


Auto-MeDiSine: an auto-tunable medical decision support engine using an automated class outlier detection method and AutoMLP

Maham Jahangir¹ · Hammad Afzal¹ · Mehreen Ahmed¹ · Khawar Khurshid¹ · Muhammad Faisal Amjad¹ · Raheel Nawaz² · Haider Abbas¹ 

Abstract

With advanced data analysis techniques, efforts for more accurate decision support systems for disease prediction are on the rise. According to the World Health Organization, diabetes-related illnesses and mortalities are on the rise. Hence, early diagnosis is particularly important. In this paper, we present a framework, Auto-MeDiSine, that comprises an automated version of enhanced class outlier detection using a distance-based algorithm (AutoECODB), combined with an ensemble of automatic multilayer perceptron (AutoMLP). AutoECODB is built upon ECODB by automating the tuning of parameters to optimize outlier detection process. AutoECODB cleanses the dataset by removing outliers. Preprocessed dataset is then used to train a prediction model using an ensemble of AutoMLPs. A set of experiments is performed on publicly available Pima Indian Diabetes Dataset as follows: (1) Auto-MeDiSine is compared with other state-of-the-art methods reported in the literature where Auto-MeDiSine realized an accuracy of 88.7%; (2) AutoMLP is compared with other learners including individual (focusing on neural network-based learners) and ensemble learners; and (3) AutoECODB is compared with other preprocessing methods. Furthermore, in order to validate the generality of the framework, Auto-MeDiSine is tested on another publicly available BioStat Diabetes Dataset where it outperforms the existing reported results, reaching an accuracy of 97.1%.

Keywords Classification · Disease prediction · Machine learning · Multilayer perceptron · Outlier detection

1 Introduction

The performance of Medical Expert Systems is continuously being improved, especially by application of novel (more accurate) pattern recognition and classification techniques. Machine learning algorithms have improved diagnostic systems that help to minimize the cost of conducting extensive medical tests. These systems not only help improve the diagnostic process, but also save the time of medical practitioners. Intelligent diagnostic systems have been applied to a range of complex diseases including cancer, liver disease, heart disease and diabetes [7, 15, 32, 48, 52, 55, 61]. In particular, during the last few decades, diabetes has become worryingly common.

WHO estimates the global number of adults suffering from diabetes to be 422 million.¹ Therefore, automated diagnosis tools tailored for diabetes are required to detect the disease at an early stage.

A number of predictive frameworks using various classification techniques such as artificial neural network (ANN), naïve Bayes (NB), support vector machine (SVM), decision trees (DT) and others are reported in the literature [35, 52, 56, 61]. Our detailed literature review indicated that ANNs achieve the best results in terms of accuracy of results. Several ANN-based frameworks have been reported for a variety of medical diagnostic tasks [8, 11, 36] demonstrating the modeling flexibility and high accuracy of the ANN approach [37, 50]. However, one of the major issues with network architectures is the optimization of parameters such as the number and composition of hidden layers, learning rate, epochs and other aspects of network topology. These parameters are to be decided before training the ANN.

✉ Haider Abbas
haider@mcs.edu.pk

¹ National University of Sciences and Technology, Islamabad, Pakistan

² Manchester Metropolitan University, Manchester, UK

¹ <http://www.who.int/diabetes/en/>.

AutoMLP, which is an auto-tunable ensemble of multilayer perceptrons (MLPs), addresses this issue by enabling automated optimization of the above parameters.

Noise in dataset due to outliers is another major challenge faced during the process of predictive modeling. The existence of outliers in dataset results in predictive models with low accuracy. The detection of outliers at preprocessing stage can cleanse the data, thus improving the performance of prediction model. Enhanced class outlier distance-based (ECODB) is a state-of-the-art outlier detection method that uses class information along with the disparity in values of attributes while detecting outliers. In particular, ECODB uses probability, deviation and distance of a particular record (with respect to the class labels of its K nearest neighbors) to detect the outlier. Varying the number and values of these factors while measuring deviation can change the performance of the prediction model. We automated the process of tuning the number of neighbors, the number of outliers and the distance metric used, thus optimizing the process of outlier detection to achieve better performance. This method is named as AutoECODB.

