Int. J. Fuzzy Syst.
https://doi.org/10.1007/s40815-020-00828-7

Coronary Heart Disease Diagnosis Through Self-Organizing Map and Fuzzy Support Vector Machine with Incremental Updates

Mehrbakhsh Nilashi[1,2] · Hossein Ahmadi[3] · Azizah Abdul Manaf[4] · Tarik A. Rashid[5] · Sarminah Samad[6] · Leila Shahmoradi[7,8] · Nahla Aljojo[9] · Elnaz Akbari[10,11]

Received: 3 June 2019 / Revised: 2 January 2020 / Accepted: 20 February 2020

Abstract The trade-off between computation time and predictive accuracy is important in the design and implementation of clinical decision support systems. Machine learning techniques with incremental updates have proven its usefulness in analyzing large collection of medical datasets for diseases diagnosis. This research aims to develop a predictive method for heart disease diagnosis using machine learning techniques. To this end, the proposed method is developed by unsupervised and supervised learning techniques. In particular, this research relies on Principal Component Analysis (PCA), Self-Organizing Map, Fuzzy Support Vector Machine (Fuzzy SVM), and two imputation techniques for missing value imputation. Furthermore, we apply the incremental PCA and FSVM for incremental learning of the data to reduce the computation time of disease prediction. Our data analysis on two real-world datasets, Cleveland and Statlog, showed that the use of incremental Fuzzy SVM can significantly improve the

✉ Mehrbakhsh Nilashi
nilashi@tdtu.edu.vn

✉ Elnaz Akbari
elnazakbari@duytan.edu.vn

[1] Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Vietnam

[2] Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam

[3] Department of Information Technology, University of Human Development, Sulaymaniyah, Iraq

[4] Department of Cybersecurity, College of Computer Science and Engineering, University Of Jeddah, Jeddah, Saudi Arabia

[5] Computer Science and Engineering Department, University of Kurdistan Hewler, Erbil, Kurdistan, Iraq

[6] Department of Business Administration, College of Business and Administration, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

[7] Halal Research Center of IRI, FDA, Tehran, Iran

[8] Health Information Management Department, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran

[9] Department of Information System and Technology, College of Computer Science and Engineering, University Of Jeddah, Jeddah 23218, Saudi Arabia

[10] Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam

[11] Faculty of Information Technology, Duy Tan University, Da Nang 550000, Vietnam

Published online: 24 March 2020

accuracy of heart disease classification. The experimental results further revealed that the method is effective in reducing the computation time of disease diagnosis in relation to the non-incremental learning technique.

Keywords Fuzzy support vector machine · Coronary heart disease · Self-organizing map · Incremental learning · Prediction accuracy

1 Introduction

Coronary Heart Disease (CHD) is a major silent killer disease and leading cause of death globally [1, 2] with the largest disease burden worldwide [3], reported by the World Health Organization. Reported in the previous research, it has been predicted that heart disease can still be the leading global cause of death in 2030 [4]. According to the report by American Heart Association, over 7 million Americans have experienced a heart attack in their lifetime [5]. Because of its importance, many studies have been conducted to help the patients from this disease [6–8].

According to [9], "Clinical decision support systems (CDSS) are computer systems designed to impact clinician decision making about individual patients at the point in time that these decisions are made." Clinical Decision Support Systems [10] developed by machine learning techniques have played an important role in estimating the presence of coronary artery disease. These systems are promoted for their potential to enhance the quality and efficiency of healthcare in heart disease diagnosis systems [8, 11]. They mainly rely on supervised and unsupervised machine learning techniques. Clinical decision support systems take the advantages of unsupervised machine learning techniques in dimensionality reduction of data for diseases diagnosis. In fact, these techniques are basically used in data manipulation, data noise removal, data similarity calculation, and data clustering [12]. However, supervised machine learning techniques are adopted in the final stage of clinical decision support systems which is called prediction stage [13]. These types of learning techniques are selected based on a main criterion, their accuracy in the disease prediction [14]. Although there are many studies on heart disease prediction using a set real-world medical data [13, 15], these studies are solely developed for the accuracy improvement and the computation time of the disease prediction is ignored. In addition, these studies mainly focused on the predictive accuracy and the data time complexity of data processing of the methods is not investigated. In fact, the trade-off between the computation time and predictive accuracy must be considered in the design and implementation of the clinical decision support systems.

This research aims to fill this gap and develop a new method for heart disease diagnosis using supervised and unsupervised learning techniques. The proposed method of this research is based on dimensionality reduction, noise removal and prediction learning techniques. In addition, we use Self-Organizing Map (SOM) for clustering the data, Principal Component Analysis (PCA) for noise removal, and Fuzzy Support Vector Machine (Fuzzy SVM) for classification of the heart disease. Furthermore, to improve the computation time of the method and enhance the efficiency of the previous methods on the scalability issue, we rely on PCA and Fuzzy SVM with incremental learning. In each stage, we accordingly evaluate the method using a set of clinical data of the patients. Overall, the contributions of this research are as follows:

(i) This study develops a new method for heart disease classification. The method is developed using Self-Organizing Map [16, 17], Principal Component Analysis [18] for noise removal, and Fuzzy Support Vector Machine [19]. It is believed that the combination of these techniques can improve the efficiency of the previous methods in predicting the disease using real-world clinical patient data.

(ii) This method is developed for accuracy improvement and enhancing the computation time of the previous methods. In relation to the previous methods which solely rely on the accuracy improvement of the clinical decision support systems using learning techniques, we use incremental leaning techniques to improve both accuracy and computation time of the previous methods. Our method is able to learn the prediction models without requiring to recompute all the data from training set which is more efficient in memory saving in relation to the non-incremental methods [20].

(iii) This study uses PCA for dimensionality reduction. As an unsupervised machine leaning technique, PCA tries to simplify the complex structure of a dataset through the variance of the dataset [18, 20]. PCA is able to reduce the data dimensionality and thereby the complexity of data [21]. This is done by linearly combining of the variables of the samples and largest amount of the structured variance [22]. This technique is widely used for dimensionality reduction [23–25]. The proposed method is based on incremental PCA and incremental SVM with fuzzy memberships which able method for online learning. In fact, it is believed that the proposed method can enhance the previous methods in terms of

computation time in noise removal and learning performance. In contract with the previous methods which are based on non-incremental PCA and Fuzzy SVM techniques, the method of this study tries to overcome the online learning by incremental version of these techniques.

(iv) We evaluate the proposed method on two real-world clinical datasets, Cleveland and Statlog. Accordingly, several comparisons with the previous methods are performed to show the effectiveness of the proposed method on these datasets.

The reminder of this paper is organized as follows. We present the related work on diseases diagnosis in Sect. 2. Section 3 presents the mathematical background of the methods. Section 4 provides the proposed method. In Sect. 5, the results and discussions are provided. Finally, we provide the conclusions in Sect. 6.

