IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 21, 2020, accepted February 6, 2020, date of publication February 12, 2020, date of current version March 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2973509

An Author Gender Detection Method Using Whale Optimization Algorithm and Artificial Neural Network

FATEMEH SAFARA[1], AMIN SALIH MOHAMMED[2,3], MOAYAD YOUSIF POTRUS[3], SAQIB ALI[4], QUAN THANH THO[5], ALIREZA SOURI[6], FERESHTEH JANENIA[1], AND MEHDI HOSSEINZADEH[7,8,9]

[1]Department of Computer Engineering, Islamshahr Branch, Islamic Azad University, Islamshahr 3314767653, Iran
[2]Department of Computer Engineering, Lebanese French University, Erbil 44001, Iraq
[3]Department of Software and Informatics Engineering, Salahaddin University-Erbil, Erbil 44001, Iraq
[4]Department of Information Systems, College of Economics and Political Science, Sultan Qaboos University, Muscat 123, Oman
[5]Department of Software Engineering, Ho Chi Minh City University of Technology–Vietnam National University, Ho Chi Minh City 76000, Vietnam
[6]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran 1477893855, Iran
[7]Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam
[8]Faculty of Electrical - Electronic Engineering, Duy Tan University, Da Nang 550000, Vietnam
[9]Health Management and Economics Research Center, Iran University of Medical Sciences, Tehran, 14496-14535, Iran

Corresponding author: Mehdi Hosseinzadeh (hosseinzadeh.m@iums.ac.ir)

ABSTRACT Author gender detection (AGD) is a serious and crucial issue in Internet security applications, in particular in email, messenger, and social network communications. Detecting the gender of communication partner helps preventing massive fraud and abuses happening through social media such as email, blogs, forums. Text and writings of people on the Internet have valuable information that can be used to identify the gender of an author. Machine learning and meta-heuristic algorithms are valuable techniques to extract hidden patterns useful for detecting gender of a text. In this paper, an artificial neural network (ANN) is employed as a classifier to detect the gender of an email author and the whale optimization algorithm (WOA) is used to find optimal weights and biases for improving the accuracy of the ANN classification. Through this combination of ANN and WOA an accuracy of 98%, precision of 97.16%, and recall of 99.67% were achieved, which indicates the superiority of the proposed method on Bayesian networks, regression, decision tree, support vector machine, and ANN examined.

INDEX TERMS Author gender detection, machine learning, artificial neural network, whale optimization algorithm.

I. INTRODUCTION

Over the last few years, the global Internet network has grown significantly, followed by a growing number of applications and communication networks such as social networking, messaging, and web sites [1]. In cyberspace and Internet, Information is known as a multi-dimensional phenomenon that can be presented in the form of multimedia, text, audio and image [2]. Text is one of the most common media of communication on the Internet, and most of the information sent to social networks [3], [4], forums and blogs are text [5]. Today, most of the services provided on the Internet can be used for fraud and abuse. Unfortunately, tracking offenders is difficult because different Internet services allow users to hide their age, gender, nationality, and location [6], [7].

One of the favorite topics to overcome the above problem is Author Gender Detection (AGD) of texts on the Internet and cyberspace [8], [9]. AGD has many usages in the web space, including the disclosure of the true identity of individuals, the prevention of scam, the analysis of customer opinions by gender, more sales and marketing, effective communication on the cyberspace. AGD can determine the level of honesty of individuals in social networks and identify those who have forged their identity. To analyze the behavior of people using the Internet, it is very important to recognize their gender. For example, you can determine what percentage of people that refer to a website are male and what percentage are female; then business service providers can figure out which of their

The associate editor coordinating the review of this manuscript and approving it for publication was K. C. Santosh.

products attracted the most attention for women or men. They can design their sale plan according to this information, and increase their sales through precise marketing.

This paper presents a hybrid algorithm to detect author gender specification from hidden patterns in the context of emails. The proposed hybrid algorithm includes the Whale Optimization Algorithm (WOA) for selecting optimal parameters for Artificial Neural Network (ANN) to recognize the author's gender. The main contributions of this research are as follows:

- Applying the Multi-layer ANN to recognize the real identity of a person in social networks and Internet.
- Presenting the hybrid ANN-WOA algorithm to find the optimal identification of ANN parameters for predicting author's gender.
- Minimizing the error rate of AGD.
- Evaluating the accuracy, precision and recall of the proposed algorithm on the Enron dataset.

The rest of the paper is organized as follows. In Section 2 related researches on author gender identification and detection using data mining methods are presented. In Section 3, the materials and the proposed method are described. In Section 4, the experimental results are presented. Section 5 provides a discussion on the results and an analytical comparison of the proposed method. In Section 6, the conclusion and future work are outlined.

II. RELATED WORK

In this section, first, a number of research studies reported before for AGD in the social networks are presented. After that, researches employing WOA are reviewed. A summary is provided at the end of the section.

According to Cheng, et al. [10], by understanding the value of the information contained in the text published on social networks, such as Facebook and Twitter, messages, emails, blogs, we can propose AGD methods for better and safer communications. A new method proposed for recognizing the author of a text by analyzing the information published by the author. Deitrick *et al.* [11] applied the ANN to recognize the gender of an email's author. In their research, they used a dataset related to the author's email recognition, Enron Corpus, with a balanced ANN learning mechanism for AGD. Murugaboopathy *et al.* [12] used a machine learning technique to recognize the gender of the email's author. They extracted a set of related email features, and with the help of a Support Vector Machine (SVM), they developed a gender stratification classifier. In their proposed method, they used a set of attributes such as emotional, narrative, and morph-based words to support a vector-machine. The results of their experiments indicate that the SVM can well identify the gender of the text author with reasonable precision.

In the research reported by Sboev *et al.* [13], for the author's gender detection of a Russian text, deep-learning-based approaches were used. Alsmearat *et al.* [14] presented two techniques of word boxes and word morphology to select the gender of Arabic texts author in cyberspace. In addition, Topaloglu and Ekmekci [15] proposed a user gender detection model based on personal handwriting in graphology science. The decision tree was considered as the applied algorithm for evaluating the proposed model. In other research studies, Sboev *et al.* [16] examined Russian textual contents based on an automatic user gender identification method using a gradient boosting algorithm. For detecting data-driven gender recognition, Tsimperidis *et al.* [17] presented a dynamic feature selection approach to identify user gender of keystroke dynamics in digital forensics. The authors used a radial basis function algorithm to evaluate classification of the proposed approach and increased the accuracy of classification.