The proposed framework constitutes a hybrid prediction model, which deploys a combination of AutoECODB (at preprocessing) and an ensemble of AutoMLP. The framework is named as Auto-MeDiSine: auto-tunable medical decision support engine. The experiments are conducted on publicly available Pima Indians Diabetes Dataset (PIDD) which is used as a benchmark in order to compare our technique with existing state-of-the-art approaches. A preliminary study on this framework is provided in [28]. A series of experiments are conducted to evaluate the proposed Auto-MeDiSine by comparing it with other individual learners (particularly focusing on neural network based) as well as ensemble learners. In order to validate the effectiveness of the preprocessing technique, AutoECODB is compared with different preprocessing techniques such as feature selection, attribute weight generation, normalization, sampling and other outlier detection methods. Results indicate that Auto-MeDiSine outperforms other reported techniques and achieves the highest accuracy (88.7%). Furthermore, in order to showcase the generality of the proposed approach, Auto-MeDiSine is applied on another diabetes dataset, i.e., BioStat, where it outperformed the existing best reported results, showing an accuracy of 97.1%. A preliminary work has been published in [28].

The key contributions of the paper are summarized below:

- A novel framework (Auto-MeDiSine) comprising auto-tunable techniques at preprocessing and predictive modeling phase: AutoECODB for preprocessing and AutoMLP for predictive modeling.
- AutoECODB automates the existing ECODB method for outlier detection.

- Auto-MeDiSine produces best results on a benchmark dataset on diabetes, i.e., PIDD, achieving an accuracy of 88.7%.
- The generality of proposed Auto-MeDiSine is showcased as it achieves the best accuracy of 97.1% on another dataset, i.e., BioStat.

The rest of the paper is organized as follows. Section 2 presents a detailed literature review covering the previous studies on diabetes prediction; Sect. 3 provides a brief description of dataset used in this study; Sect. 4 discusses the proposed framework in detail; and the experimental details are covered in Sect. 5 along with discussion of results. Finally, the conclusions are presented in Sect. 6.

2 Literature review

This section presents related work that employs machine learning techniques in the design of intelligent healthcare applications, particularly for the prediction of diabetes. We have primarily focused on studies that use preprocessing techniques before applying the learners as they closely associate with our proposed method. The literature survey shows that Pima Indians Diabetes Dataset.² (PIDD) is the most commonly used dataset for research in diabetes prediction. This is a benchmark dataset, commonly used to compare prediction models. There are other studies reported as well that use privately created datasets; however, the prediction models applied to private datasets cannot directly be compared due to unavailability of these datasets. Therefore, our main focus has been on publicly available datasets.

In terms of most used learning techniques, ANN is the most popular prediction model followed by ensemble-based methods [1, 2, 6, 40]. SVM and DTs are also reported to produce good results. A comparison showing the number of studies (reviewed during our research work) using individual and ensemble-based classifiers is illustrated in Fig. 1. The following text provides an overview of existing state-of-the-art learning techniques applied on PIDD. ANN and ensemble-based techniques, being the most popular and most related, are discussed in detail. Table 1 summarizes the existing research in terms of pre-processing techniques, the classification techniques and performance measures.

2.1 Disease prediction using artificial neural network

ANNs have widely been used for prediction of diseases [14, 17, 19, 46, 53]. One of the earlier works on diabetes prediction is reported in 2003 in which authors trained

² <http://archive.ics.uci.edu/ml/>.

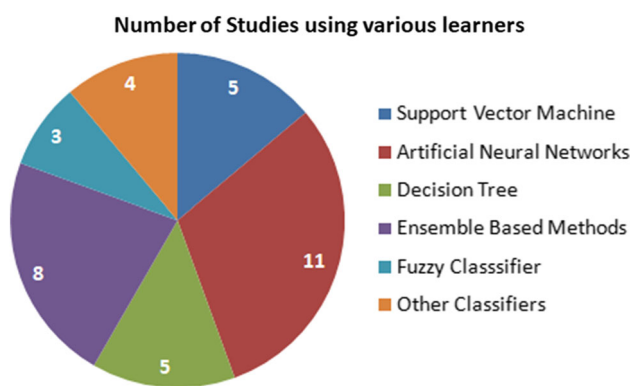


Fig. 1 Number of publications (reviewed during this work) related to diabetes prediction using various machine learning methods

