägerstad M, Anderson I. Vitamins in milk and milk products. In: Fox PF, editor. Advanced Dairy Chemistry, Vol. 3 Lactose, Water, Salts and Vitamins. 2nd ed. London: Chapman & Hall; 1997

[9] Cook NB, Bennett TB, Nordlund KV. Monitoring indices of cow comfort in free-stall-Houtilized dairy herds. Journal of Dairy Science. 2005;**88**(11):3876-3885. https://doi.org/10.3168/jds.S0022-0302(05)73073-3

[10] Krawczel PD, Klaiber LB, Butzler RE, Klaiber LM, Dann HM, Mooney CS, Grant RJ. Short-term increases in stocking density affect the lying and social behavior, but not the productivity, of lactating Holstein dairy cows. Journal of Dairy Science. 2012;**95**(8):4298-4308. https://doi.org/10.3168/jds.2011-4687

[11] Vazquez OP, Smith TR. Factors affecting pasture intake and Total dry matter intake in grazing dairy cows. Journal of Dairy Science. 2000;**83**(10):2301-2309. https://doi.org/10.3168/jds.S0022-0302(00)75117-4

[12] Jim R. Dairy Facilities and Cow Comfort. Tulare, CA: Veterinary Medicine Teaching and Research Center; 2011

[13] Bridger R. Introduction to Egonomics. Engineering. Vol. 8. 2003. https://doi.org/10.4324/9780203426135

[14] International Organization for Standardization (ISO) (2008), Basic human body measurements for technological design (ISO 7250-1;2008)

[15] Taifa IW, Desai DA. Anthropometric measurements for ergonomic design of students' furniture in India. Engineering Science and Technology: An International Journal. 2017;**20**(1):232-239. https://doi.org/10.1016/j.jestch.2016.08.004

[16] Masson AE, Hignett S, Gyi DE. Anthropometric study to understand body size and shape for plus size people at work. Procedia Manufacturing. 2015;**3**:5647-5654. https://doi.org/10.1016/j.promfg.2015.07.776

[17] Pheasant S, Bodyspace Anthropometry, Ergonomics and the Design of Work Second Edition; 2003

[18] Zhang P, Qint S, Wright DK Novel method of capturing static and dynamic anthropometric data for home design, IEEE; 2005. https://doi.org/10.1109/EURCON.2005.1629990

[19] Kroemer KHE. Engineering anthropometry. Ergonomics. 1989;**32**(7):767-784. https://doi.org/10.1080/00140138908966841

[20] Grant R. Taking advantage of natural behavior improves dairy cow performance. In: Western Dairy Management Conference. 2007. pp. 1-13

[21] Marino L, Allen K. The psychology of cows. Animal Behavior and Cognition. 2017;**4**(44): 474-498. https://doi.org/10.26451/abc.04.04.06.2017

[22] Krawczel P, Grant R. Effects of cow comfort on milk quality, productivity, and behavior. 2013. http://articles.extension.org:80/pages/70107/effects-of-cow-comfort-on-milk-quality-productivity-and-behavior

[23] Dairy NZ. Dairy NZ Economic Survey 2007-08. In Report; 2009

[24] Anonymous, Interdisciplinary report. Housing Design for Cattle – Danish Recommendation. Third Edition 2001. The Danish Agricultural Advisory Center, Translated into English and Issued in 2002. 122 pp

[25] Phillips C. Cattle Behaviour and Welfare: 2nd Edition. 2007. https://doi.org/10.1002/9780470752418

[26] Currie WB. Structure and Function of Domestic Animals. CRC Press; 1995

[27] Benham PFJ. Synchronization of behaviour in grazing cattle. Applied Animal Ethology. 1982;**8**(4):403-404. https://doi.org/10.1016/0304-3762(82)90075-X

[28] Michael BJ, John SF, Joseph HP. Facility and Climate Effect on Dry Matter Intake of Dairy Cattle. 5th Western Dairy Management Conference, April 4-6. Las Vegas: Nevada; 2001

[29] Gurney K. An introduction to neural networks. Neurocomputing (Vol. 14). 1997. https://doi.org/10.1016/S0925-2312(96)00046-X

[30] Yu X, Efe MO, Kaynak O. A general backpropagation algorithm for feedforward neural networks learning. IEEE Transactions on Neural Networks/a Publication of the IEEE Neural Networks Council. 2002;**13**(1):251-254. https://doi.org/10.1109/72.977323

[31] Laurene F. Fundamentals of Neural Network. Prentice Hall International, Inc, USA: Florida Institute of Technology; 1994

[32] Rajkumar T, Bardina J. Training Data Requirement for a Neural Network to Predict Aerodynamic Coefficient. NASA Ames research center, Moffet Field, California, USA

[33] Khaw JFC, Lim BS, Lim LEN. Optimal design of neural networks using the Taguchi method. Neurocomputing. 1995;**7**(3):225-245. https://doi.org/10.1016/0925-2312(94)00013-I

[34] Sugiono, Wu MH, Oraifige I. Applying the design of experiment (DoF) to optimise the NN architecture in the car body design system. In: Proceedings of 2011 17th International Conference on Automation and Computing, ICAC 2011

[35] Sugiono, Wu MH, Oraifige I. Employ the Taguchi Method to Optimize BPNN's Architectures in Car Body Design System, American Journal of Computational and Applied Mathematics. 2012;**2**(4):140-151. https://doi.org/10.5923/j.ajcam.20120204.02

[36] Sugiono S, Soenoko R, Riawati L. Investigating the impact of physiological aspect on cow milk production using artificial intelligence. International Review of Mechanical Engineering. 2017;**11**(1). https://doi.org/10.15866/ireme.v11i1.9873

Edited by Marco Antonio Aceves-Fernandez

Artificial intelligence (AI) is taking an increasingly important role in our society. From cars, smartphones, airplanes, consumer applications, and even medical equipment, the impact of AI is changing the world around us. The ability of machines to demonstrate advanced cognitive skills in taking decisions, learn and perceive the environment, predict certain behavior, and process written or spoken languages, among other skills, makes this discipline of paramount importance in today's world. Although AI is changing the world for the better in many applications, it also comes with its challenges. This book encompasses many applications as well as new techniques, challenges, and opportunities in this fascinating area.

IntechOpen