Rangel *et al.* [18] presented an overview of author profiling and deception detection in Arabic on news headlines and Twitter. They found out that SVM and deep learning were the most popular and successful classifiers in author profiling. Alvarez-Carmona *et al.* [19] evaluated author profiling detection based on textual and visual resources on Twitter to recognize some demographic aspects such as age and gender. The authors applied a dynamic feature selection on images posted on Twitter. SVM was applied to classify the posts. After that, they evaluated the proposed author profiling detection approach and achieved acceptable accuracy.

Vogel and Jiang [20] proposed a system on Twitter data for author profiling on the bot and gender identification with the use of SVM. In addition, Customer profiling is one of the research areas that gender detection has an important role. Hirt *et al.* [21] proposed a meta-classifier for customer profiling composing of three classifiers: name classifier, text classifier and image classifier. They improved the knowledge used in their proposed method by a cognitive method that allows the integration of current, as well as emerging customer profiling classifiers to enhance the prediction performance.

Vicente *et al.* [22] presented a combined classifier to detect the gender of English and Portuguese Tweets. A combined classifier consists of Multinomial Naive Bayes, Logistic Regression, and SVM. The accuracy of 93.2% and 96.9% were achieved in English and Portuguese using a combined classifier, respectively. Qian [23] presented research on the difference between males and females in terms of gender stereotypes. They found out that female's writing has fewer stereotypes than that of male's writing.

Evolutionary and meta-heuristic algorithms are usually used to improve the results of a classier. WOA is a relatively new evolutionary algorithm that gained researchers' attention for enhancing their classification results. WOA is used for feature selection [24]. Two approaches are followed: First, Tournament and Roulette Wheel selection mechanisms and second, crossover and mutation operators are used to improve the manipulation of the WOA. Also, Saidala and Devarakonda [25] proposed a feature selection method based on WOA. WOA is also used in the cloud computing research environment. A conceptual framework is proposed by [26] to solve scheduling problems of multi-objective

virtual machines. They used WOA and presented a problem formulation for the framework to achieve multi-objective functions.

Table 1 presents a side-by-side comparison for a number of existing related works on gender detection using data mining methods. The existing case studies and applied classification algorithms are compared in the table.

III. MATERIALSAND METHODS

In this section, the dataset used for this study is introduced first. Then the preliminaries about WOA and the proposed method are presented. We used WOA as an algorithm for minimizing the output error of an ANN. The weights and the biases of ANN to create a model must be well-defined so that output error of the classification model is minimized. One of the methods for improving the output of ANN is the use of meta-heuristic or evolutionary algorithms. In this research, WOA is used for the improvement.

A. MATERIALS

A large-scale public email dataset is not readily available for research in the field of gender detection. Enron dataset is a public email dataset that most of the researches on email data has been conducted on it [27]. This dataset was originally made public and posted to the web by the Federal Energy Regulatory Commission (FERC). The size of the dataset is 0.5M messages. Each record composed of 48 linguistic features. The features are divided into four categories: character-based features, word-based features, syntax-based features, and structure-based features. Each category has its own features as explained in the followings [28]:

- Character-based features such as total number of letters, total number of lower cases, total number of characters in a word, the total number of upper cases.
- Word-based features such as total number of words, average length per word, words longer than 6 characters, vocabulary richness.
- Syntax-based features or total number of each special characters such as total number of single quotes, total number of periods, total number of commas, total number of colons
- Structure-based features such as total number of sentences, total number of lines, total number of paragraphs, average number of sentences per paragraph.
- Function words such as total number of pronoun words, such as auxiliary-verbs, total number of article words.

B. WHALE OPTIMIZATION ALGORITHM (WOA)

Whales are impressive creatures. They are considered the largest creatures in the world. An adult whale can be up to 30 meters long and 180 tons weight. Humpback whales are a special type of whales [29]. The most stimulating thing about this type of whales is their distinctive hunting method. These hunting manners are called bubble-net feeding method. Humpback whales are favor to hunt krill or small fishes

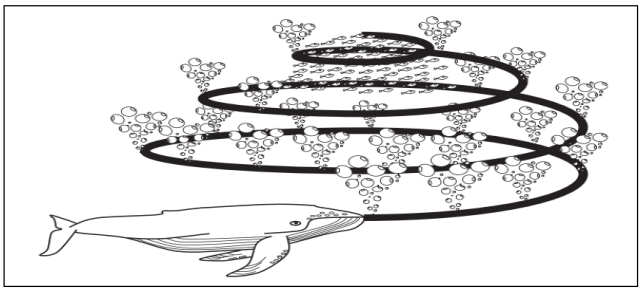

FIGURE 1. Bubble-net feeding behavior of humpback whales [33].

near to the water surface. This hunting is done by creating typical bubbles along a circular or spiral shaped path as illustrated in Figure 1 [30]. Before 2011, this behavior was only inspected based on the observation from the surface. Goldbogen *et al.* [31] inspected this behavior by employing tag sensors. They found two tactics related to bubble and named them 'upward-spirals' and 'double-loops' behaviors [30], [32]. Bubble-net feeding is an exclusive behavior that can only be seen in humpback whales.

In order to analyze and mathematically model the Humpback whale hunting behavior three phases could be considered: 1) the phase of finding prey and encircle them, 2) bubble-net attacking phase called exploitation phase, and 3) exploration phase. In the first phase, Humpback whales can spot the location of prey and surround them. The situation of the ideal design in the search space is not identified from before, the WOA algorithm assumes that the current optimal probable solution is the target prey or is close to it. After identifying the best search agent, the reminder of search agents would update their positions according to the position of the best search agent. This behavior is expressed by the following equations [29]

$$\vec{D} = \left|\vec{C}.\vec{X}^*(t) - \vec{X}(t)\right| \tag{1}$$

$$\vec{X}(t+1) = \vec{X}^*(t) - \vec{A}\cdot\vec{D} \tag{2}$$

where t is the current iteration, $\vec{A}$ and $\vec{C}$ are coefficient vectors, X^* is the position vector of the best solution found so far, $\vec{X}$ is the position vector, $|\ |$ is the absolute value, and $\cdot$ is an element-by-element multiplication. Provided a better solution is found, X^* should be updated in each iteration. The vectors $\vec{A}$ and $\vec{C}$ are obtained as follows

$$\vec{A} = 2\vec{a}\cdot\vec{r} - \vec{a} \tag{3}$$

$$\vec{C} = 2\vec{r} \tag{4}$$

where $\vec{a}$ is linearly decreased from 2 to 0 over the sequence of iterations, and $\vec{r}$ is a random vector in [0,1] [34].