2 Related Work on Heart Disease Diagnosis

In the study by [26], the authors developed a method using Fuzzy Petri net and fuzzy rule-based reasoning algorithm to predict the heart disease presence. They evaluated their method on Long Beach and Cleveland clinic datasets. [27] developed a fuzzy expert system for the diagnosis of heart disease by a set of medical data from Long Beach and Cleveland clinic datasets. Their method was based on Mamdani inference system. For each variable of disease diagnosis, they considered an appropriate membership function in fuzzy inference system. Accordingly, they could generate a set of valuable decision rules for disease diagnosis. The variable for the heart disease diagnosis were Sex. Age, Pain type, Cholesterol, Blood pressure, Maximum heart rate, Resting blood sugar, Resting electrocardiography, Thallium scan, Exercise, and Old peak. [7] used ensembles of neural network for heart disease diagnosis. The proposed method used several individual classifiers for disease diagnosis. Then, they combined the results of the individual classifiers to obtain the final result. They implemented the method in the SAS software. Their method was evaluated by Cleveland heart disease dataset. They used 70% of the data for training the neural network models and 30% of the data for model validation. They found that 3 neural network classifiers can obtain the best prediction accuracy. In addition, the results of their study showed that accuracy of the ensembles of neural network is much better than single learning techniques such as Naive Bayes and Logistic Regression. Their method accuracy was about 89.01%. The method proposed by [28] was based on association rule mining. For the prediction task, they used Apriori, Predictive Apriori, and Tertius algorithms. Different decision rules were generated by these algorithms. In these algorithms, the generated rules with confirmation levels above 79%, confidence levels above 90%, and accuracy levels above 99% were selected. [29] developed a hybrid method using neural network and fuzzy neural network. They used the backpropagation algorithm to train these techniques. To evaluate the method, they used Cleveland heart disease dataset and trained the model under a 10-cross validation approach. In their experiments, the symptoms of heart disease were Sex, Age, Pain type, Cholesterol, Blood pressure, Maximum heart rate, Resting blood sugar, Resting electrocardiography, Thallium scan, Exercise, and Old peak. They accuracy of their method was 86.8% on Cleveland heart disease dataset. They found that the hybrid of fuzzy neural network and neural network can outperform other proposed methods on heart disease diagnosis such as method developed by *k*-nearest neighbor, C4.5, regression trees, Naive Bayes, and SOM. [30] developed a study for heart disease diagnosis using a set of machine learning techniques. They used Logistic Regression, Rough Set, Multivariate Adaptive Regression Splines, and Artificial Neural Network. Cleveland heart disease dataset was used for their method evaluation. [31] developed a method by machine learning techniques. The aim of their method was to generate the best feature subset and improve the performance of cardiovascular diseases diagnosis systems. Their method was able to find smaller subsets of the feature to accordingly improve the predictive accuracy of the disease diagnosis. They implemented three algorithms for feature selection by a SVM classifier. The three feature selection algorithms were Forward Feature Selection, Forward Feature Inclusion, and Back-elimination Feature Selection. They found that feature selection can significantly improve the accuracy of diseases diagnosis systems. [11] developed a decision support system for heart disease diagnosis using genetic algorithm and SVM. The method was tested on Cleveland Heart Database which included Sex, Age, Resting blood sugar, Pain type, Blood pressure, Maximum heart rate, Cholesterol, Resting electrocardiography, Thallium scan, Exercise, and Old peak as symptoms of heart disease. Their method accuracy was 72.55% on the selected features with RBF kernel of SVM. [8] developed a method for heart disease diagnosis using SVM and neural network techniques. They used 3-layer neural network with backpropagation algorithm to train the neural network. In addition, they used RBF kernel in SVM. The results showed that neural network can provide a better accuracy in relation to the SVM for two-class classification. [32] conducted a study on heart disease diagnosis and developed a method using weighted fuzzy rule-based technique. The method was evaluated on Cleveland Heart Database.

3 Mathematical Background of the Methods

3.1 Self-Organizing Map

As a fascinating neural network method, Self-Organizing Map (SOM) was originally proposed by [16, 17] for supervised and unsupervised learning. SOM learns to cluster the data for discovering the similar patterns form high-dimensional input space [33–35]. The typical structure of a SOM includes two layers: input and Kohonen (output) layer. One neuron in input layer is considered for each input in the vectors space. In output layer, adjustable weights or network parameters are used to connect Kohonen layer neurons to every neuron in the input layer [36, 37]. The weight vectors in the output (Kohonen) layer are description of the distribution of the input samples as vectors of the dataset in an ordered fashion. The SOM algorithm includes three main steps, neuron evaluation, finding the Best Matching Unit (BMU), and updating the neurons. The procedure of SOM for data clustering includes these steps for unsupervised learning [33, 38].

3.2 Support Vector Machine

The Support Vector Machine approach is considered as a supervised machine learning technique [39, 40]. This technique is applied on both classification and regression problems [41]. SVM is widely used in diseases diagnosis decision support systems [42–44].

Considering two-class classification with two classes of data $\mathcal{C}1$ and $\mathcal{C}2$, m samples $i(i = 1, \ldots, m)$ along with n features in a matrix X with $m \times n$ dimensions. For each object i, y_i defines its label, $y_i = 1$ the class belongs to $\mathcal{C}1$ and $y_i = -1$ the class belongs to $\mathcal{C}2$. SVM tries to construct a separating hyperplane $w^{\mathrm{T}}x + b = 0$ that can classify the sample x_i correctly based on its class label, while maximizing the separation margin between the classes. In fact, SVM finds w and b that minimize $\frac{2}{\|w\|}$. To do so, in SVM, the following quadratic programming (Q) [45] problem is solved which is presented in Eq. (1):

$$\begin{aligned} &\min_{w.b.\xi} \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{m} \xi_i \\ &\text{s.t} \quad y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\ &\xi_i \geq 0 \quad i = 1, \ldots, m \end{aligned} \tag{1}$$

where ξ_i represents the slack variable and $C > 0$ is the appropriately selected parameter.

For non-linear SVM, the projection function Ø (.) is used to map the training samples onto a feature space. The kernel-based version for the SVM is introduced In Eq. (2).

The kernel function $\mathrm{K}(x_i, x_s) = \emptyset(x_i).\emptyset\left(x_j\right)$ can be a Gaussian function [46, 47] which is presented in Eq. (3).

$$\begin{aligned} &\max_{a} \sum_{i=1}^{m} a_i - \frac{1}{2}\sum_{i,s=1}^{m} a_i a_s y_i y_s K(x_i, x_s) \\ &\text{s.t.} \quad \sum_{i=1}^{m} a_i y_i = 0,\ 0 \leq a_i \leq C \quad i = 1, \ldots, m \end{aligned} \tag{2}$$

$$K\left(x_{i,} x_s\right) = \exp\left(-\frac{x_i - x_s^2}{2\sigma^2}\right) \tag{3}$$

where a is the Lagrange multiplier and σ is the width parameter. Accordingly, for a new sample x^* with an unknown label, we have

$$y^* = \ \mathrm{sign}\left(\sum i' \in SV a_{i'} a_{i'} K\left(x_{i'}, x^*\right) + b\right) \tag{4}$$

where SV is the set of support vectors.

