*Computers, Materials & Continua*
DOI:10.32604/cmc.2022.023848
***Article***

Tech Science Press

# A Hybrid Meta-Classifier of Fuzzy Clustering and Logistic Regression for Diabetes Prediction

**Altyeb Altaher Taha*** **and Sharaf Jameel Malebary**

Department of Information Technology, Faculty of Computing and Information Technology in Rabigh,
King Abdulaziz University, Jeddah, Saudi Arabia
*Corresponding Author: Altyeb Altaher Taha. Email: aaataha@kau.edu.sa

Received: 23 September 2021; Accepted: 02 November 2021

**Abstract:** Diabetes is a chronic health condition that impairs the body's ability to convert food to energy, recognized by persistently high levels of blood glucose. Undiagnosed diabetes can cause many complications, including retinopathy, nephropathy, neuropathy, and other vascular disorders. Machine learning methods can be very useful for disease identification, prediction, and treatment. This paper proposes a new ensemble learning approach for type 2 diabetes prediction based on a hybrid meta-classifier of fuzzy clustering and logistic regression. The proposed approach consists of two levels. First, a base-learner comprising six machine learning algorithms is utilized for predicting diabetes. Second, a hybrid meta-learner that combines fuzzy clustering and logistic regression is employed to appropriately integrate predictions from the base-learners and provide an accurate prediction of diabetes. The hybrid meta-learner employs the Fuzzy C-means Clustering (FCM) algorithm to generate highly significant clusters of predictions from base-learners. The predictions of base-learners and their fuzzy clusters are then employed as inputs to the Logistic Regression (LR) algorithm, which generates the final diabetes prediction result. Experiments were conducted using two publicly available datasets, the Pima Indians Diabetes Database (PIDD) and the Schorling Diabetes Dataset (SDD) to demonstrate the efficacy of the proposed method for predicting diabetes. When compared with other models, the proposed approach outperformed them and obtained the highest prediction accuracies of 99.00% and 95.20% using the PIDD and SDD datasets, respectively.

**Keywords:** Ensemble learning; fuzzy clustering; diabetes prediction; machine learning

## 1 Introduction

Diabetes Mellitus (DM) is a common chronic disease that affects approximately 425 million people worldwide, and this figure is expected to rise to 629 million by 2045 [1]. It is a metabolic condition characterized by elevated blood glucose levels. Diabetes is classified into three types: type 1 diabetes (T1D), (ii) type 2 diabetes (T2D), and (iii) gestational diabetes (GD). T1D is most commonly found in young adults under the age of 30. Polyuria, thirst, constant hunger, weight loss, vision changes, and fatigue are all symptoms of T1D. T2D affects adults over the age of 45 and is frequently associated with obesity, hypertension, dyslipidemia, arteriosclerosis, and other diseases. Gestational diabetes is the third type of diabetes. Actually, gestational diabetes affects pregnant women and its diagnostic criteria vary by country, illustrating the lack of consensus around the most effective way to screen in routine clinical care [2,3]. This research focus on the prediction of Type 2 diabetes, which is distinguished by relative insulin deficiency due to pancreatic-cell dysfunction and insulin resistance in target organs. The consequences include a variety of health-related issues, such as early death, blindness, cardiac disease, and kidney problems. The seriousness of this disorder can be compounded by the fact that it can be ppdiagnosis of diabetes will help decrease the occurrence of the disease and related complications.

Over the last few years, there has been increasing interest in the use of machine learning techniques for the early diagnosis of diabetes to improve classification accuracy. Many classification methods based on machine learning, including Naive Bayes (NB), Neural Networks (NN), Support Vector Machines (SVM), Fuzzy Decision Trees (FDT), K-Nearest Neighbor (K-NN) and Decision Trees (DT), have been used for the diagnosis of diabetes [4,5]. However, several researchers in this field have concluded that there is no single classifier that can significantly boost the prediction accuracy. Because of this challenge, there has been much focus on ensemble-based learners to improve prediction efficiency. This type of learner assumes that various base classifiers create different classification mistakes and that classification accuracy can be greatly enhanced by the integration of these particular classifiers. It is also widely acknowledged that an ensemble learner outperforms an individual learner in terms of robustness and precision, as shown by the results in a variety of research domains [6]. Ensembles of different individual classifiers benefit from the various biases and variances of each classification algorithm. Stacking is an example of an ensemble learning strategy that uses a two-stage classification system: the base-learner stage and the meta-learner stage. Identifying the appropriate configuration of the base-learner and meta-learner models is a challenging but critical process because the efficiency of the classification scheme is heavily dependent on the accurate determination of both models [7,8]. Ensemble strategies train multiple learners and then combine them to achieve a better predictive performance. In this research, we compared the proposed approach with the soft voting and hard voting methods. Each individual classifier in soft voting offers a probability value that a certain data point belongs to a specified target class. The predictions are weighted by the significance of the classifier and summed. The target label with the highest sum of weighted probability then wins the vote. In a hard voting ensemble, the votes for crisp class labels from different models are added together and the class with the most votes is predicted.

Although machine learning approaches for diabetes prediction have been developed recently, existing models rely heavily on a single classifier trained on a single dataset, which is incapable of accurately predicting diabetes. Therefore, this research proposes an ensemble approach based on the principles of the stack method, which combines multiple base models for sample prediction and one meta-model for achieving the ultimate predictions by combining the base-level classifiers. The main goal of this study is to explore whether integrating fuzzy clustering and logistic regression at the meta-learner level improves ensemble stacking efficiency. The proposed ensemble approach utilizes a combination of fuzzy clustering and logistic regression to create a highly accurate diabetes model. The hybrid meta-learner employs the Fuzzy C-means clustering (FCM) algorithm to generate highly significant clusters of predictions from base-learners. The predictions of the base-learners and their fuzzy clusters are then employed as inputs to a logistic regression (LR) algorithm, which generates the final diabetes prediction result. The base-learner model includes six different machine learning algorithms: KNN, radial SVM, NB, linear SVM, NN, and light gradient boosting machine

(LightGBM). The proposed approach can be leveraged on health care systems to enhance the accuracy of disease prediction.