different types of ANNs [31] on PIDD and performed comparative analysis among radial basis function (RBF), general regression neural network (GRNN) [51] and multilayer perceptron (MLP). GRNN outperformed other networks by achieving the highest accuracy of 80.2%. In 2005, [18] applied ANN on the PIDD and used 600 (78% approximately) randomly selected cases for training set and 168 (22% approximately) for test set. Two different experiments are carried out, one with 8 input variables and the other with 4 input variables. They reported highest performance using 8 inputs with 3 hidden layers. In 2006, [59] proposed a method of linguistic rule extraction from nodes of ANN and tested it on several UCI benchmark datasets including PIDD. The rules in this paper are extracted from neural network pruning using frequency interval data representation. They reported an accuracy of 74% on the PIDD. Applications of real-valued neural network (RVNN) and complex-valued neural network (CVNN) were reported by [47]. They experimented with several normalization techniques including min-max, z, complex along with unitary data normalization. They reported accuracies ranging between 80% and 81%, depending on the parameter combinations. The authors further extended their work in [3] proposing enhancements through application of complex-valued pseudo-autoregressive (CAR) technique, where adaptive coefficients are obtained from the trained network. They reported an accuracy of 81.28% on PIDD. In another study, [56] reported an accuracy of 82.37% using Levenberg–Marquardt (LM) [51] algorithm with probabilistic neural network on PIDD. Multilayer neural network is trained using LM algorithm.

ANN and its variations have been used in diabetes prediction using private datasets. Among earlier studies, [37] proposed an application of sequential MLP (SMLP) using a dataset collected from a US company. Stratified random sampling and random shuffling of inputs are used

as preprocessing steps to achieve a sensitivity of 86.04% and gain (average profit 0.18). In 2006, [38] performed experiments on Juvenile Diabetes Dataset for prediction and reported an accuracy of 99.72% using ANN. In the same year, [57] applied RBF on a private dataset and reported an accuracy of 97.0%, followed by sensitivity 97.3% and specificity 96.8%.

2.2 Disease prediction using ensemble-based learners

Ensemble classifiers have emerged as a popular technique during the last few years in the field of medical diagnostics. The ensemble-based classifier, as explained by [40], is the idea of using a combination of individual classifiers in order to get a classifier that performs better than any of the individual classifiers. In 2014, [35] improved the accuracy of diabetes prediction by proposing combination of individual classifiers on PIDD dataset. The following five classifiers were combined: sequential minimal optimization, RBF, C4.5, NB and RIPPER. The use of synthetic minority over-sampling technique as preprocessing step served the purpose. SMOTE helped increase the minority class. The highest accuracy of 77.9% was produced by C4.5, whereas lowest by RBF, i.e., 73.6%. They trained a metamodel and reported an accuracy of 77.0%. In the same year, [34] used weight-based voting approach during training. They reported an accuracy of 77.0% by using ANN, NB and SVM as an ensemble.

Among the studies on private datasets, [45] presented a combination of random forest and CART on a dataset collected from medical records of chronic disease of patients from Banjarnegara. They used a different number of trees and attribute selection and reported an accuracy of 83.8%. In 2015, [24] applied an ensemble-based learning using SVM and RBF. The model was trained on a data collected for China Health and Nutrition Survey. The dataset was first trained using SVM. The next step was to extract rules using RBF. Then, best extracted rules after tuning rule induction parameters were used to predict the class tuples from test data. Vacant data exclusion, feature selection and noise data canceling were used as preprocessing steps. Scores for precision, recall and f value calculated are 81.8%, 75.6% and 0.786, respectively.