Fuzzy logic proposed by has been effective in the implementation of decision support systems [48–51]. In this study, we use fuzzy SVM technique [19] for heart disease classification. Considering the first p instances of the dataset (i.e., $t_i = 1, i = 1, 2, \ldots, p$) as positive examples and the remain samples (i.e., $t_i = -1, i = p + 1, p + 2, \ldots, 1$) as negative examples, the fuzzy Support Vector Machine is formulated as follows:

$$\begin{aligned} &\underbrace{\min}_{\omega,\xi}(1/2)\|\omega\|^2 + C^+\sum_{i=1}^{p} s_i^+ \xi_i + C^-\sum_{i=p+1}^{1} s_i^- \xi_i \\ &\text{s.t.} \quad t_i(\omega \cdot \Phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, \quad i = 1, 2, \ldots, l \end{aligned} \tag{5}$$

In the above equation, the importance of x_i in its own class are defined by the fuzzy memberships s_i^+ and s_i^-. The fuzzy membership is given as follows [52]:

$$\begin{aligned} s_i^+ &= 1 - \frac{d_i^{cel+}}{\underbrace{\max}_{j}\left(d_j^{cel+}\right) + \delta}, \quad i = 1, 2, \ldots, p \\ s_i^- &= 1 - \frac{d_i^{cel-}}{\underbrace{\max}_{j}\left(d_j^{cel-}\right) + \delta}, \quad i = p + 1, p + 2, \ldots, l \\ d_i^{cel+} &= \left|\left|x_i - \frac{1}{p}\sum_{j=1}^{p} x_j\right|\right| \\ d_i^{cel-} &= \left|\left|x_i - \frac{1}{1-p}\sum_{j=p+1}^{1} x_j\right|\right| \end{aligned} \tag{6}$$

In the above equation, to fuzzy membership always be higher than zero , δ is used which is a very small positive value. However, as suggested by [19], for irregular distributed samples, we use a K-nearest neighbor strategy for estimating the closeness around training samples. Accordingly, for positive and negative samples x_i, we can

calculate the average distance between x_i and, $N_K^+(x_i)$ and $N_K^-(x_i)$, respectively, through Eqs. (7) and (8).

$$D_i^+ = \frac{1}{K}\sum\nolimits_{x_j \in N_K^+(x_j)} \left\| x_i - x_j \right\| \tag{7}$$

$$D_i^+ = \frac{1}{K}\sum\nolimits_{x_j \in N_K^+(x_j)} \left\| x_i - x_j \right\| \tag{8}$$

According to the above equations, the fuzzy membership is defined as follows:

$$s_i^+ = \left(1 - \alpha * \frac{d_i^{cel+}}{\underbrace{\max}_{j}\left(d_j^{cel+}\right) + \delta} - (1-\alpha) * \frac{D_i^+ - \min D_j^+}{\underbrace{\max}_{j} D_j^+ - \min D_j^+ + \delta} \right)^m, \quad i = 1,2,\ldots,p \tag{9}$$

$$s_i^- = \left(1 - \alpha * \frac{d_i^{cel-}}{\underbrace{\max}_{j}\left(d_j^{cel-}\right) + \delta} - (1-\alpha) * \frac{D_i^- - \min D_j^-}{\underbrace{\max}_{j} D_j^- - \min D_j^- + \delta} \right)^m, \quad i = p+1,p+2,\ldots,1 \tag{10}$$

where $\alpha \in [0,1], m > 0$.

Speeding up the SVM models in the training procedure will result in an important improvement in time efficiency. Incremental SVM classifier is one of the important components of our proposed method. In conventional SVM, batch mode is needed for training model construction. Hence, certain amount of training data must be available before a model can be constructed. In fact, for conventional SVM, a large amount of memory is needed to keep all data for prediction models. In addition, incremental SVM is able to keep the computational process and memory requirements at a minimal level. Accordingly, several efforts have been made for the development of online (incremental) SVM algorithms [53, 54]. [55] proposed an elegant algorithm for online learning by the SVM technique. In this study, we apply their approach with fuzzy memberships for incremental learning in heart disease diagnosis method.

3.3 Imputation Methods

This research applies two imputation techniques for treatment of experimental datasets containing missing values in the features which are hot-deck and *k*-Nearest Neighbor *(k*-NN).

3.3.1 Hot-Deck

For the imputation of an incomplete case, hot-deck approach [56, 57] uses a completely observed donor case. In fact, the imputation is performed through the corresponding values of the donor case. In addition, the most similar case is identified and the missing case is replaced with the identified case as follows:

$$X_i^j = X_k^j, k = \arg \underbrace{\min}_{p} \sqrt{\sum\nolimits_{j \in I(complete)} \mathrm{std}_j \left(X_i^j - X_p^j\right)^2} \tag{11}$$

where std_j indicates the standard deviation of non-missing j th attribute.

3.3.2 k-NN

In this method, among non-missing attributes the search is done by *k*-NN [58]. The *k* most similar instances are selected as follows:

$$X_i^j = \sum\nolimits_{p \in k-NN(X_i)} k\left(X_i^{I(\text{complete})}, X_p^{I(\text{complete})}\right).X_P^j \tag{12}$$

where $k\left(X_i, X_j\right)$ indicates the kernel function for similarity calculation of two vectors X_i and X_j in the dataset, and $k - \mathrm{NN}(X_i)$ indicates the index set in the *k* th nearest neighbors.

4 Research Methodology

Machine learning techniques have been effective in diseases diagnosis [29, 59–64]. This research applies supervised and unsupervised machine learning techniques for heart disease diagnosis. The method is developed by clustering technique, dimensionality reduction, and classification learning techniques. The schematic diagram of the proposed method is shown in Fig. 1. The method performs the classification of heart disease in several stages which are data pre-processing, clustering, dimensionality reduction, and classification. Such stages have been relatively performed in several studies on disease prediction and classification [65–69].

Step 1 (data pre-processing): The data are pre-processed as suggested by the previous studies [70, 71]. In this study, the aim of data pre-processing is to handling the null values of the dataset. Generally, we considered pre-processing stage in our method as this stage is practically performed in the first stage of data analysis. Then, the pre-processed data are used in the next stages of data analysis such as clustering and classification.