Diabetes prediction experiments were used to assess the proposed ensemble method based on five performance measures: accuracy, precision, recall, f1-score, and area under the receiver operating characteristic curve (AUC). The main contributions of this study and its significance for improving type 2 diabetes prediction can be summarized as follows:

- Based on the excellent performance of ensemble learning, a hybrid ensemble approach for diabetes prediction is suggested, which can integrate the base learners to create a more efficient learner.
- The proposed approach employs a hybrid meta-learner that combines fuzzy clustering and logistic regression to integrate base-learner results appropriately and provide an accurate diabetes prediction.
- In order to improve the accuracy of diabetes prediction, the suggested approach uses the fuzzy clustering method to explore the hidden information in base-learner predictions.
- The proposed approach outperformed the individual classifiers, soft and hard voting ensemble approaches, and achieved the highest diabetes predictive accuracy of 99% and 95% using two publicly available datasets.
- When compared to experimental results from other researchers utilizing the same datasets, the proposed approach achieved superior results.

The remainder of this paper is organized as follows. Section 2 introduces studies related to ensemble learning approaches for diabetes prediction. Section 3 describes the proposed ensemble learning approach for diabetes prediction, which is based on the hybrid meta-classifier of fuzzy clustering and logistic regression. Experimental results and discussion are provided in Section 4. Finally, conclusions and future work are presented in Section 5.

## 2 Related Works

Several ensemble learning techniques for predicting diabetes have recently been formulated. These techniques have received significant attention because they are more efficient than individual learners in achieving high classification accuracy and generalization capacity. Ensemble learning constructs a model from multiple classifiers, which are combined to create a stronger model with better prediction efficiency [9].

Zolfaghari [10] proposed a stack of an SVM and artificial neural network for diabetes diagnosis based on the Pima Indians Diabetes Database (PIDD) dataset; the proposed approach achieved an accuracy of 88.04% and outperformed the single classifiers. Saleh et al. [11] introduced an ensemble classifier for diabetic detection based on a dominance-based rough set and fuzzy random forests, and the ensemble classifier achieved an accuracy of 77%. Ali et al. [12] applied an ensemble classifier on a dataset from a local medical center for diabetes diagnosis and achieved an accuracy of 81%. Bashir et al. [13] suggested a three-layer ensemble classifier based on a majority voting method. The proposed classifier achieved accuracies of 93% and 77.08% on the PIDD and the Biostat Diabetes Dataset, respectively. El-Baz et al. [14] introduced two neural network ensemble classifiers based on the cascade-forward back-propagation network (CFBN) and multilayer perceptron. Majority voting was used to integrate the ultimate prediction class of each classifier. The classifier based on the CFBN achieved an accuracy of 96.88%, while the other classifier achieved an accuracy of 95.31% with the PIDD dataset. Alehegn et al. [15] proposed an ensemble classifier for diabetes prediction using the SVM, NB, and

DT algorithms. The results based on the PIDD dataset showed that the proposed ensemble achieved an accuracy of 90.36%.

In Singh et al. [16], introduced a stacking-based evolutionary ensemble learning scheme for diabetes prediction using the PIDD dataset. A multi-objective optimization procedure was used for base-learner selection, while KNN was utilized as a meta-classifier that ensembled the predictions of the base-learners. Their comparison findings show that the proposed approach outperformed several single machine learning approaches and achieved the accuracy of 83.8%. Using machine learning techniques, Nilashi et al. [17] created a hybrid smart decision support framework for diabetes prediction. The experimental results on the PIDD dataset demonstrated that the proposed approach outperformed previously developed approaches in terms of prediction accuracy. Onan [18] presented a hybrid intelligent classification model for breast cancer diagnosis. The model consists of three phases: instance selection, feature selection and classification. The fuzzy-rough instance selection method based on weak gamma evaluator is utilized for instance selection. In feature selection, the consistency-based feature selection method is used in conjunction with a re-ranking algorithm. In the classification phase of the model, the fuzzy-rough nearest neighbor algorithm is utilized. The accuracy of the proposed approach is 99.7151%. Onan [19] presented a comprehensive analysis of Turkish sentiment analysis using nine supervised and unsupervised term weighting schemes. The predictive efficiency of term weighting schemes is explored using four supervised learning algorithms and three ensemble learning methods (AdaBoost, Bagging, and Random Subspace). Onan [20] presented an ensemble scheme based on hybrid supervised clustering for text classification. In the presented scheme, supervised hybrid clustering is utilized to partition the data samples of each class into clusters so that training subsets with higher diversities can be provided. Each classifier is trained on the diversified training subsets and the predictions of individual classifiers are combined by the majority voting rule. The authors in [21] proposed an approach for improving the diagnostic accuracy and reducing diagnostic misclassification based on the extracted significant diabetes features. The proposed approach is an integration between the SVM technique and K-means clustering algorithms to diagnose diabetes disease. The authors in [22] suggested a model for the diagnosis of breast cancer based on integrating RBF neural network with ensemble features using the boosting method. The highest diagnosis accuracy achieved by the proposed model is 98.4%.

Most existing stacking-based research has a significant issue with the improper integration of the base- and meta-learners. To address this limitation, we proposed a hybrid level-1 meta-learner that integrates fuzzy clustering and logistic regression to classify the predictions from level-0 base-learners. As a result, this research presents a stacking-based ensemble learning approach for diabetes prediction, in which the hybrid meta-learner was used for model combination.

## 3 Ensemble Learning for Diabetes Prediction Based on Hybrid Meta-Classifier of Fuzzy Clustering and Logistic Regression

The ensemble learning model is a well-known approach for improving performance by combining a group of classifiers [23,24]. This section explains the proposed approach for diabetes prediction, which consists of two levels. First, a base-learner comprising six machine learning algorithms is utilized for predicting diabetes. Second, a hybrid meta-learner that combines fuzzy clustering and logistic regression is employed to appropriately integrate predictions from the base-learners and to provide the final prediction of diabetes. The hybrid meta-learner generates highly significant clusters of predictions from the base-learners using the FCM algorithm, and then outputs the final diabetes prediction result by employing the predictions of the base-learners and their clusters resulting from FCM as inputs to the LR algorithm. Fig. 1 shows the proposed approach for the prediction of diabetes.