In the second phase, which is the bubble-net attacking or exploitation phase, the circle made around the prey would be smaller and smaller. This would be happened with decreasing $\vec{a}$. Then, the distance between the whale and the prey is calculated, and according to the distance obtained the spiral position of the whale would be updated. Equation 6 is used

TABLE 1. Comparison of related works on the author gender detection.

Reference	Data source	Language	Reported Results	Applied algorithm/technique
Cheng, et al. [10]	Enron email dataset	English	85.1	SVM, Bayesian logistic regression and AdaBoost decision tree
Deitrick, et al. [11]	Enron email dataset	English	95	Winnow Neural Network
Murugaboopathy, et al. [12]	Enron email dataset	English	83	SVM and Bayesian Logistic Regression
Sboev, et al. [13]	Russian texts	Russian	86	Convolutional neural network
Alsmearat, et al. [14]	Arabic texts	Arabic	80.4	Bag-Of-Words
Topaloglu and Ekmekci [15]	Personal handwriting	English	93.75	Decision tree
Sboev, et al. [16]	Russian textual contents	Russian	Not reported	Gradient Boosting
Tsimperidis, et al. [17]	Keystroke dynamics	English	94.2	Radial basis function network
Alvarez-Carmona, et al. [19]	Author profiling in Twitter	English/Spanish	95.7/75.7	SVM
Hirt, et al. [21]	German-speaking in Twitter	German	90.21	Meta-classifier (constructed on three base classifiers)
Vogel and Jiang [20]	Twitter	English/Spanish	92/91	SVM
Vicente, et al. [22]	Twitter	English/Portuguese	93.2/96.9	Multinomial Naive Bayes
Mafarja and Mirjalili [24]	Eighteen UCI benchmark dataset	English	99.14	Particle Swarm Optimization (PSO), Genetic Algorithm(GA), the Ant Lion Optimizer (ALO)
Saidala and Devarakonda [25]	Enron email dataset	English	98.50	SVM and WOA

to calculate the distance and update the position:

$$\vec{X}(t+1) = \vec{D'} e^{bt} \cos(2\pi t) + \vec{X^*}(t) \quad (5)$$

$$\vec{D'} = \left| \vec{X^*}(t) - \vec{X}(t) \right| \quad (6)$$

where $\vec{D'}$ is the distance of the best solution obtained, and t is a random number in the range of $[-1, 1]$.

The third phase, the exploration phase, involves the refinement of the solutions found from the previous phase as stated in the following equations:

$$\vec{D} = \left| \vec{C} \overrightarrow{X_{rand}} - \vec{X} \right| \quad (7)$$

$$\vec{X}(t+1) = \overrightarrow{X_{rand}} - \vec{A} \cdot \vec{D} \quad (8)$$

where $\overrightarrow{X_{rand}}$ is the random position vector, and $\vec{X}(t+1)$ is the position vector found in the exploration phase.

C. PROPOSED METHOD

AGD is, a double class classification problem, and it is assumed that each data is in a male or female class. If a sample is male or female, the class number is zero or one, respectively. If a sample is male and it has classified in the female class, an error is appeared [35]. We are looking for a method to minimize these errors. Our proposed method for AGD is a hybrid method based on a multi-layer ANN and WOA optimization algorithm, which is illustrated in Figure 2.

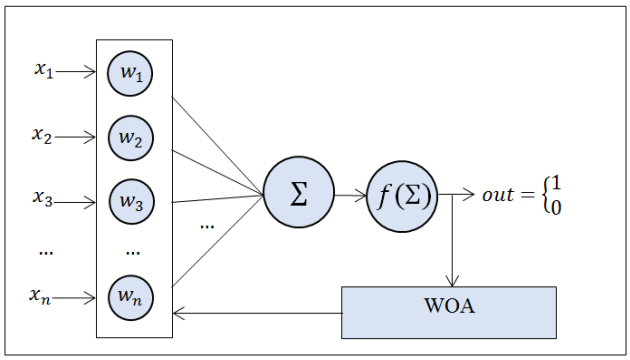

FIGURE 2. The usage of WOA to select optimal weights and biases for ANN.

An ANN is constructed with 48 input neurons as in the Enron dataset each record has 48 features, one output neuron to show the gender of an email author, and two hidden layers with 20 and 11 neurons in the first and the second layer with Sigmoid activation function.