The dataset for method evaluation includes several null

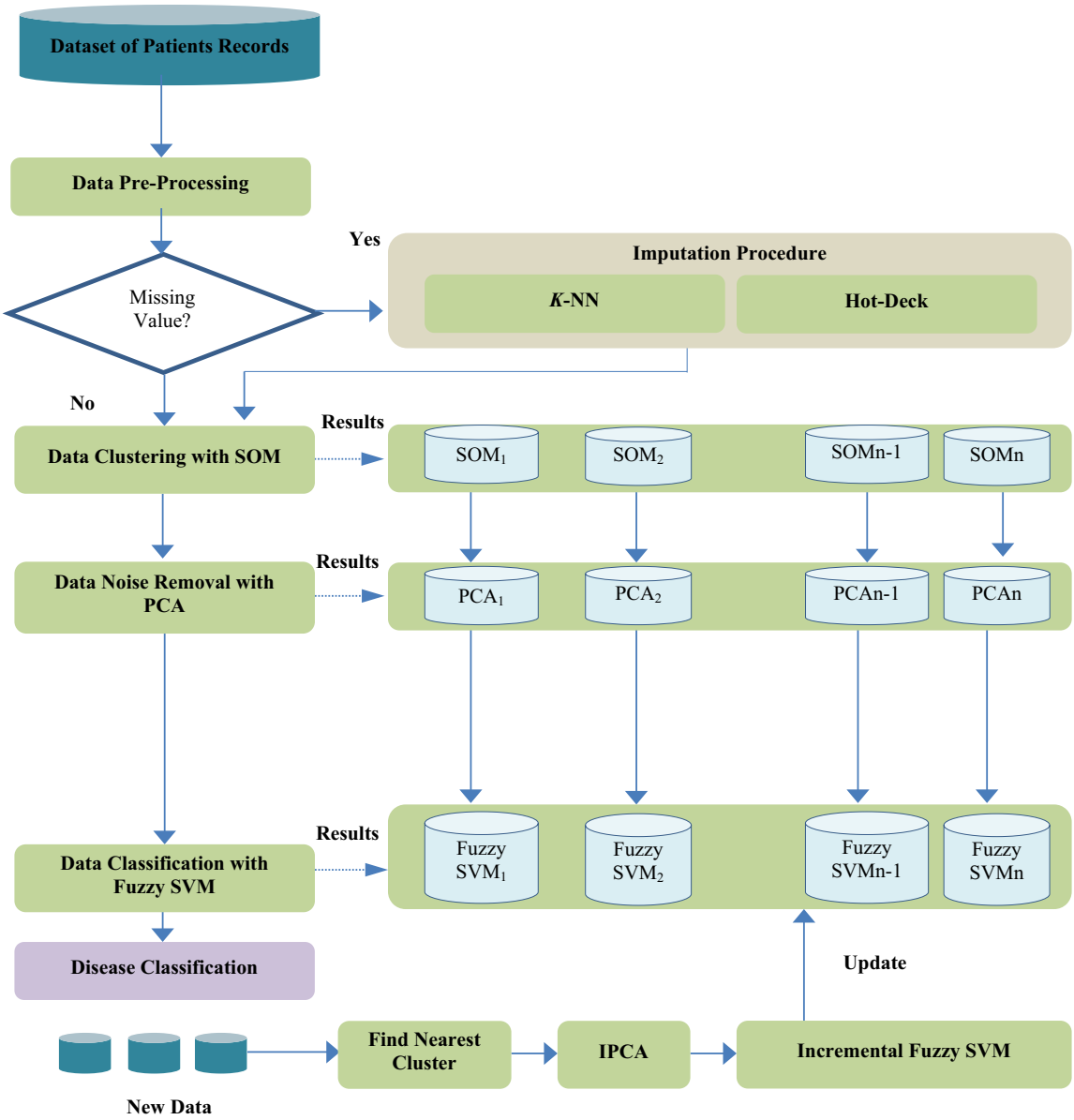


Fig. 1 Schematic diagramm of proposed method

values. These values must be imputed before clustering and classification tasks. This research investigates also the effect of null values on the classification accuracy of the data. Accordingly, we try to solve this issue and apply two algorithms, hot-deck and k-NN, for null value imputation.

Step 2 (clustering): In this stage, we apply an unsupervised machine learning technique to cluster the data. The aim of this stage is to improve the readability of the patients' records by clustering them into several distinct groups. This research applies SOM for this task.

Step 3 (dimensionality reduction): In this stage, a noise removal technique is applied for dimensionality reduction task. Multicollinearity is one of the important issues within the data [72] which significantly impacts on the accuracy of predictors [72–74]. It is found that multicollinearity has negative impact on the accuracy of SVM predictor [74]. Accordingly, we use PCA as the most popular noise removal technique to overcome the multicollinearity issue. We also apply incremental PCA to perform the dimensionality reduction for improving a real-time performance over the traditional PCA.

Step 4 (classification): This stage is applied for classification task. The aim is to classify the heart disease from a set of patient records. In this stage, Fuzzy SVM is applied on the clusters generated by SOM in the second stage. Fuzzy SVM is trained on the training datasets to

construct the classification models. In addition, we also apply incremental Fuzzy SVM for online learning. The method is able to incrementally learn the classification models from new data.

5 Experiments and Results

5.1 Dataset

The main objective of experiments in this research is to evaluate the effectiveness of the proposed method for the heart disease classification using real-world data. To do so, two heart disease datasets are selected which are freely available in Data Mining Repository of the University of California, Irvine (UCI) (https://archive.ics.uci.edu/ml/datasets). The datasets are Cleveland and Statlog. The first dataset for our experiment is Cleveland in which the number of observations in this dataset is 303 in two classes (sick with 164 samples and normal with 139 samples) with 13 sampled features. This dataset includes missing values. The second dataset is Statlog which is in a slightly different format from the above dataset. In this dataset, there are 270 observations in two classes (150 samples with class 1 and 120 samples with class 2) with 13 sampled features and no missing values. Heart dataset component introduced above hold several features for the symptoms of a heart attack that are used to identify healthy persons and patients. In addition, these datasets have one column that indicates the class labels. The class labels are used for prediction using supervised machine learning methods such as Fuzzy SVM that we incorporated to the proposed model.

5.2 Experimental Setup

The experimental setup of our experiment on the dataset with missing values is as follows. The dataset is divided into two subsets, 70% of the dataset as training subset and 30% of dataset as test subset. Then we applied the imputation procedure through hot-deck and k-NN on the training set for missing value imputation. The assumption behind the use of k-NN for missing values is that a null value in the dataset can be approximated by the values of the points that are closest to it, the variable k is set to the most similar cases from non-missing data. The mean rule was used to predict missing numerical features. In hot-deck, the most similar records were imputed to missing values. The resulting dataset was used in SOM for clustering task. In this stage, different map sizes of SOM were considered. We then applied PCA on each cluster which does not include the missing values. In the final stage, the classification models were constructed by Fuzzy SVM. To obtain the classification accuracy, the Fuzzy SVM classification model was evaluated on the test set. We then compared the results of our experiments with the dataset which is clustered and the imputation was not performed, and the dataset without clustering and the imputation was not performed. We also compared our method with the proposed methods on the heart disease diagnosis in the literature. The comparisons have been made by three evaluation measures which are accuracy, sensitivity, and specificity through the confusion matrix as shown in Table 1 [75]. The sensitivity indicates the percentage of subjects (patients) which are diagnosed correctly as patients. The specificity indicates the percentage of subjects (healthy people) which are correctly diagnosed as healthy. Accuracy indicates the percentage of subject (healthy individuals and patients) who are diagnosed correctly [75]. The formulas for these measures are provided as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{13}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{15}$$

where TN indicates true negatives, TP indicates true positives, FN indicates false negatives, and FP indicates false positives.

Table 1 Confusion matrix

Actual class	Predicted class	
Positive	TP	FN
Negative	FP	TN

5.3 Experimental Results

The data were clustered after dataset imputation. We clustered the data using SOM with different SOM sizes. We found that the best cluster size obtained by $SOM2 \times 2$ and $SOM1 \times 3$, indicating 4 and 3 clusters, respectively, for Cleveland and Statlog. These clusters include the attributes and the corresponding class labels. All clusters have been generated with learning rate $\alpha = 0.05$. The map qualities for Cleveland and Statlog were, respectively, measured $\text{MapQ} = 0.8574$ and $\text{MapQ} = 0.824$. As we selected SVM classification method, linearly dependent between the variables can usually cause accuracy issue in the classifier. To overcome this issue, we need to reduce the multicollinearity among the features in the dataset. Accordingly, to remove correlated features or reduce the correlations between the variables we applied PCA on the clusters. We examined the datasets for correlation

coefficients between the input variables as shown in Fig. 2. Overall, from these tables it can be found that, a strong dependency can be found between some features of both datasets. It is believed that the reducing the multicollinearity among the features will enhance the classification accuracy.