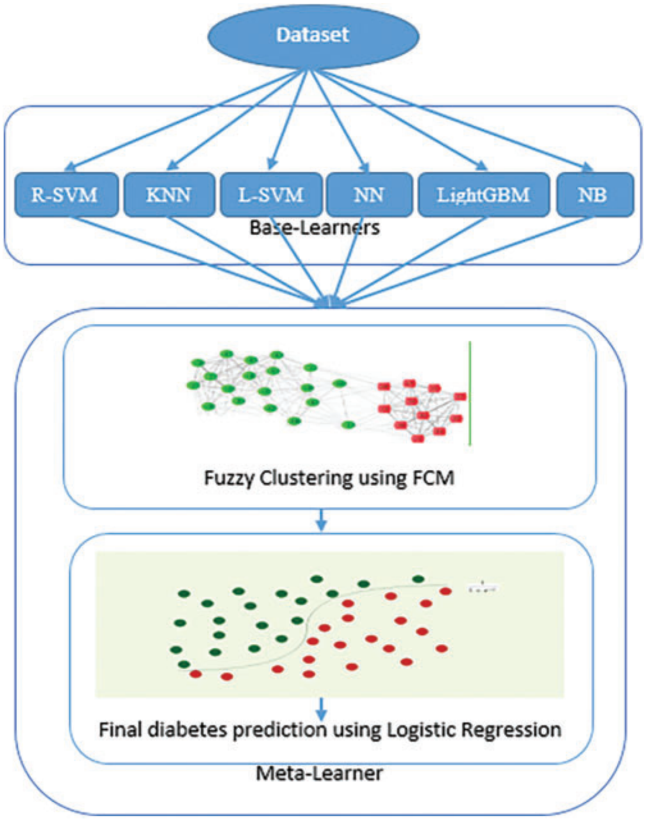
Dataset
R-SVM
KNN
L-SVM
NN
LightGBM
NB
Base-Learners
Fuzzy Clustering using FCM
Final diabetes prediction using Logistic Regression
Meta-Learner

**Figure 1:** The proposed approach for diabetes prediction

The overall process flow of the proposed approach is shown in Fig. 1 and summarized in the following steps:

Step 1: The dataset is preprocessed by removing the lost and incorrect values due to errors or deregulation, and then it is then divided into training and testing sets using the 5-fold cross-validation technique.

Step 2: The base learners (KNN, radial SVM, NB, linear SVM, NN and LightGBM) are trained and tested using the training and testing sets.

Step 3: FCM algorithm is utilized to cluster the prediction probabilities generated by the base learners.

Step 4: The Logistic Regression algorithm takes the prediction probabilities generated by the base learners and their clusters produced by the FCM as inputs, and provides the final diabetes prediction result.

### *3.1 Datasets and Data Preprocessing*

The proposed approach was implemented using the PIDD with type 2 diabetes from the University of California Irvine repository. Each dataset instance consists of class and eight features; the class indicates whether the patient has diabetes or not: the value of class “1” characterizes diabetic cases, and "0" characterizes nondiabetic cases. The features used in the dataset are the number pregnancies, plasma glucose concentration, diabetes pedigree function, triceps skin fold thickness, diastolic blood pressure, 2-h serum insulin, body max index, and age [25]. The accuracy of the data significantly affects the performance efficiency. This indicates that data preprocessing is critical to the success of the model [26]. The dataset was preprocessed by removing the lost and incorrect values due to errors or deregulation. The dataset contains 768 instances, 376 instances were deleted from the 768 in the PIDD dataset because they had one or more attributes with missing values. As a result of the data preparation, the PIDD dataset has a total of 392 instances. Among the 392 instances, 255 are non-diabetic, and the remaining 137 are diabetic cases. To ensure that the classification model was not simply fitted to the PIDD, the proposed hybrid ensemble approach was used to analyze Schorling Diabetes Dataset (SDD), a public diabetes dataset type 2 diabetes, donated by the Virginia School of Medicine [27]. The dataset includes data on 403 subjects who were interviewed as part of a study to determine the rate of diabetes, obesity, and other cardiovascular risk factors in central Virginia. Among the 403 instances, 330 are non-diabetic, and the remaining 73 are diabetic cases. In this study all the attributes from PIDD and SDD were used. Tab. 1 shows the independent attributes in the used datasets.

**Table 1:** Independent attributes used in PIMA and Schorling datasets

| Sr. no. | Independent attributes in PIMA indian dataset | Description of independent attributes |
|---|---|---|
| 1 | Pregnancy | Number of times a participant is pregnant |
| 2 | Glucose | Plasma glucose concentration in an oral glucose tolerance test |
| 3 | Diastolic blood pressure | Diastolic blood pressure (mm/Hg) |
| 4 | Skin thickness | Triceps skin fold thickness (mm). |
| 5 | Serum insulin | 2-h serum insulin (mu U/ml) |
| 6 | BMI | Body mass index (kg/m2) |
| 7 | Diabetes pedigree function | An attribute used in diabetes prognosis |
| 8 | Age | Participants' ages |
| | Independent attributes in schorling dataset | Description of independent attributes |
| 1 | Stab.glu | Stabilized glucose (mg/dL) |
| 2 | Age | Age (years) |
| 3 | Ratio | Cholesterol/High Density Lipoproteins (HDL) ratio |
| 4 | Waist | Waist (inches) |
| 5 | Chol | Total cholesterol (mg/dL) |
| 6 | Bp. s | Systolic blood pressure (mmHg) |
| 7 | Bp. d | Diastolic blood pressure (mmHg) |
| 8 | Frame | A factor with levels (small, medium, large) |
| 9 | Gender | Gender of subject (male, female) |
| 10 | Hdl | High density lipoprotein (mg/dL) |
| 11 | Height | Height (inches) |

(Continued)

**Table 1:** Continued

| Sr. no. | Independent attributes in PIMA indian dataset | Description of independent attributes |
|---|---|---|
| 12 | Hip | Hip (inches) |
| 13 | Weight | Weight (pounds) |

The K-fold cross-validation strategy is among the most widely used techniques for model selection and classifier error estimation [28]. The 5-fold cross-validation procedure was utilized in this research to split the dataset into training and testing sets. The 5-fold cross-validation is used in this research due to that 5-fold cross-validation generally performs quite well for middle size datasets [29]. In addition, Ayon et al. [30] applied both the five-fold and ten-fold cross-validation procedure on the same dataset and found that the model performed better using the five-fold procedure.

The dataset was split into five mutually exclusive subsamples of equal size. The proposed approach was trained five times. Each time, we used four folds for the training and left one fold for testing. This approach has the advantage of reducing the bias associated with random sampling [31].