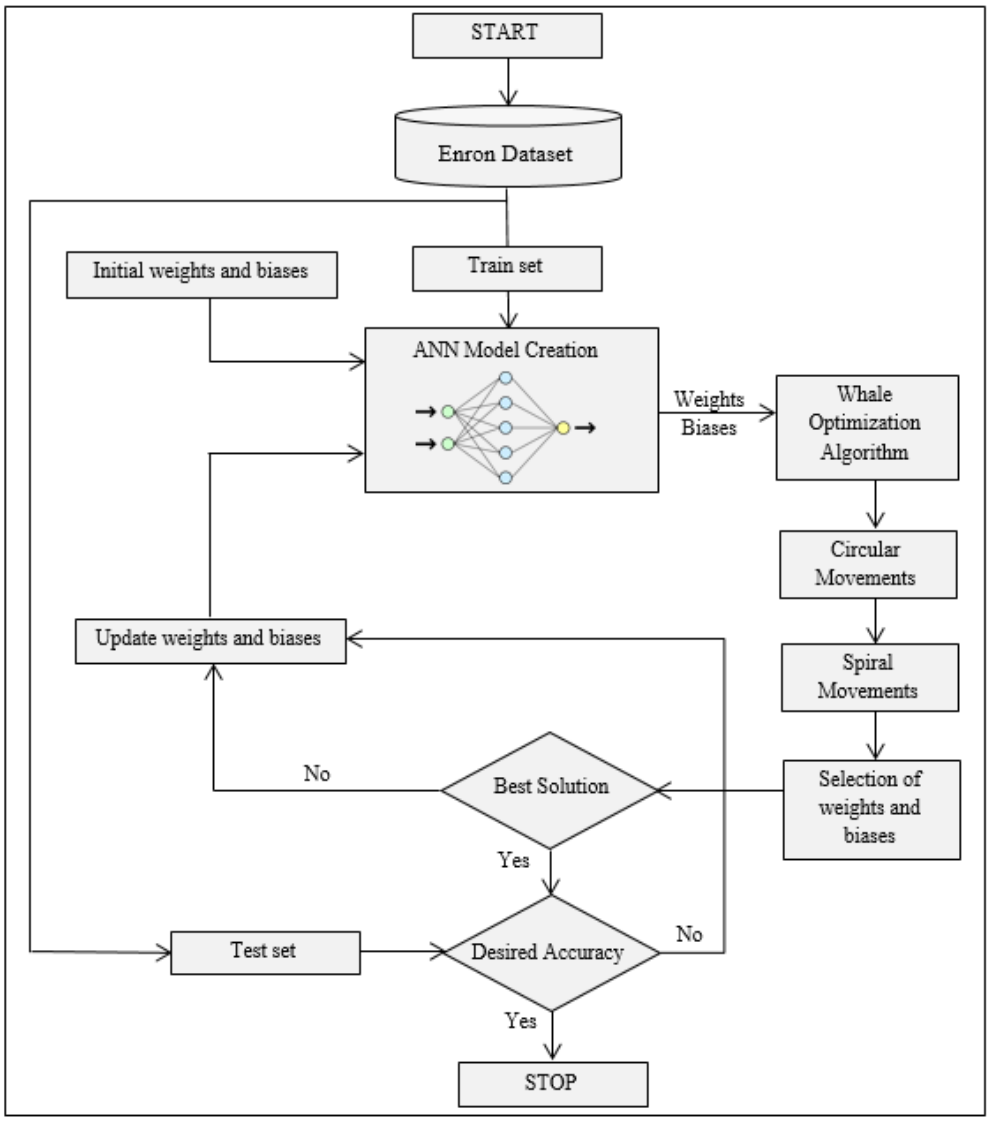

FIGURE 3. The flowchart of the proposed ANN-WOA method in detecting the gender of an email author.

In our proposed method the weights and biases of ANN are optimized through WOA, where in previous researches statistical methods were used for selecting weights and biases. With ANN-WOA, weights and biases are considered as members of the whale's population, and at each step of the algorithm, weights and biases are modified to reach to their optimal values. Then, the weights and biases are fed into ANN again. This process is repeating until the desired accuracy will be achieved.

WOA investigates the problem state to find the optimal solution. This investigation includes simultaneous local and global search in the problem space. This property has caused WOA to be able to solve optimization issues more precisely than other evolutionary algorithms such as genetics algorithm, particle swarm optimization, and bat algorithm.According to Figure 2, different weights and biases are evaluated by WOA in each iteration, and a set of weights and biases with the least classification error is considered as the optimal solution for constructing ANN in the next iteration. Therefore, WOA participates in the learning of ANN in order to improve its accuracy of classification.

The above hybrid method, ANN-WOA, is used to classify the Enron records for AGD. The flowchart of the proposed AGD is provided in Figure 3. The Enron dataset is randomly divided into a train set and a test set. 70% of the records are used in the training phase and the rest of 30% is utilized in the test phase. Training data is used to train multi-layer ANN and test data is used to evaluate the model constructed in the training phase. In ANN training phase, the ANN output error is minimized by a WOA. An arrow between ANN and whale shows that there is two-way communication between these two techniques to look for the best weights and biases with

maximum accuracy. The role of the WOA here is to help the ANN to minimize the gender detection error.

D. EVALUATION PARAMETERS

The proposed method is evaluated through four parameters including Mean Square Error (MSE) rate, accuracy, precision, and recall. The parameters are formulated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \grave{Y}_i\right)^2 \tag{9}$$

where n is the number of data points, $\grave{Y}_i$ is the observed output and Y_i is the predicted output.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FP} \tag{12}$$

where TP is true positive rate, TN is true negative rate, FP is false positive rate, and FN is false negative rate of classification results. TP is the rate of records which are male and correctly classified as male. TN is the rate of records that are female and correctly classified as female. FP is the rate of records which are male, however incorrectly classified as female, and TN is the rate of female that are incorrectly classified as male.

IV. EXPERIMENTAL RESULTS

To implement the proposed method, MATLAB programming environment has been used in this research. For implementation, 70% of the dataset is used for training, and the remaining 30% is used to test and validation of the proposed method. There are two classes that are respectively zero and one indicating male and female classes. To evaluate the proposed method, classification criteria including MSE, accuracy, precision, and recall were used.

The size of the initial population of whales is equal to 5 and the number of iterations of the algorithm is equal to 20. From Figure 4, it is could be observed that the average error of detecting male and female samples in WOA, is constantly reduced as the number of iterations increased.

The evaluation is repeated with whale's initial population of 10 and the number of iterations of the algorithm is equal to 20. As shown in Figure 5, the average error of the detecting male and female samples in WOA constantly reduced. However, the reduction is started in earlier iterations.

Figure 6 presents the MSE reduction in the case of 20 nodes as the whale's initial population 20 iterations of the algorithm. In this specification of MSE reduces considerably even in the 2nd iteration.

The evaluation is repeated with whale's initial population of 20 and the number of iterations of the algorithm is equal to 30. As shown in Figure 7, the average error of the detecting male and female samples in WOA starts at 0.217 which is the minimum value between four evaluations illustrated in Figures 4, 5, 6, and 7.

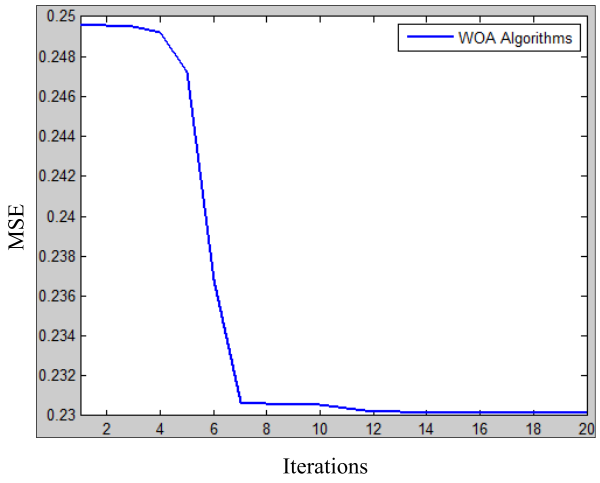

FIGURE 4. MSE of the proposed method with 5 nodes and 20 iterations.