The results for PCA are presented in Fig. 5a, b for Statlog and Cleveland. These figures further reveal the number of PCs generated by PCA technique. As we are looking for trade-off between the accuracy and computation time, accordingly the best PCs are selected in PCA for each cluster. This is done by the method proposed by [76]. As suggested by [76], the PCs are selected based on their percentage of information they have provided. To do so, we provide the Scree plot for each cluster for PCs selection. In Fig. 3, we provide the results for first cluster of the datasets. In these figures, it is clear that PC1-PC8 have provided more than 80% of the information of the original datasets. In the final stage of our experiments, we applied SVM and incremental Fuzzy SVM for constructing the batch and incremental training models. For both approaches, we use the Radial Basis Function (RBF) kernel with a fixed parameter $C = 0$ and $\gamma = 2^3$. The Receiver Operating Characteristic (ROC) curve is provided for the first cluster of Cleveland and Statlog in Fig. 4, with the Area Under the Curve (AUC) of 0.914 and 0.921, respectively.

The average results for classification of heart disease by the methods are presented in Tables 2 and 3. In these tables, the accuracy, sensitivity, and specificity values of the classification for Cleveland and Statlog are provided. The results show that the best classification accuracy, sensitivity, and specificity are obtained for *k-NN+SOM+PCA+Fuzzy SVM* method for both datasets. In addition, the results reveal that the accuracy of the classification with the use of SOM and PCA is higher than other methods. This shows that overcoming the multicollinearity issue in medical data can significantly improve the predictive accuracy of the classification methods. Our results further reveal that imputing the missing values with a robust method is also important before classification task. Our experiment on Cleveland dataset which included missing values approved the usefulness of *k*-NN and hot-deck as two imputation techniques in predicting missing values in medical datasets and accordingly improving the efficiency of disease classification.

We also evaluate the method for online learning. Accordingly, we provide an experiment for computational efficiency on both datasets when new data are incrementally added to the datasets. For computation time, the increment size (incsize = 5) for training the incremental Fuzzy SVM is kept constant. For each incremental adding the sample, the computation time of the training is calculated for both incremental Fuzzy SVM and non-incremental SVM. The results for training the models for first clusters of Cleveland and Statlog are presented in Fig. 5a, b, respectively. It can be observed that the computation time for the incremental Fuzzy SVM is much lower than the method which does not support online learning. In addition, when new training samples are added to the dataset, the incremental Fuzzy SVM does not need to training whole data in constructing the classification models. For large training sets, it was found that the difference in processing time is considerable, the incrementally Fuzzy SVM showed on average 4 times faster than non-incremental SVM in constructing the classification models.

6 Summary and Conclusions

In this paper, a new method was proposed for heart disease diagnosis through supervised and unsupervised learning techniques. The method used two imputation approaches, Hot-Deck and *k*-NN, to impute missing values in the experimental datasets. We then performed clustering through SOM technique. In our experiments, we found that there was a correlation between the features of the dataset. Accordingly, to reduce the dimensionality of the data and