### *3.2 The Base-Learner Model*

Several studies have shown that the accuracy and variety of the base-learners influence the effectiveness of any ensemble of classifiers [32]. Thus, an ensemble of classifiers might enhance the accuracy of any of its base-learners if their errors are not coincident (are diverse) and they have a low error rate (are accurate).

To obtain the diversity, the proposed approach used heterogeneous base-learner model consists of six different machine learning algorithms. These algorithms were selected because they represent a wide variety of fields: neural networks, probabilistic models, statistical models, decision trees and ensemble learning. Moreover, the prediction models generated with these algorithms presented a good performance in several previous works.

The 5-fold cross-validation procedure was utilized to divide the dataset into five training and testing subsets, which increases the diversity of the input samples of the base learner. Then base learners, i.e., KNN, radial SVM, NB, linear SVM, NN and LightGBM take the five training and testing subsets as inputs and output the initial predictions of diabetes. The predictions of base-learners and their fuzzy clusters are then employed as inputs to the Logistic Regression (LR) algorithm, to achieve more generalized performance and reliability. LR algorithm generates the final diabetes prediction result.

*Neural Networks:*

The Artificial Neural Networks (ANN) technique is a well-known machine learning approach for dealing with complicated pattern-oriented problems in both categorization and time-series data types. An ANN algorithm has three major levels: the input layer, the hidden layer(s), and the output layer. By recognizing the intrinsic connections between different features, ANNs attempt to prepare a mapping between the input layer and the output layer. The hidden layer(s) analyses the information obtained from the input layer and then sends it to the output layer [33].

k-nearest neighbors

k-nearest neighbors (KNN) is a technique that uses a similarity measure such as distance functions to assign a new feature vector to a class in all possible cases. After determining the distance between the feature vector and all training samples, the new case is allocated to the class with the greatest probability.

*Naive Bayes (NB)*

NB is a classifier that is based on the Bayes theorem. It employs the concept of probability and implies that features are independent of each other [34]. Assume C is a class variable with n features. Then, according to Bayes' theorem, the probability of an instance belonging to class C is given by Eq. (1).

$$p(c|k_1, k_2, k_n) = \frac{(p(c)p(k_1, k_2, k_n|c))}{(p(k_1, k_2, k_n))} \tag{1}$$

where the value of instances for feature $i$ is represented by $\mathrm{k}_i$. Each feature is presumed to be independent of the others.

The highest posterior of the classification variable is chosen using Eq. (2).

$$arg_c\ maxp\ (C = c) \prod_{i=1}^{n} p(K_i = k_i\ |C = c) \tag{2}$$

Support vector machine (SVM)

Decision hyperplanes are employed in SVM classification. In input space or high-dimensional space, these hyperplanes define the decision boundaries. SVM creates linear functions (hyperplanes) using labeled training samples in order to divide data into two groups (positive or negative). The samples closest to the hyperplanes are known as support vectors. The margin of the SVM is defined as the distance between the support vectors and hyperplanes. SVM aims to increase this margin as much as possible [35].

*Light Gradient Boosting Machine (LightGBM)*

LightGBM is a boosting approach for improving a model's performance by merging a group of weak classifiers into a strong classifier. The concept entails selecting weak classifiers in such a way that their performance is considerably enhanced when combined. Gradient boosting is what LightGBM is all about. It starts with decision trees as a weak learner and then uses gradient boosting to iteratively fit a sequence of trees. LightGBM is a decision tree-based model that is built leaf-by-leaf instead of depth-by-depth (as in other decision tree-based methods). As a result of this leaf-by-leaf generation, more complex trees are produced with greater accuracy [36]. LightGBM is a strong machine-learning approach that produces cutting-edge results in a wide range of practical problems. It is utilized in this research because it has the ability to strategically evaluate the previously constructed decision trees and construct new decision trees to enhance results. Therefore, it has the advantage of potentially improving the prediction accuracy.

### *3.3 Hybrid Meta-Classifier of Fuzzy Clustering and Logistic Regression*

#### *3.3.1 Fuzzy C-Means Clustering Algorithm*

Clustering is an unsupervised machine learning approach for identifying natural groupings or patterns in a dataset. The clustering methods ensure that all of the allocated observations in a group or cluster are related. Clustering methods can also aid in the discovery of hidden information in the data [37]. As a result, combining base-learner predictions with their clusters can improve the hybrid meta-classification learner's performance and offer a more accurate diabetes prediction. Clustering methods divide the samples into distinct clusters so that samples inside a single cluster are more closely linked to one another than those allocated to dissimilar clusters [38]. The FCM algorithm is among the most widely used clustering algorithms, and it is a clustering method that allows a single piece of data to be assigned to two or more clusters. Dunn [39] introduced this method, which was later enhanced by Bezdek [40]. It is commonly employed for pattern recognition. In the FCM approach, each provided data item does not belong exclusively to a well-defined cluster, but it can be located in a middle way. Probability coefficients reflect this characteristic of belonging to more than one cluster.

The FCM is based on the objective function minimization described below:

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \,.\, \|x_i - \ C_j\|^2, \tag{3}$$

where $u_{ij}$ denotes the probability that person $x_i$ is part of the *j*th cluster, $x_i$ denotes the *i*th person, $C_j$ denotes the coordinates of the cluster's *j*th center, || || denotes the Euclidean norm, N denotes the number of persons in the database, *m* presents the index of fuzziness ($m = 2$) and C denotes the total number of considered clusters. Fuzzy partitioning is accomplished by iterative optimization of the objective function, with membership $u_{ij}$ and cluster centers $C_j$ updated by

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}} \tag{4}$$

$$C_j = \frac{\sum_{i=1}^{N} u_{ij} \ .\ \ x_i}{\sum_{i=1}^{N} u_{ij}} \tag{5}$$

The probability coefficients reflect that a data item can belong to several clusters. Thus, person $x_i$ has a probability of $u_{ij}$ being a member of cluster j considering that

$$\sum_{j=1}^{C} u_{ij} = 1 \tag{6}$$

The FCM algorithm assists in the detection of hidden data. Therefore, integrating base-learner predictions with their clusters can increase the performance of the hybrid meta-classification learner and provide a more accurate diabetes prediction. FCM employed to obtain the membership values of the base-learner predictions' probabilities to which of the two clusters (diabetic and non-diabetic) they belong. The prediction probability shares a similarity with each class or cluster is represented using membership functions whose value ranges between 0 and 1. In this paper, the membership function's value of the prediction probability for a class is computed using Eq. (4), based on an iterative measure of the distance between the prediction probability and the class center using Eq. (5). Membership function value close to zero indicates a low degree of similarity between the prediction probability and the class. While if the membership function value is unity, it implies a high degree of similarity between the prediction probability and the class. Tab. 2 shows examples for the FCM clusters' results and prediction probabilities from the base-learners. These membership values are then used with the predictions' probabilities from the base-learner in the second stage of the proposed approach, as additional informative features to improve the classification process completed by LR classifier. which utilized as the input data to the logistics regression (LR) classifier.