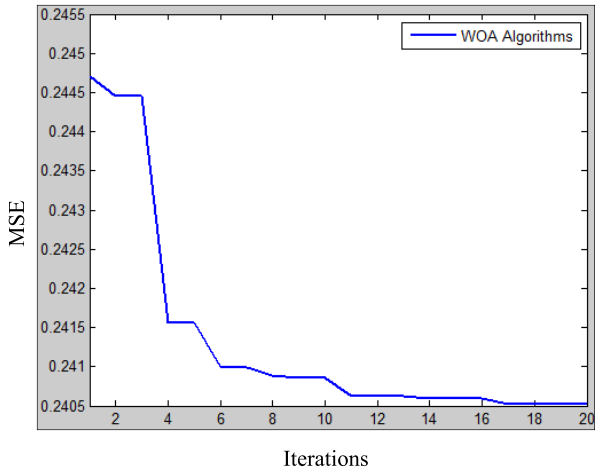

FIGURE 5. MSE of the proposed method with 10 nodes and 20 iterations.

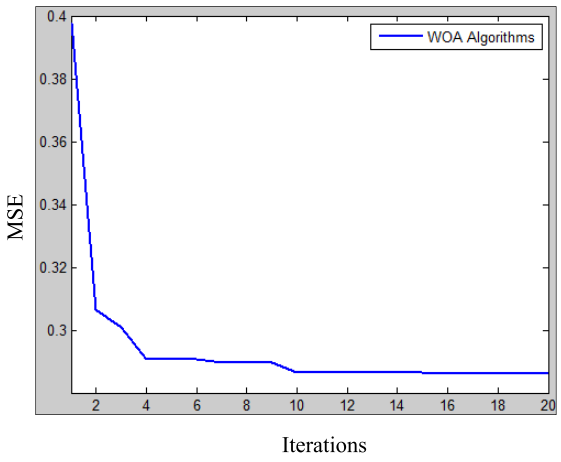

FIGURE 6. MSE of the proposed method with 20 nodes and 20 iterations.

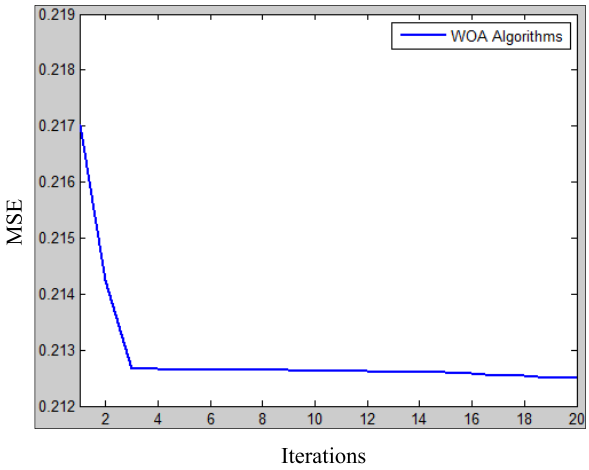

FIGURE 7. MSE of the proposed method with 20 nodes and 30 iterations.

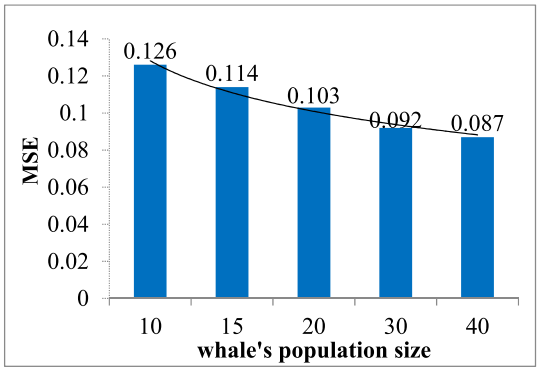

FIGURE 8. MSE reduction of the proposed AGD algorithm in terms of the initial whales' population in 30 iterations.

Error reduction process in the proposed method diagram shows that WOA has been able to reduce the amount of ANN error to distinguish AGD, as shown in Figures 4, 5, 6 and 7 with 5, 10 and 20 nodes in 30 iterations.

The examination of WOA with the same test conditions is repeated with the different number of nodes (10, 15, 20, 30 and 40 population sizes) in 30, 50 and 100 iterations. As an example, Figures 8, 9 and 10 illustrate the mean square error of the proposed method with 20 nodes and different population sizes in 30, 50 and 100 iterations.

Figure 8 indicates that, with 10 nodes and 30 iterations MSE is 0.126, however as the whale's population size increases to 40, MSE decreases to 0.087.

In Figure 9, as whale's population is increased from 10 to 40, the MSE decreased from 0.12 to 0.067. The difference between Figure 8 and Figure 9, indicates that increasing the number of iteration from 30 to 50 caused a reduction in MSE from 0.087 to 0.067.

Reduction of the MSE continues in Figure 10 where with the increase of whale's population size from 10 to 40, the MSE decreases from 0.1 in Figure 8 to 0.037 in Figure 10. Overall, MSE deceased from 0.126 to 0.036 as the result of increasing the number of iterations from 30 to 100.

Reducing the ANN classification error throughout the iterations indicates that WOA was successful to set weights

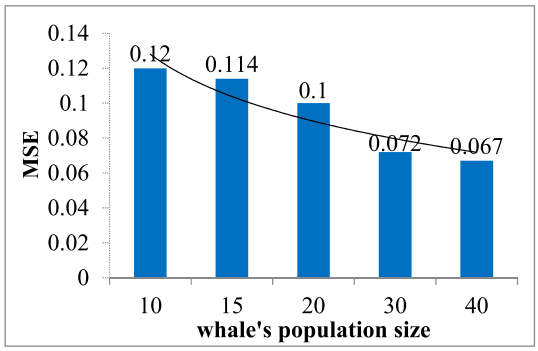

FIGURE 9. MSE reduction of the proposed AGD algorithm in terms of the initial whales' population in 50 iterations.

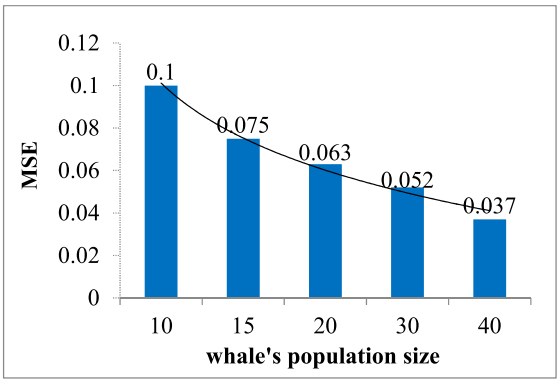

FIGURE 10. MSE reduction of the proposed AGD algorithm in terms of the initial whales' population in 100 iterations.

and biases optimally. Accuracy, precision and recall of the proposed method are measured by applying the proposed method on the test dataset as well. The accuracy, precision and recall of the proposed method for author gender detection are 93.20%, 93.24% and 93.18%, respectively, illustrated in Figure 11, Figure 12 and Figure 13, discussed in the next section.