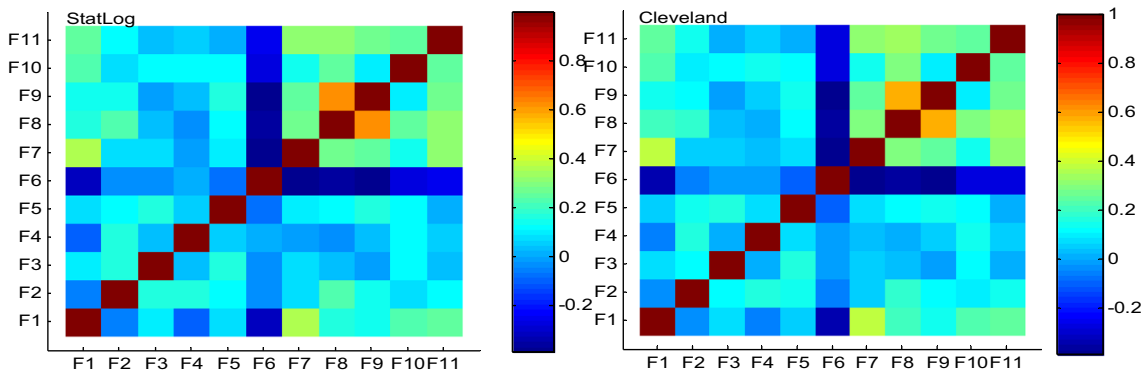


Fig. 2 Correlation coefficients for Statlog and Cleveland

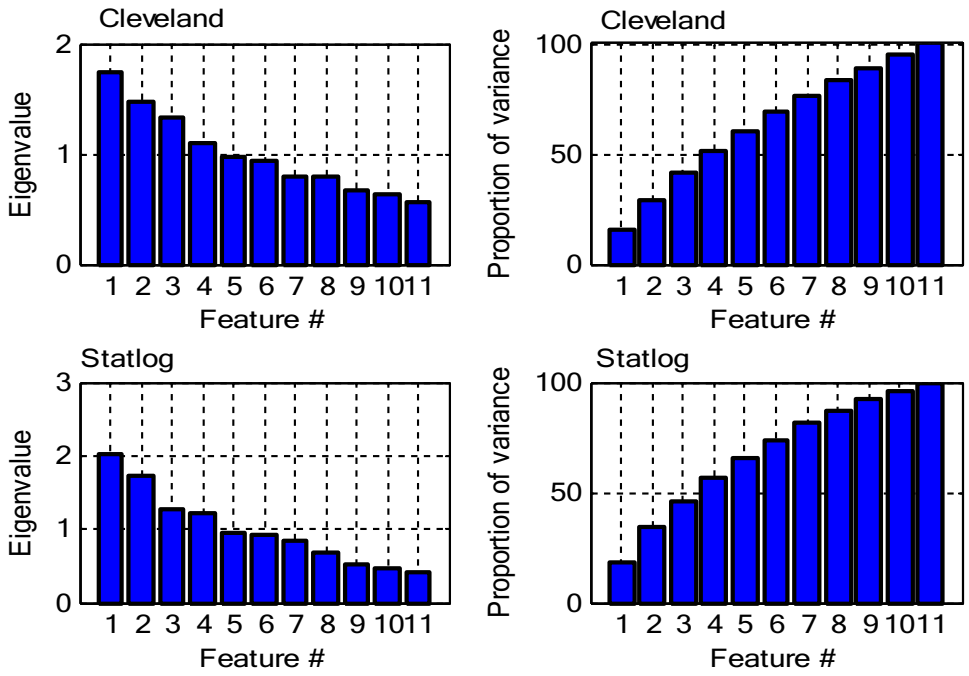


Fig. 3 PCs selection for first cluster of Cleveland and Statlog

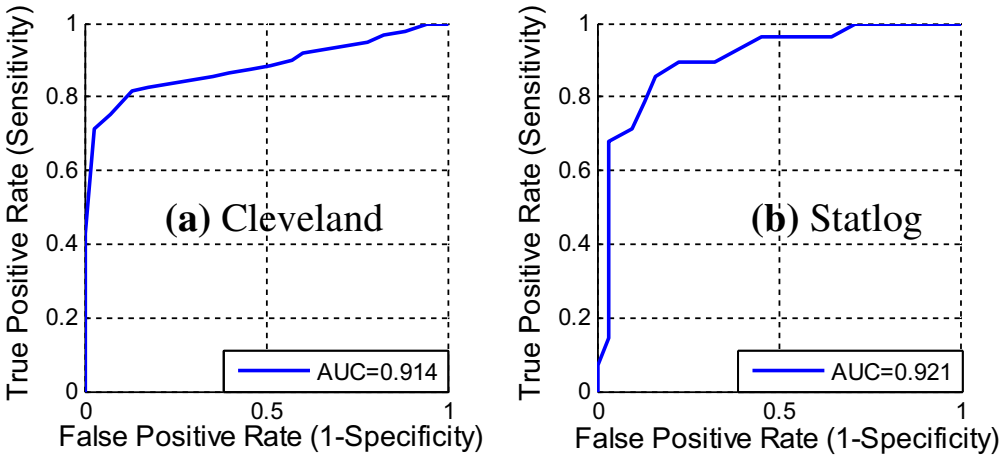


Fig. 4 Receiver Operating Characteristic (ROC) curve of first cluster of **a** Cleveland and **b** Statlog

overcome the multicollinearity issue within the data, PCA was applied on each cluster generated by SOM. Then, we used Fuzzy SVM to construct the classification models for heart disease diagnosis. Finally, we evaluated our method through two real-world datasets, Statlog and Cleveland. We provided accuracy, sensitivity, and specificity values for each method for comparisons. The results revealed that the dataset imputation has a positive relationship with the accuracy of the Fuzzy SVM classifier. In addition, we found that the methods which rely on PCA provide better accuracy in relation to the other methods. In fact, it was found that, in the medical dataset the multicollinearity can significantly affect the predictive accuracy of the classifiers. Our experimental findings on two datasets also showed that the use of the methods with incremental techniques can have advantages on enhancing the computation time of disease prediction.

This study has some limitations. This study has used only two medical datasets for method evaluation. The datasets used in this study are not large enough to better reflect its usefulness in processing big datasets. In addition, large clinical data will help better reveal the shortcomings and advantages of the proposed method in terms of computation time and prediction accuracy. Hence, future study may implement the proposed method for large datasets with many features. In addition, our method has applied

Table 2 Results of method for Cleveland dataset

Method	Specificity	Sensitivity	Accuracy
KNN+SOM+PCA+Fuzzy SVM	0.9435	0.9606	0.9686
Hot-Deck+SOM+PCA+Fuzzy SVM	0.9334	0.9540	0.9568
SOM+PCA+SVM	0.9143	0.9412	0.9449
PCA+SVM	0.8571	0.9020	0.8976
SOM+SVM	0.8131	0.8693	0.8661
SVM	0.7642	0.8377	0.8268
NB	0.6909	0.7778	0.7677
DT	0.6727	0.7616	0.7441
NN	0.6330	0.7351	0.7087

Table 3 Results of method for Statlog dataset

Method	Specificity	Sensitivity	Accuracy
KNN+SOM+PCA+Fuzzy SVM	0.9697	0.9697	0.9787
Hot-Deck+SOM+PCA+ Fuzzy SVM	0.9697	0.9355	0.9583
SOM+PCA+SVM	0.9697	0.9032	0.9388
PCA+SVM	0.8529	0.8529	0.8958
SOM+SVM	0.8235	0.8235	0.8723
SVM	0.7941	0.7941	0.8511
NB	0.7429	0.7429	0.8085
DT	0.7297	0.7297	0.7727
NN	0.6842	0.6842	0.7273

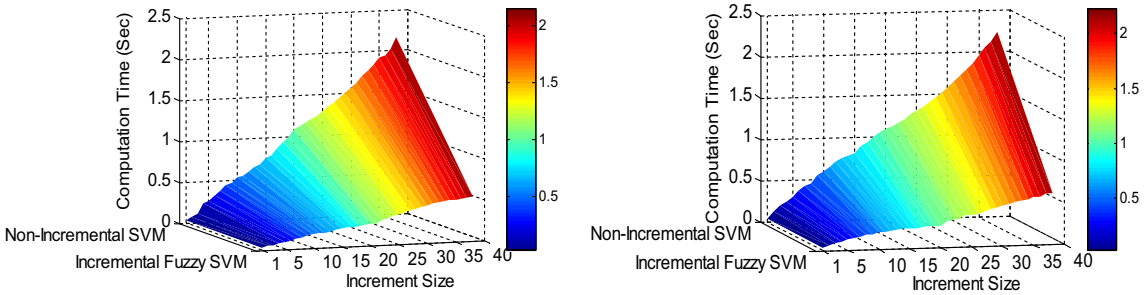


Fig. 5 Computation time of non-incremental SVM and incremental Fuzzy SVM

PCA and SOM for noise removal and clustering task. Other dimensionality and clustering techniques are also suggested for further investigations. Furthermore, our method used crisp type of *k*-NN and Hot-Deck as two imputation techniques for missing value prediction. Fuzzy version of these techniques may provide better results for imputing the missing values. Finally, our method by SVM was developed for classification of heart disease. Hence, it is suggested to implement the SVM for its regression application with the aid of clustering and noise removal techniques for predicting other types of diseases.

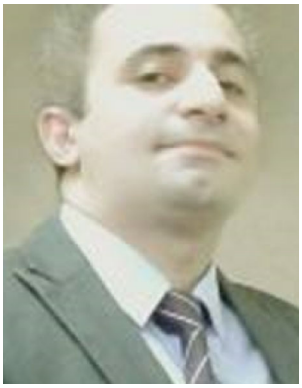
Acknowledgements This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Fast-track Research Funding Program.

References

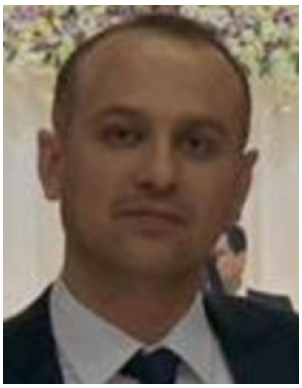
1. Mendis, S., et al.: Global atlas on cardiovascular disease prevention and control. World Health Organization, Geneva (2011)
2. Paul, A.K., et al.: Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease. Appl. Intell. **48**(7), 1739–1756 (2018)
3. McAloon, C.J., et al.: The changing face of cardiovascular disease 2000–2012: an analysis of the world health organisation global health estimates data. Int. J. Cardiol. **224**, 256–264 (2016)
4. Luengo-Fernandez, R., Leal, J., Gray, A.: UK research expenditure on dementia, heart disease, stroke and cancer: are levels of spending related to disease burden? Eur. J. Neurol. **19**(1), 149–154 (2012)
5. El-Bialy, R., et al.: Feature analysis of coronary artery heart disease data sets. Procedia Comput. Sci. **65**, 459–468 (2015)