**Table 2:** FCM clustering results based on predictions' probabilities from the base-learner

| Prediction by (Linear SVM) | Predictions by (NB) | Prediction by (NN) | Prediction by (KNN) | Prediction by (Radial SVM) | Prediction by *(Light-GBM)* | Cluster 1 membership | Cluster 0 membership | Final cluster |
|---|---|---|---|---|---|---|---|---|
| 0.0362 | 0.0004 | 0.0000 | 0.0383 | 0.0000 | 0.0304 | 0.0145 | 0.9854 | 0 |
| 0.0217 | 0.0001 | 0.0001 | 0.0026 | 0.0000 | 0.0507 | 0.0121 | 0.9878 | 0 |
| 0.1685 | 0.5000 | 0.0207 | 0.0401 | 0.0160 | 0.2354 | 0.8469 | 0.1530 | 1 |
| 0.0215 | 0.0196 | 0.0020 | 0.0005 | 0.0000 | 0.0417 | 0.0142 | 0.9857 | 0 |
| 0.2419 | 0.5395 | 0.0086 | 0.1728 | 0.1044 | 0.1727 | 0.9227 | 0.0772 | 1 |
| 0.0194 | 0.0125 | 0.0051 | 0.0119 | 0.0000 | 0.0481 | 0.0116 | 0.9883 | 0 |
| 0.1763 | 0.0612 | 0.0040 | 0.1231 | 0.0479 | 0.3956 | 0.3857 | 0.6142 | 0 |
| 0.0232 | 0.0741 | 0.0021 | 0.0273 | 0.0000 | 0.0686 | 0.0072 | 0.9927 | 0 |
| 0.1945 | 0.0423 | 0.0090 | 0.1714 | 0.9228 | 0.2371 | 0.5549 | 0.4450 | 1 |

### *3.3.2 LR Algorithm*

The classification method aims to create a model capable of assigning data elements to a specific class based on the current data. It was used to derive important elements from the model or forecast the data trend. Usually, the dependent variable of the logistic regression algorithm is binary classification. In other words, the logistic regression method is often employed to solve two-category problems. The primary objective of our experiment is to predict whether an individual has diabetes, and this is a classic binary classification issue.

We selected logistic regression for our research from several supervised machine learning techniques. Logistic regression is a well-established technique that produces simple-to-understand models that have been demonstrated to be successful in a variety of situations and cases [41]. Additionally, logarithmic regression returns the probability of an instance being belongs to class rather than a label, which provides additional insight into the prediction.

The logistic regression method is built on a linear regression model, which is explained in Eq. (7)

$$P = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m \tag{7}$$

The classification issue is similar to the linear regression issue. Only continuous values can be predicted using a linear regression. As the predictive value of the classification problem can only be 0 or 1, we can establish a critical point. If the value exceeds the threshold, 1 is returned; otherwise, 0 is returned. Logistic regression is a type of regression model in which the prediction scope is reduced, and the prediction score is limited to [0, 1]. Based on linear regression, logistic regression incorporates a sigmoid function level (non-linearity). Initially, the attributes are linearly added together before being predicted using the sigmoid function. The key formulas of the logistic regression algorithm are shown in Eqs. (8)–(10).

$$\Pr(Y = +1|X) \sim \beta.X \; and \; Pr(Y = -1|X) = 1 - Pr(Y = +1|X) \tag{8}$$

$$\downarrow \ \sigma(x) := \frac{1}{1+e^{-x}} \in [0,1] \ \ (\textit{the sigmoid function}) \tag{9}$$

$$\Pr(Y=+1|X) \sim \ \sigma(\beta.X) \ \textit{and} \ Pr(Y=-1|X) = 1 - Pr(Y=+1|X) \tag{10}$$

There are two classes in this research: diabetic and nondiabetic. The letter Y denotes that the person has diabetes. The features in the datasets are represented by X independent variables. Each dependent variable X is associated with the coefficient value $\beta$, which is referred to as the weight coefficient. After the logistic regression algorithm processed the dataset, the weight value for each variable was calculated. Different weights correspond to different degrees of similarity between X and Y.

In this research, the prediction probabilities from the base learners (KNN, radial SVM, NB, linear SVM, NN and LightGBM) and their clusters generated by the Fuzzy C-means algorithm, were utilized as input to the Logistic regression algorithm, to get the final diabetes prediction result.

### 3.4 Evaluation Metrics

In our experiments, we evaluated the classification efficiency of the proposed hybrid ensemble scheme and comparison models using five performance assessment measures: accuracy, recall, precision, f1-measure, and AUC. These measures were determined as shown in Tab. 3.

**Table 3:** Performance evaluation metrics

| Metrics of performance | Mathematical formula | Notes |
|---|---|---|
| Accuracy | $\frac{TP+TN}{TP+TN+FP+FN}$ | Calculates the percentage of properly classified instances in total. |
| Recall | $\frac{TP}{TP+FN}$ | Calculates the proportion of diabetic samples that are correctly classified. |
| Precision | $\frac{TP}{TP+FP}$ | Calculates the ratio between all the diabetic samples that are correctly classified and the total number of diabetic samples. |
| F-measure | $2 * \frac{Precision * Recall}{Precision + Recall}$ | Precision and recall are averaged harmonically. |
| Area under receiver operating characteristic curve (AUC) | $\frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$ | Determines the diagnostic ability of the scheme to distinguish between persons with diabetes and without diabetes. |

## 4 Results and Discussion

Our proposed ensemble approach outperformed all the base-learner algorithms with the PIDD dataset, and achieved the highest accuracy value of 99%, followed by the soft voting ensemble method (98%), LightGBM classifier (93.46%), and hard voting ensemble method (90%). The worst accuracies were shown by NN (64.59%) and linear SVM (75.30%). In recall terms, the proposed approach achieved the highest value of 99%, followed by the soft voting ensemble method (97.21%) and LightGBM (82.38%). In precision terms, the suggested approach attained the highest value of 99%, followed by LightGBM (98.97%) and soft voting ensemble (97.51%). The proposed approach also achieved an average f-measure of 98%, outperforming the soft voting ensemble (97.55%) and the LightGBM classifier (89.88%).