V. DISCUSSION

To consider whales' population size and the number of iterations are two effective factors on meta-heuristic algorithms, experiments with the same condition has been performed. In the first case, the size of the population is 10, and in the latter case, this value has been increased to 20 to determine the effect of this parameter on MSE of the gender detection of the text author.

It is observed that increasing the size of the initial whales' population causes the search space to be larger, and the chance to find optimal weights and bias to decrease the ANN error increased. In fact, the increase in the size of the whales' population is caused that more weights and biases to be considered in each iteration and resulted in selecting optimal weights and bias. Therefore, it can be concluded that an increase in the size of the initial whales' population leads to a decrease in the error rate and increases the accuracy of the proposed system. Another influential parameter in the

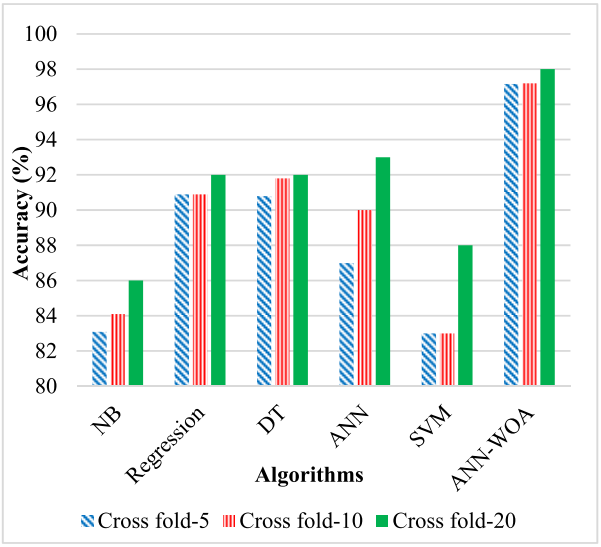

FIGURE 11. Comparison of the accuracy of the existing data mining methods in different cross folds.

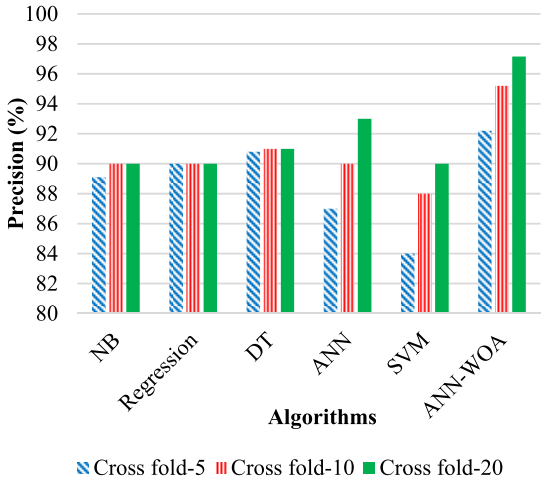

FIGURE 12. Comparison of the precision of the existing data mining methods in different cross folds.

proposed ANN-WOA is the number of iterations that causes members of the population to have more opportunities to converge and to find optimal thresholds and weight.

The results of our experiments show that WOA is well merged with ANN to find optimal weights and biases. In addition, increasing the size of the initial whales' population and the number of repetitions is an important factor in increasing the accuracy of the proposed method, so that increasing population size from 10 to 40 could reduce the error of the ANN-WOA diagnosis.

To compare and evaluate the proposed method, data-mining techniques such as Bayesian network, Regression, Decision tree (DT), ANN, and SVM were examined on the same dataset. Precisions and recalls of the methods and proposed ANN-WOA method are compared in different cross folds 5, 10 and 20. Graphic comparison of the proposed

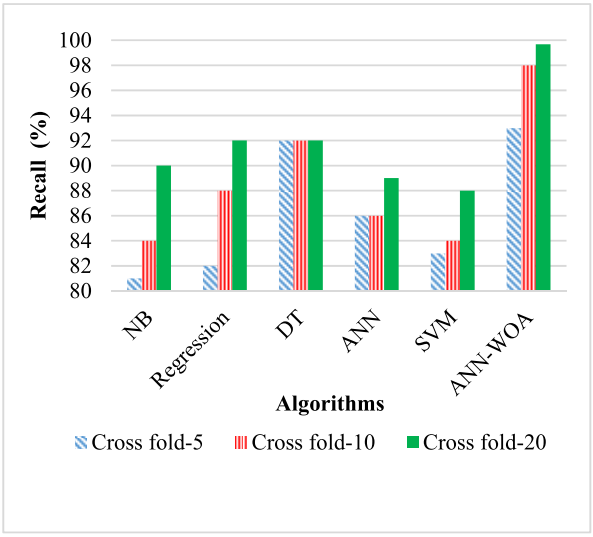

FIGURE 13. Comparison of the recall of the existing data mining methods in different cross folds.

method with the mentioned classifiers, in terms of accuracy of the ANN-WOA in author's gender identification, shows the proposed method detects the author's gender more accurately than the Bayesian network, Regression, DT, ANN and SVM, as illustrated in Figure 11.

Also, the precision of the ANN-WOA in author's gender identification is higher than the Bayesian network, Regression, DT, ANN and SVM, as illustrated in Figure 12 in different cross folds.

Figure 13 illustrates the recall value for existing classification algorithms. The recall factor in the proposed hybrid algorithm has the highest percentage for detecting true genders based on applied different cross folds.

VI. CONCLUSION

Reducing the output error of the ANN in choosing the text writer in terms of gender is an optimization problem. WOA is a meta-heuristic optimization algorithm with a collective and group intelligence approach. It is has been modeled from the whales' hunting behavior by creating bubbles. The results of the implementation of the proposed method show that the error rate of multilayer ANN along with the WOA algorithm has a significant decrease, and this reduction indicates the optimal determination of the weight and biases of the neural network through WOA. In addition, our experiments showed that the magnitude of the final error on the detecting of the gender of the male from the woman depends on the size of the initial whales' population and the number of iterations in different cross folds. For example, increasing the size of the population from 10 to 40 could reduce the gender diagnostic error of the text writer by about 25%.