6. Metwally, A.H., Elgamal, M.-A.F.: The relation between hepatitis C virus and coronary heart disease. Med. Hypotheses **82**(4), 505 (2014)
7. Das, R., Turkoglu, I., Sengur, A.: Effective diagnosis of heart disease through neural networks ensembles. Expert Syst. Appl. **36**(4), 7675–7680 (2009)
8. Gudadhe, M., Wankhade, K., Dongre, S.: Decision support system for heart disease based on support vector machine and artificial neural network. In: 2010 International Conference on Computer and Communication Technology (ICCCT), IEEE (2010)
9. Berner, E.S.: Clinical decision support systems, vol. 233. Springer, Berlin (2007)
10. Johnston, M.E., et al.: Effects of computer-based clinical decision support systems on clinician performance and patient outcome: a critical appraisal of research. Ann. Intern. Med. **120**(2), 135–142 (1994)
11. Bhatia, S., Prakash, P., Pillai, G.: SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features. In: Proceedings of the world congress on engineering and computer science (2008)
12. Mantel, N.: The detection of disease clustering and a generalized regression approach. Cancer Res. **27**(2), 209–220 (1967)
13. Nahato, K.B., Nehemiah, K.H., Kannan, A.: Hybrid approach using fuzzy sets and extreme learning machine for classifying clinical datasets. Inform. Med. Unlocked **2**, 1–11 (2016)
14. Nilashi, M., et al.: A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics Inform. **34**(4), 133–144 (2017)
15. Fida, B., et al.: Heart disease classification ensemble optimization using genetic algorithm. In: 2011 IEEE 14th International Multitopic Conference, IEEE (2011)
16. Kohonen, T.: Analysis of a simple self-organizing process. Biol. Cybern. **44**(2), 135–140 (1982)
17. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biol. Cybern. **43**(1), 59–69 (1982)
18. Maćkiewicz, A., Ratajczak, W.: Principal components analysis (PCA). Comput. Geosci. **19**(3), 303–342 (1993)
19. Ju, Z., Cao, J.-Z., Gu, H.: Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC. J. Theor. Biol. **397**, 145–150 (2016)
20. Nilashi, M., et al.: A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques. Biocybern. Biomed. Eng. **38**(1), 1–15 (2018)
21. Abdi, H., Williams, L.J.: Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. **2**(4), 433–459 (2010)
22. Ringnér, M.: What is principal component analysis? Nat. Biotechnol. **26**(3), 303 (2008)
23. Ding, C., He X.: K-means clustering via principal component analysis. In: Proceedings of the twenty-first international conference on Machine learning, ACM (2004)
24. Wall, M.E., Rechtsteiner, A., Rocha, L.M.: Singular value decomposition and principal component analysis. In: A practical approach to microarray data analysis. 2003, Springer. p. 91–109
25. Wright, J., et al.: Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization. In: Advances in neural information processing systems (2009)
26. Übeyli, E.D., Doğdu, E.: Automatic detection of erythemato-squamous diseases using k-means clustering. J. Med. Syst. **34**(2), 179–184 (2010)
27. Adeli, A., Neshat, M.: A fuzzy expert system for heart disease diagnosis. In: Proceedings of International Multi Conference of Engineers and Computer Scientists, Hong Kong (2010)
28. Nahar, J., et al.: Association rule mining to detect factors which contribute to heart disease in males and females. Expert Syst. Appl. **40**(4), 1086–1093 (2013)
29. Kahramanli, H., Allahverdi, N.: Design of a hybrid system for the diabetes and heart diseases. Expert Syst. Appl. **35**(1–2), 82–89 (2008)
30. Shao, Y.E., Hou, C.-D., Chiu, C.-C.: Hybrid intelligent modeling schemes for heart disease classification. Appl. Soft Comput. **14**, 47–52 (2014)
31. Shilaskar, S., Ghatol, A.: Feature selection for medical diagnosis: evaluation for cardiovascular diseases. Expert Syst. Appl. **40**(10), 4146–4153 (2013)
32. Anooj, P.: Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules. J. King Saud Univ. Comput. Inform. Sci. **24**(1), 27–40 (2012)
33. Mangiameli, P., Chen, S.K., West, D.: A comparison of SOM neural network and hierarchical clustering methods. Eur. J. Oper. Res. **93**(2), 402–417 (1996)
34. Nilashi, M., et al.: An analytical method for measuring the Parkinson's disease progression: a case on a Parkinson's telemonitoring dataset. Measurement **136**, 545–557 (2019)
35. Ahani, A., et al.: Revealing customers' satisfaction and preferences through online review analysis: the case of Canary Islands hotels. J. Retail. Consum. Serv. **51**, 331–343 (2019)
36. Chen, D.-R., Chang, R.-F., Huang, Y.-L.: Breast cancer diagnosis using self-organizing map for sonography. Ultrasound Med. Biol. **26**(3), 405–411 (2000)
37. Kiviluoto, K.: Predicting bankruptcies with the self-organizing map. Neurocomputing **21**(1–3), 191–201 (1998)
38. Vesanto, J., Alhoniemi, E.: Clustering of the self-organizing map. IEEE Trans. Neural Networks **11**(3), 586–600 (2000)
39. Brown, M.P., et al.: Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. **97**(1), 262–267 (2000)
40. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: European conference on machine learning, Springer (1998)
41. Joachims, T.: Learning to classify text using support vector machines, vol. 668. Springer, Berlin (2002)
42. Guyon, I., et al.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**(1–3), 389–422 (2002)
43. Orru, G., et al.: Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci. Biobehav. Rev. **36**(4), 1140–1152 (2012)
44. Samanta, B., Al-Balushi, K., Al-Araimi, S.: Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection. Eng. Appl. Artif. Intell. **16**(7–8), 657–665 (2003)
45. Gunn, S.R.: Support vector machines for classification and regression. ISIS Tech. Rep. **14**(1), 5–16 (1998)
46. Keerthi, S.S., Lin, C.-J.: Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput. **15**(7), 1667–1689 (2003)
47. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, IEEE (2004)
48. Mardani, A., et al.: Application of decision making and fuzzy sets theory to evaluate the healthcare and medical problems: a review of three decades of research with recent developments. Expert Syst. Appl. **137**, 202–231 (2019)
49. Nilashi, M., et al.: Factors influencing medical tourism adoption in Malaysia: a DEMATEL-Fuzzy TOPSIS approach. Comput. Ind. Eng. **137**, 106005 (2019)
50. Nilashi, M., et al.: Measuring sustainability through ecological sustainability and human sustainability: a machine learning approach. J. Clean. Prod. **240**, 118162 (2019)