Using the SDD, the proposed approach also outperformed all the base-learner algorithms and other ensemble methods. The proposed approach achieved the highest accuracy value of 95.20%, followed by the NB classifier (94.89%), the soft voting ensemble method (94.31%), and the hard voting ensemble method (93.38%). In recall terms, the proposed approach achieved the highest value of 95.40%, followed by the soft voting ensemble method (94.39), and the linear SVM classifier (93.75%). In precision terms, the suggested approach attained the highest value of 95.10%, followed by the soft voting ensemble (94.18%) and hard voting ensemble (93.41%). The proposed approach also achieved an average f-measure of 95.11%, outperforming the soft voting ensemble (94.01%) and hard voting ensemble (92.02%).

The promising and competitive performance results of the proposed approach demonstrate its superiority over traditional ensemble methods. The capability of the proposed approach to combine fuzzy clustering and logistic regression enables it to appropriately integrate the predictions from base-learners and provide a more accurate prediction of diabetes. As a result, the proposed hybrid approach outperforms the six individual classifiers and two ensemble approaches in terms of classification efficiency, as shown in Tab. 4.

**Table 4:** Performance comparison between the proposed ensemble approach, the individual classifiers and other ensemble methods

| Dataset | Approach | Accuracy | Recall | Precision | f1 score | AUC |
|---|---|---|---|---|---|---|
| Pima Indians Diabetes Database (PIDD) | KNN | 0.8203 | 0.7964 | 0.8780 | 0.8352 | 0.9450 |
| | Radial SVM | 0.7599 | 0.7683 | 0.6634 | 0.7120 | 0.8850 |
| | Naive bayes | 0.7582 | 0.6888 | 0.6048 | 0.6441 | 0.8210 |
| | Linear SVM | 0.7530 | 0.7134 | 0.5951 | 0.6489 | 0.8410 |
| | Neural network | 0.6459 | 0.7743 | 0.7463 | 0.7598 | 0.9090 |
| | LightGBM | 0.9346 | 0.8238 | 0.9897 | 0.8988 | 0.9886 |
| | Soft voting | 0.9800 | 0.9721 | 0.9751 | 0.9755 | 0.9790 |
| | Hard voting | 0.9000 | 0.8752 | 0.9134 | 0.8900 | 0.9194 |
| | Proposed ensemble approach | 0.9931 | 0.9910 | 0.9920 | 0.9814 | 0.9910 |
| Schorling Diabetes Dataset (SDD) | KNN | 0.8203 | 0.7964 | 0.5001 | 0.6250 | 0.7370 |
| | Radial SVM | 0.9285 | 0.8871 | 0.4021 | 0.5517 | 0.6940 |
| | Naive bayes | 0.9489 | 0.8947 | 0.8532 | 0.8717 | 0.8210 |
| | Linear SVM | 0.9387 | 0.9375 | 0.7527 | 0.8333 | 0.8690 |
| | Neural network | 0.9356 | 0.9165 | 0.6536 | 0.6451 | 0.7190 |
| | LightGBM | 0.9046 | 0.8238 | 0.8897 | 0.8988 | 0.9301 |
| | Soft voting | 0.9431 | 0.9439 | 0.9418 | 0.9401 | 0.9217 |

(Continued)

**Table 4:** Continued

| Dataset | Approach | Accuracy | Recall | Precision | f1 score | AUC |
|---|---|---|---|---|---|---|
| | Hard voting | 0.9338 | 0.9317 | 0.9341 | 0.9202 | 0.8631 |
| | Proposed ensemble approach | 0.9520 | 0.9540 | 0.9510 | 0.9511 | 0.9410 |

The AUC was generated for both the individual and ensemble classifiers to demonstrate the separation and discrimination capabilities of the models and to compute their specificity and sensitivity at various class prediction score thresholds. In general, an AUC curve that is increasingly closer to the top-left corner demonstrates an improved classification efficiency. Fig. 2 shows that the proposed hybrid ensemble approach performs better than the six-base-learners and the other two ensemble methods.

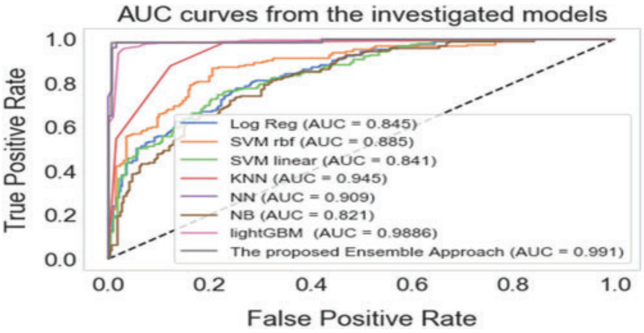

**Figure 2:** The areas under the receiver operating curves (AUCs) of the proposed ensemble approach and the individual classifiers

When combined, precision and recall are valuable metrics to use for unbalanced data; precision denotes the appropriateness of the result scale and its closeness to the anticipated solution, while recall denotes the number of related results. A high recall score indicates a low rate of false negatives, while a high precision score indicates a low rate of false positives. High precision and recall scores suggest that the classifier returns results accurately and retrieves the majority of positive results [42]. As a result, the precision-recall curve provides a comprehensive picture of the classifier's accuracy and is efficient even when the dataset is unbalanced [43]. Fig. 3 depicts the precision-recall curves for the proposed scheme and the other machine learning algorithms. The precision-recall curve of the proposed approach is positioned closest to the top-right corner of the graph, indicating that the proposed ensemble hybrid approach for diabetes prediction performed better than the rest of the approaches.

Moreover, we compared the performance of our proposed approach with experiments of other researchers using the same dataset PIDD to show that our proposed approach achieved a certain level of improvement. The proposed ensemble approach achieved an accuracy of 99%. The results of the other studies are listed in Tab. 5.