The proposed method is compared with common machine learning techniques such as Bayesian network, Regression, DT and Multi-layer ANN and SVM; the accuracy of the proposed method is more than that of the machine learning methods mentioned. Our tests also showed that the

accuracy, precision, and recall of the proposed method are 98%, 97.16%, and 99.67% respectively.

One of the problems with the proposed method was the execution time of ANN itself and then in combination with WOA. It could be reduced in future works. Moreover, the volume of text data is increasing every day, therefore, more powerful data mining techniques are necessary for processing big data, in particular in cloud and fog environments. Other datasets than Enron could be examined as well in order to show the effectiveness of the proposed method for different email datasets.

ACKNOWLEDGMENT

This paper derives from the Research Project with code 98-2-37-15607 and Approval ID IR.IUMS.REC.1398.798.

REFERENCES

[1] L. Liu, X. Bai, and Z. Jiang, "The generic technology identification of saline–alkali land management and improvement based on social network analysis," *Cluster Comput.*, vol. 22, no. S6, pp. 13167–13176, Nov. 2019.
[2] D. T. Ramotsoela, G. P. Hancke, and A. M. Abu-Mahfouz, "Attack detection in water distribution systems using machine learning," *Hum.-Centric Comput. Inf. Sci.*, vol. 9, p. 13, Apr. 12 2019.
[3] L. Liang and X. Qin, "Research on consumers online shopping decision-making and recommendation of commodity based on social media network," *Cluster Comput.*, vol. 22, no. S3, pp. 6529–6539, May 2019.
[4] S. Razzaghzadeh, A. H. Navin, A. M. Rahmani, and M. Hosseinzadeh, "Probabilistic modeling to achieve load balancing in expert clouds," *Ad Hoc Netw.*, vol. 59, pp. 12–23, May 2017.
[5] L. Srinivasan and C. Nalini, "An improved framework for authorship identification in online messages," *Cluster Comput.*, vol. 22, no. S5, pp. 12101–12110, Sep. 2019.
[6] A. Souri, S. Hosseinpour, and A. M. Rahmani, "Personality classification based on profiles of social networks' users and the five-factor model of personality," *Hum.-Centric Comput. Inf. Sci.*, vol. 8, p. 24, Aug. 2018.
[7] M. R. Mesbahi, A. M. Rahmani, and M. Hosseinzadeh, "Reliability and high availability in cloud computing environments: A reference roadmap," *Hum.-Centric Comput. Inf. Sci.*, vol. 8, no. 1, p. 20, Dec. 2018.
[8] J. Soler-Company and L. Wanner, "On the role of syntactic dependencies and discourse relations for author and gender identification," *Pattern Recognit. Lett.*, vol. 105, pp. 87–95, Apr. 2018.
[9] A. Sboev, I. Moloshnikov, D. Gudovskikh, A. Selivanov, R. Rybka, and T. Litvinova, "Deep learning neural nets versus traditional machine learning in gender identification of authors of RusProfiling texts," *Procedia Comput. Sci.*, vol. 123, pp. 424–431, Jan. 2018.
[10] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi, "Author gender identification from text," *Digit. Invest.*, vol. 8, pp. 78–88, Jul. 2011.
[11] W. Deitrick, Z. Miller, B. Valyou, B. Dickinson, T. Munson, and W. Hu, "Author gender prediction in an email stream using neural networks," *J. Intell. Learn. Syst. Appl.*, vol. 4, no. 3, pp. 169–175, 2012.
[12] N. Cheng, R. Chandramouli, K. P. Subbalakshmi, "Author gender identification from text," *Digit. Invest.*, vol. 8, no. 1, pp. 78–88, 2011.
[13] A. Sboev, T. Litvinova, D. Gudovskikh, R. Rybka, and I. Moloshnikov, "Machine learning models of text categorization by author gender using topic-independent features," *Procedia Comput. Sci.*, vol. 101, pp. 135–142, Jan. 2016.
[14] K. Alsmearat, M. Al-Ayyoub, R. Al-Shalabi, and G. Kanaan, "Author gender identification from Arabic text," *J. Inf. Secur. Appl.*, vol. 35, pp. 85–95, Aug. 2017.
[15] M. Topaloglu and S. Ekmekci, "Gender detection and identifying one's handwriting with handwriting analysis," *Expert Syst. Appl.*, vol. 79, pp. 236–243, Aug. 2017.
[16] A. Sboev, I. Moloshnikov, D. Gudovskikh, A. Selivanov, R. Rybka, and T. Litvinova, "Automatic gender identification of author of Russian text by machine learning and neural net algorithms in case of gender deception," *Procedia Comput. Sci.*, vol. 123, pp. 417–423, Jan. 2018.
[17] I. Tsimperidis, A. Arampatzis, and A. Karakos, "Keystroke dynamics features for gender recognition," *Digit. Invest.*, vol. 24, pp. 4–10, Mar. 2018.
[18] F. Rangel, P. Rosso, A. Charfi, W. Zaghouani, B. Ghanem, and J. Snchez-Junquera, "Overview of the track on author profiling and deception detection in arabic," in *Proc. Working Notes Forum Inf. Retr. Eval. (FIRE)*, Kolkata, India, 2019, pp. 1–14.
[19] M. A. Alvarez-Carmona, L. Pellegrin, M. Montes-y-Gómez, F. Sánchez-Vega, H. J. Escalante, A. P. López-Monroy, L. Villaseñor-Pineda, and E. Villatoro-Tello, "A visual approach for age and gender identification on Twitter," *J. Intell. Fuzzy Syst.*, vol. 34, no. 5, pp. 3133–3145, May 2018.
[20] I. Vogel and P. Jiang, "Bot and gender identication in Twitter using word and character N-grams," Lugano, Switzerland, Tech. Rep., Sep. 2019.
[21] R. Hirt, N. Kühl, and G. Satzger, "Cognitive computing for customer profiling: Meta classification for gender prediction," *Electron. Markets*, vol. 29, no. 1, pp. 93–106, Mar. 2019.
[22] M. Vicente, F. Batista, and J. P. Carvalho, "Gender detection of Twitter users based on multiple information sources," in *Interactions Between Computational Intelligence and Mathematics*. Cham, Switzerland: Springer, 2019, pp. 39–54.
[23] Y. Qian, "Gender stereotypes differ between male and female writings," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, Student Res. Workshop*, 2019, pp. 48–53.
[24] M. Mafarja and S. Mirjalili, "Whale optimization approaches for wrapper feature selection," *Appl. Soft Comput.*, vol. 62, pp. 441–453, Jan. 2018.
[25] R. K. Saidala and N. R. Devarakonda, "Bubble-net hunting strategy of whales based optimized feature selection for e-mail classification," in *Proc. 2nd Int. Conf. Converg. Technol. (ICT)*, Apr. 2017, pp. 626–631.
[26] N. Rana and M. S. A. Latiff, "A cloud-based conceptual framework for multi-objective virtual machine scheduling using whale optimization algorithm," *Int. J. Innov. Comput.*, vol. 8, no. 3, pp. 1–6, 2018.
[27] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *Proc. Eur. Conf. Mach. Learn.*, 2004, pp. 217–226.
[28] A. Souri and R. Hosseini, "A state-of-the-art survey of malware detection approaches using data mining techniques," *Hum.-Centric Comput. Inf. Sci.*, vol. 8, no. 1, p. 3, Dec. 2018.
[29] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Adv. Eng. Softw.*, vol. 95, pp. 51–67, May 2016.
[30] M. M. Mafarja and S. Mirjalili, "Hybrid whale optimization algorithm with simulated annealing for feature selection," *Neurocomputing*, vol. 260, pp. 302–312, Oct. 2017.
[31] J. A. Goldbogen, J. Calambokidis, E. Oleson, J. Potvin, N. D. Pyenson, G. Schorr, and R. E. Shadwick, "Mechanics, hydrodynamics and energetics of blue whale lunge feeding: Efficiency dependence on krill density," *J. Exp. Biol.*, vol. 214, no. 4, pp. 698–699, Feb. 2011.
[32] S. Mirjalili, "SCA: A sine cosine algorithm for solving optimization problems," *Knowl.-Based Syst.*, vol. 96, pp. 120–133, Mar. 2016.
[33] I. Aljarah, H. Faris, and S. Mirjalili, "Optimizing connection weights in neural networks using the whale optimization algorithm," *Soft Comput.*, vol. 22, no. 1, pp. 1–15, Jan. 2018.
[34] R. Barham and I. Aljarah, "Link prediction based on whale optimization algorithm," in *Proc. Int. Conf. New Trends Comput. Sci. (ICTCS)*, Oct. 2017, pp. 55–60.
[35] O. Giannakopoulos, N. Kalatzis, I. Roussaki, and S. Papavassiliou, "Gender recognition based on social networks for multimedia production," in *Proc. IEEE 13th Image, Video, Multidimensional Signal Process. Workshop (IVMSP)*, Jun. 2018, pp. 1–5.