51. Yadegaridehkordi, E., et al.: The impact of big data on firm performance in hotel industry. Electron. Commer. Res. Appl. **40**, 100921 (2019)
52. Lin, C.-F., Wang, S.-D.: Fuzzy support vector machines. IEEE Trans. Neural Netw. **13**(2), 464–471 (2002)
53. Wang, W., Men, C., Lu, W.: Online prediction model based on support vector machine. Neurocomputing **71**(4–6), 550–558 (2008)
54. Zhang, Z., Shen, H.: Application of online-training SVMs for real-time intrusion detection with different considerations. Comput. Commun. **28**(12), 1428–1442 (2005)
55. Cauwenberghs, G., Poggio, T.: Incremental and decremental support vector machine learning. In: Advances in neural information processing systems (2001)
56. Andridge, R.R., Little, R.J.: A review of hot deck imputation for survey non-response. Int. Stat. Rev. **78**(1), 40–64 (2010)
57. Myers, T.A.: Goodbye, listwise deletion: presenting hot deck imputation as an easy and effective tool for handling missing data. Commun. Methods Meas. **5**(4), 297–310 (2011)
58. Sim, J., Lee, J.S., Kwon, O.: Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications. Math. Probl. Eng. (2015). https://doi.org/10.1155/2015/538613
59. Nahato, K.B., Harichandran, K.N., Arputharaj, K.: Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Comput. Math. Methods Med. (2015). https://doi.org/10.1155/2015/460189
60. Nilashi, M., Ibrahim, O., Ahani, A.: Accuracy improvement for predicting Parkinson's disease progression. Sci. Rep. **6**, 34181 (2016)
61. Nilashi, M., et al.: A soft computing method for mesothelioma disease classification. J. Soft Comput. Decis. Support Syst. **4**(1), 16–18 (2017)
62. Nilashi, M., et al.: An analytical method for diseases prediction using machine learning techniques. Comput. Chem. Eng. **106**, 212–223 (2017)
63. Nilashi, M., et al.: Accuracy improvement for diabetes disease classification: a case on a public medical dataset. Fuzzy Inform. Eng. **9**(3), 345–357 (2017)
64. Ahmadi, N., et al.: An intelligent method for iris recognition using supervised machine learning techniques. Opt. Laser Technol. **120**, 105701 (2019)
65. Santhanam, T., Padmavathi, M.: Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Comput. Sci. **47**, 76–83 (2015)
66. Zheng, B., Yoon, S.W., Lam, S.S.: Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst. Appl. **41**(4), 1476–1482 (2014)
67. Hariharan, M., Polat, K., Sindhu, R.: A new hybrid intelligent system for accurate detection of Parkinson's disease. Comput. Methods Programs Biomed. **113**(3), 904–913 (2014)
68. Ortiz, A., et al.: LVQ-SVM based CAD tool applied to structural MRI for the diagnosis of the Alzheimer's disease. Pattern Recogn. Lett. **34**(14), 1725–1733 (2013)
69. Toro, C., et al.: Supervoxels-based histon as a new alzheimer's disease imaging biomarker. Sensors **18**(6), 1752 (2018)
70. Ahmed, H., et al.: Heart disease identification from patients' social posts, machine learning solution on Spark. Future Gener. Comput. Syst. (2019). https://doi.org/10.1016/j.future.2019.09.056
71. Pumo, D., et al.: Sensitivity of extreme rainfall to temperature in semi-arid Mediterranean regions. Atmos. Res. **225**, 30–44 (2019)
72. Lee, J.-H., et al.: Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods. Geomorphology **303**, 284–298 (2018)
73. Aguilera, A.M., Escabias, M., Valderrama, M.J.: Using principal components for estimating logistic regression with high-dimensional multicollinear data. Comput. Stat. Data Anal. **50**(8), 1905–1924 (2006)
74. Kim, M.-J., Kim, H.-B., Kang, D.-K.: Optimizing SVM ensembles using genetic algorithms in bankruptcy prediction. J. Inform. Commun. Converg. Eng. **8**(4), 370–376 (2010)
75. Ramezani, M., Karimian, A., Moallem, P.: Automatic detection of malignant melanoma using macroscopic images. J. Med. Signals Sensors **4**(4), 281 (2014)
76. Cattell, R.B.: The scree test for the number of factors. Multivar. Behav. Res. **1**(2), 245–276 (1966)



Mehrbakhsh Nilashi received his Ph.D. degree in Computer Science in the faculty of Computing, Universiti Teknologi Malaysia in 2014. His researches are mainly in the fields of Soft Computing, Machine Learning, Multi-criteria Decision Making, Information Retrieval, Recommender Systems with a special focus on Multi-Criteria Recommender Systems, Health Informatics, tourism management, and Decision Support Systems. His contributions have been published in prestigious peer-reviewed journals and international conferences.



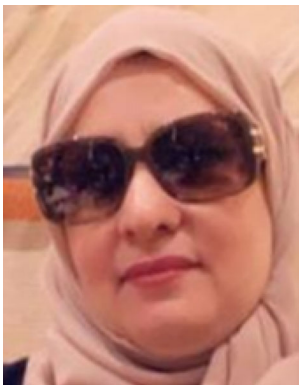
Hossein Ahmadi received his Ph.D. degree in Information Systems in the faculty of Computing, Universiti Teknologi Malaysia in 2016. His research interests are mainly in the field of information and communication technologies, innovation adoption and implementation, artificial intelligence, marketing and strategy, big data analysis, disease diagnosis, internet of things, fuzzy logic, sustainability development, tourism development, e-Health in public sector, and multi-criteria decision making. His contributions have been published in prestigious peer-reviewed journals.

Azizah Abdul Manaf received her Ph.D. in Computer Science from UTM, Malaysia in 1995. Her research interests are mainly in the field of Image Processing, Multimedia Security, Computer Forensics, and Software Development. Her contributions have been published in prestigious peer-reviewed journals.

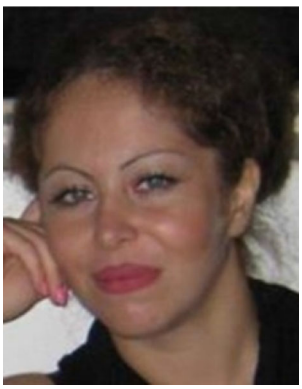
Tarik A. Rashid received his Ph.D. in Computer Science and Informatics degree from College of Engineering, Mathematical and Physical Sciences, University College Dublin (UCD) in 2001–2006. He pursued his Post-Doctoral Fellow at the Computer Science and Informatics School, College of Engineering, Mathematical and Physical Sciences, University College Dublin (UCD) from 2006 to 2007. His research interests are mainly in the field of artificial intelligence, marketing and strategy, big data analysis, and tourism development. His contributions have been published in prestigious peer-reviewed journals.

Sarminah Samad is an Associate Professor in CBA Research Centre, Department of Business Administration, Collage of Business and Administration, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia. Her research expertise is focused on Technology Adoption, Social Informatics, and Business Research. Her contributions have been published in prestigious peer-reviewed journals and international conferences.

Leila Shahmoradi has Ph.D. in Health Information Management. She studied and published on medical informatics area. Her research interests are mainly in the field of Knowledge Management, Machine learning, and Decision Support Systems.



Nahla Aljojo obtained her PhD in Computing at Portsmouth University. She is working currently as Associate Professor at College of Computer Science and Engineering, Information system and information Technology Department, University of Jeddah, Jeddah, Saudi Arabia. Her research interests include adaptivity in web-based educational systems, E-business, leadership's studies, information security and data integrity, E-Learning, education, machine learning, health informatics, environment and ecology, and logistics and supply chain management. Her contributions have been published in prestigious peer-reviewed journals.



Elnaz Akbari is a post-doctoral researcher in the field of electrical engineering at the University of California, Merced. Her major fields of research activity include decision making methods, machine learning, and artificial intelligence. Her contributions have been published in prestigious peer-reviewed journals and international conferences.