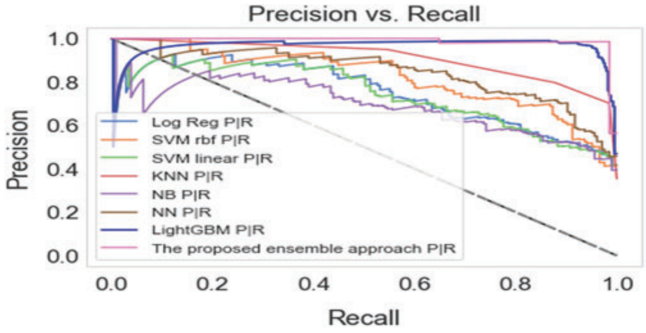

**Figure 3:** Precision-recall curves of the proposed approach and the base-learners

**Table 5:** Performance comparison between the proposed ensemble approach and other studies

| No | Reference | Method | Accuracy |
|---|---|---|---|
| 1 | [10] | Stack of SVM and ANN | 88.04% |
| 2 | [11] | Ensemble classifier based on dominance-based rough set and fuzzy random forests | 77% |
| 3 | [13] | Three-layer ensemble classifier based on a majority voting method | 93% |
| 4 | [14] | Two neural network ensemble classifiers based on cascade-forward back-propagation network (CFBN) and multilayer perceptron (MLP) | 96.88% |
| 5 | [15] | Ensemble classifier based on SVM, naive bayes, and decision tree algorithms | 90.36% |
| 6 | [16] | Method of evolutionary ensemble learning based on stacking | 83.8% |
| 7 | [44] | Competitive co-evolutionary neural networks | 78.2% |
| 8 | [45] | Hybrid system | 80.99% |
| 9 | [46] | Weighted multilayer classifier ensemble approach | 78.21 |
| 10 | This research | Ensemble learning based on hybrid meta-classifier of fuzzy clustering and logistic regression | 99% |

## 5 Conclusions and Recommendations

Diabetes mellitus refers to a group of metabolic disorders that are defined by persistently high blood glucose levels. Diabetes that remains undiagnosed may result in a variety of complications, including retinopathy, nephropathy, neuropathy, and other vascular disorders. This paper proposes a novel ensemble learning method for diabetes prediction that is based on a hybrid meta-classifier composed of fuzzy clustering and logistic regression. The proposed method is divided into two stages. The first stage is a base-learner, which consists of six machine learning algorithms for predicting diabetes. Second, a hybrid meta-learner is used to incorporate predictions from base-learners to provide the final prediction of diabetes. The hybrid meta-learner employs the FCM algorithm to generate highly significant clusters of predictions from base-learners. The predictions of the base-learner and their fuzzy clusters are then employed as inputs to the LR algorithm, which generates the final diabetes prediction result. Experiments were performed using two publicly available diabetes datasets to demonstrate the effectiveness of the proposed method for predicting diabetes. Comparing the proposed approach to other approaches revealed that the proposed approach outperformed the others and achieved the highest prediction accuracies of 99% and 95.20% using the PIDD and SDD datasets, respectively. Future research will focus on effective approaches for weighting base classifiers according to cluster attributes. The limitation of this study is considering only one dataset, it is recommended to consider various category diabetes datasets that have different sizes, and include all thefactors that may contribute to the development of diabetes mellitus.

**Acknowledgement:** The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia "for funding this research work through the project number IFPHI-193-830-2020" and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

**Funding Statement:** The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia "for funding this research work through the project number IFPHI-193-830-2020" and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] N. H. Cho, J. E. Shaw, S. Karuranga, Y. D. Huang, J. D. da Rocha Fernandes *et al.,* "IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Research and Clinical Practice*, vol. 138, no. 1, pp. 271–281, 2018.

[2] M. Maniruzzaman, M. J. Rahman, B. Ahammed and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Health Information Science and Systems*, vol. 8, no. 1, pp. 1–14, 2020.

[3] B. A. Mateen, A. L. David and S. Denaxas, "Electronic health records to predict gestational diabetes risk," *Trends in Pharmacological Sciences*, vol. 41, no. 5, pp. 301–304, 2020.

[4] I. Kavakiotis, T. Olga, A. Salifoglou, N. Maglaveras, I. Vlahavas *et al.,* "Machine learning and data mining methods in diabetes research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017.

[5] A. Alharbi and M. Alghahtani, "Using genetic algorithm and ELM neural networks for feature extraction and classification of type 2-diabetes mellitus," *Applied Artificial Intelligence*, vol. 33, no. 4, pp. 311–328, 2019.

[6] L. I. Kuncheva, J. C. Bezdek and R. P. Duin, "Decision templates for multiple classifier fusion: An experimental comparison," *Pattern Recognition*, vol. 34, no. 2, pp. 299–314, 2001.

[7] P. Shunmugapriya and S. Kanmani, "Optimization of stacking ensemble configurations through artificial bee colony algorithm," *Swarm and Evolutionary Computation*, vol. 12, no. 1, pp. 24–32, 2013.

[8] Y. Chen, M. L. Wong and H. Li, "Applying ant colony optimization to configuring stacking ensembles for data mining, " *Expert Systems with Applications*, vol. 41, no. 6, pp. 2688–2702, 2014.

[9] D. S. Nascimento, A. L. Coelho and A. M. Canuto, "Integrating complementary techniques for promoting diversity in classifier ensembles: A systematic study," *Neurocomputing*, vol. 138, no. 1, pp. 347–357, 2014.

[10] R. Zolfaghari, "Diagnosis of diabetes in female population of pima Indian heritage with ensemble of bp neural network and svm," *International Journal of Computational Engineering & Management*, vol. 15, no. 4, pp. 2230–7893, 2012.

[11] E. Saleh, J. Błaszczyński, A. Moreno, A. Valls, P. Romero-Aroca *et al.,* "Learning ensemble classifiers for diabetic retinopathy assessment," *Artificial Intelligence in Medicine*, vol. 85, no. 3, pp. 50–63, 2018.

[12] R. Ali, M. H. Siddiqi, M. Idris, B. H. Kang and S. Lee, "Prediction of diabetes mellitus based on boosting ensemble modeling," in *Proc. of Int. Conf. on Ubiquitous Computing and Ambient Intelligence*, Belfast, UK, pp. 25–28, 2014.

[13] S. Bashir, U. Qamar, F. H. Khan and L. Naseem, "HMV: A medical decision support framework using multi-layer classifiers for disease prediction," *Journal of Computational Science*, vol. 13, no. 1, pp. 10–25, 2016.