FATEMEH SAFARA received the B.Sc. degree (Hons.) in applied mathematics from Islamic Azad University, the M.Sc. degree in data warehousing from the Tarbiat Modares University, Tehran, and the Ph.D. degree in biological signal processing from University Putra Malaysia, in 2014. She was a Faculty Member with the Iran Telecommunication Research Center, Information Technology Department, from 2000 to 2010, doing research on image mining and biometric signals. She joined Islamic Azad University, Islamshahr Branch, on 2010, where she is currently an Assistant Professor of computer engineering faculty. Her current research interests include signal processing and in particular biological signal processing, data mining, the Internet of Things, and cloud computing.

AMIN SALIH MOHAMMED received the bachelor's and master's degrees in engineering and the Ph.D. degree in computer engineering from the Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, as a tech-savvy. He is currently serving as the Vice President for Scientific Affairs of Lebanese French University, Erbil, Iraq. Before being promoted as the Vice President, he has served as the Dean of the College of Engineering and Computer Science and a Lecturer with Salahaddin University-Erbil. He is also having more than 15 years of teaching and research experience. He is also an Active Researcher and acted as a Resource Person for various workshops and faculty development programs organized by different institutions. His research fields are computer networks, wireless networks, and cloud computing. So far, he has published more than 40 research articles in various reputed international journals indexed in SCI and Scopus services and reputed conferences.

MOAYAD YOUSIF POTRUS received the B.Sc. degree in electrical engineering and the M.Sc. degree in computer engineering from the University of Baghdad, Iraq, in 1997 and 2000, respectively, and the Ph.D. degree in computer Engineering from University Sains Malaysia, in 2012. He is currently a full-time Associated Professor with the Department of Software and Informatics Engineering, Salahaddin University-Erbil, Iraq. His research interests are in the field of machine learning, pattern recognition, global optimization, and software testing.

SAQIB ALI received the B.S. degree in computer software engineering from SYICT University, Shenyang, China, and the M.Sc. degree in information systems and Ph.D. degree in computer science from La Trobe University, Australia. He is currently an Associate Professor with the Department of Information Systems, Sultan Qaboos University, Muscat, Sultanate of Oman. His research interest includes industrial informatics and cyber security for cyber physical systems and business processes.

QUAN THANH THO received the B.Eng. degree in information technology from HCMUT, in 1998, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2006. He is currently an Associate Professor with the Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Vietnam. His current research interests include formal methods, program analysis/verification, the semantic web, machine learning/data mining, and intelligent systems. He is also the Vice Dean of Academic Affairs of the Faculty.

ALIREZA SOURI received the B.S. degree in software engineering from the University College of Nabi Akram, Iran, and the M.Sc. and Ph.D. degrees in software engineering from the Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran. He is currently an Associate Editor of *Human-Centric Computing and Information Sciences* (Springer), *Cluster Computing* (Springer), and *IET Communications* (IEEE) journals. He has authored or coauthored more than 55 scientific articles. His research interests include formal specification and verification, model checking, fog and cloud computing, the Internet of Things, data mining, and social networks.

FERESHTEH JANENIA received the B.S. and M.Sc. degrees in computer engineering from Islamic Azad University. His research interests are in machine learning, the IoT, cloud computing, and distributed systems.

MEHDI HOSSEINZADEH received the B.S. degree in computer hardware engineering from Islamic Azad University, Dezfol Branch, Iran, in 2003, and the M.Sc. and Ph.D. degrees in computer system architecture from the Science and Research Branch, Islamic Azad University, Tehran, Iran, in 2005 and 2008, respectively. He is currently an Associate professor with the Iran University of Medical Sciences (IUMS), Tehran. His research interests include SDN, information technology, data mining, big data analytics, e-commerce, e-marketing, and social networks.