[14] A. H. El-Baz, A. E. Hassanien and G. Schaefer, "Identification of diabetes disease using committees of neural network-based classifiers," in *Proc. Machine Intelligence and Big Data in Industry*, Cham, Switzerland, pp. 65–74, 2016.

[15] M. Alehegn, R. Joshi and P. Mulay, "Analysis and prediction of diabetes mellitus using machine learning algorithm," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 9, pp. 871–878, 2018.

[16] N. Singh and P. Singh, "Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 1–22, 2020.

[17] M. Nilashi, O. Ibrahim, M. Dalvi, H. Ahmadi and L. Shahmoradi, "Accuracy improvement for diabetes disease classification: A case on a public medical dataset," *Fuzzy Information and Engineering*, vol. 9, no. 9, pp. 345–357, 2017.

[18] A. Onan, "Fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer," *Expert Systems with Applications*, vol. 4, no. 20, pp. 6844–6852, 2015.

[19] A. Onan, "Ensemble of classifiers and term weighting schemes for sentiment analysis in turkish," *Scientific Research Communications*, vol. 1, no. 1, pp. 1–12, 2021.

[20] A. Onan, "Hybrid supervised clustering based ensemble scheme for text classification," *Kybernetes*, vol. 1, no. 1, pp. 330–348, 2017.

[21] A. H. Osman and H. M. Aljahdali, "Diabetes disease diagnosis method based on feature extraction using K-sVM," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 1, pp. 236–244, 2017.

[22] A. H. Osman and H. M. Aljahdali, "An effective of ensemble boosting learning method for breast cancer virtual screening using neural network model," *IEEE Access*, vol. 8, pp. 39165–39174, 2020.

[23] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks," *Journal of Biomedical Informatics*, vol. 86, no. 1, pp. 25–32, 2018.

[24] S. L. Hsieh, S. H. Hsieh, P. H. Cheng, C. H. Chen, K. P. Hsu *et al.,* "Design ensemble machine learning model for breast cancer diagnosis," *Journal of Medical Systems*, vol. 36, no. 5, pp. 2841–2847, 2012.

[25] C. Blake, "UCI repository of machine learning databases," 1998. [Online]. Available: http://archive.ics.uci.edu/ml/index.php.

[26] K. M. Orabi, Y. M. Kamal and T. M. Rabah, "Early predictive system for diabetes mellitus disease," in *Proc. Industrial Conf. on Data Mining*, Cham, Switzerland, pp. 420–427, 2016.

[27] J. P. Willems, J. T. Saunders, D. E. Hunt and J. B. Schorling, "Prevalence of coronary heart disease risk factors among rural blacks: A community-based study," *Southern Medical Journal*, vol. 90, no. 8, pp. 814–820, 1997.

[28] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, no. 1, pp. 40–79, 2010.

[29] C. Y. Ma, S. Y. Yang, H. Zhang, M. L. Xiang, Q. Huang *et al.,* "Prediction models of human plasma protein binding rate and oral bioavailability derived by using GA–CG–SVM method," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 47, no. 4–5, pp. 677–682, 2008.

[30] S. I. Ayon and M. Islam, "Diabetes prediction: A deep learning approach, "*International Journal of Information Engineering & Electronic Business*, vol. 11, pp. 21–7, 2019.

[31] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000.

[32] H. Wu, S. Yang, Z. Huang, J. He and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics in Medicine Unlocked*, vol. 10, no. 1, pp. 100–107, 2018.

[33] G. S. Fesghandis, A. Pooya, M. Kazemi and Z. N. Azimi, "Comparison of multilayer perceptron and radial basis function neural networks in predicting the success of new product development," *Engineering, Technology & Applied Science Research*, vol. 7, no. 1, pp. 1425–1428, 2017.

[34] Z. Zhang, Q. Zhu and Y. Xie, "A novel image matting approach based on naive Bayes classifier," in *Proc. Int. Conf. on Intelligent Computing*, Springer, Berlin, Heidelberg, pp. 433–441, 2012.

[35] F. F. Chamasemani and Y. P. Singh, "Multi-class support vector machine (SVM) classifiers–an application in hypothyroid detection and classification," in *Proc. 2011 Sixth Int. Conf. on Bio-Inspired Computing: Theories and Applications*, Penang, Malaysia, IEEE, pp. 351–356, 2011.

[36] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen *et al.,* "LightGBM: A highly efficient gradient boosting decision tree, " in *Proc. 31st Annual Conference on Neural Information Processing Systems*, California, USA, pp. 3146–3154, 2017.

[37] F. Franceschi, M. Cobo and M. Figueredo, "Discovering relationships and forecasting PM10 and PM2. 5 concentrations in bogotá, Colombia, using artificial neural networks, principal component analysis, and k-means clustering," *Atmospheric Pollution Research*, vol. 9, no. 5, pp. 912–922, 2018.

[38] G. Gan, C. Ma and J. Wu, "Fuzzy clustering algorithms," in *Data Clustering: Theory, Algorithms, and Applications*, 2nd ed., vol. 1. Philadelphia, USA: Society for Industrial and Applied Mathematics, pp. 152–159, 2020.

[39] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, pp. 32–57, 1973.

[40] J. C. Bezdek, "Pattern recognition with fuzzy objective function algorithms," *Springer Science & Business Media*, 2013.

[41] M. Zhu and M. Gribskov, "Mipepid: MicroPeptide identification tool using machine learning," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–11, 2019.

[42] M. A. Al-Shabi, "Credit card fraud detection using autoencoder model in unbalanced datasets," *Journal of Advances in Mathematics and Computer Science*, vol. 33, no. 5, pp. 1–16, 2019.

[43] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. The 23rd Int. Conf. on Machine Learning*, Pennsylvania, USA, pp. 233–240, 2006.

[44] B. Y. Hiew, S. C. Tan and W. S. Lim, "A double-elimination-tournament-based competitive co-evolutionary artificial neural network classifier," *Neurocomputing*, vol. 249, no. 2, pp. 345–356, 2017.

[45] M. R. Nalluri and D. S. Roy, "Hybrid disease diagnosis using multiobjective optimization with evolutionary parameter optimization," *Journal of Healthcare Engineering*, vol. 1, pp. 1–27, 2017.

[46] S. Bashir, U. Qamar and F. H. Khan, "Intellihealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework," *Journal of Biomedical Informatics*, vol. 59, no. 2, pp. 185–200, 2016.