2.3 Disease prediction using other learners

One of the earliest works reported in a disease prediction in 2002, [42], used critical SVM without kernel function to a number of benchmark datasets. The proposed algorithm is applied on PIDD as well where reported accuracy is 82.3% without any cross-validation on the PIDD dataset. In 2008, [39] presented the application of generalized discriminant analysis (GDA) [9] and least square SVM (LS-SVM) [54]

Table 1 Performance comparison of several techniques applied on PIDD

Year	Preprocessing technique	Prediction technique	Accuracy
<i>ANN-based techniques</i>			
2003 [31]	None	General regression neural network (GRNN)	80.21
2006 [59]	None	ANN	74
2009 [56]	None	Levenberg–Marquardt (LM) with probabilistic ANN	82.37
2010 [47]	Normalization, formatting of data	Complex-valued neural network (CVNN)	80–81
2011 [3]	None	CVNN [24]-based CAR model	81.28
<i>Ensemble-based techniques</i>			
2014 [35]	SMOTE	An ensemble of 5 classifiers	77
2014 [34]	Missing value imputation, wrapper method feature selection	Majority voting (SVM + ANN + NB)	77
<i>Other techniques</i>			
2002 [42]	None	SVM	82.29
2007 [58]	None	Ontology-based fuzzy inference agent system	74.2
2008 [23]	Feature identification and categorization, outlier removal and feature selection, data normalization, numerical discretization	ID3	80
2008 [39]	None	GDA + LS – SVM	79.16
2012 [20]	Normalization, discretization feature selection	NB network	72.3
2013 [13]	None	Neuro-fuzzy classifier	82.3
2013 [33]	None	Support vector machine with RBF	78
2014 [30]	None	Neuro-fuzzy inference system	80
2015 [49]	None	C4.5	81.3

to predict diabetes using PIDD. GDA is used for preprocessing, followed by LS-SVM for classification. They reported accuracy, sensitivity and specificity at 79.16%, 83.3% and 82.05%, respectively. In 2013, [33] applied SVM with RBF on PIDD and reported an accuracy of 78%.

Decision tree and its variants have also been extensively used in diabetes diagnosis. A maximum of 81% accuracy is reported in studies on PIDD. [23] proposed an application of various data preprocessing techniques combined with decision trees for classification to predict diabetes using PIDD. They reported maximum accuracy of 80% using ID3. In another study in 2011, [4] applied DT on PIDD. The dataset is trained using J48 algorithm and reported an accuracy of 78.2%. In a recent study, [49] achieved an accuracy of 81.3% by using C4.5 to extract rules.

3 Datasets

Pima Indians Diabetes Dataset (PIDD) is available on UCI³ machine learning repository. PIDD consists of data of 21 years or older females. Dr John Schorling from University of Virginia donated BioStat Diabetes Dataset (BDD). It contains the records of persons screened for diabetes. The

value of glycosylated hemoglobin > 7.0 is usually considered as a positive diagnosis of diabetes. Detailed statistics of datasets are listed in Table 2.

4 Proposed framework: Auto-MeDiSine

Auto-MeDiSine uses a novel automated version of a class outlier detection method as a major preprocessor, followed by an ensemble of AutoMLPs to create a prediction model. The process starts with splitting of dataset into training, validation and test set. The training dataset is processed initially using *Data Transformation* (conversion of nominal data into numeric). It is followed by the application of AutoECODB algorithm that removes noisy and unimportant incidences from datasets based on class outlier factor. We enhanced the ECODB by automating it to auto-tune itself to determine the best values of parameters involved in finding outliers. The outliers detected from the training set are discarded, thus giving a subset of original training set that is noise free. This dataset is then used for training the classifier for prediction. At prediction phase, an ensemble of AutoMLPs is used. AutoMLP trains on the dataset by performing auto-tuning of parameters and adjusting the size of MLPs, thus minimizing the human intervention in getting best prediction model. A complete block diagram

³ <http://archive.ics.uci.edu/ml/>.

Table 2 Description of Pima Indian Diabetes Datasets (PIDD) and BioStat dataset

Data set	Instances	Attributes	Prevalence of diabetes (%)	Features
PIDD	768	8	34.89	Concentration of plasma glucose, 2-h oral glucose tolerance test, diastolic BP, skin fold thickness of triceps (mm), 2-h serum insulin (mu U/ml), BMI, diabetes pedigree function, age (years), no. of pregnancies
BioStat	403	18	14.8	Stabilized glucose, total cholesterol, cholesterol/HDL ratio, glycosylated hemoglobin, location, age, gender, waist, hip, height, weight, frame, 1st SBP, 1st DBP, 2nd SBP, 2nd DBP, high-density lipoprotein, postprandial time when laboratories were drawn (min)

of the proposed framework is shown in Fig. 2. Each step involved in the framework is described in detail in the following subsections.

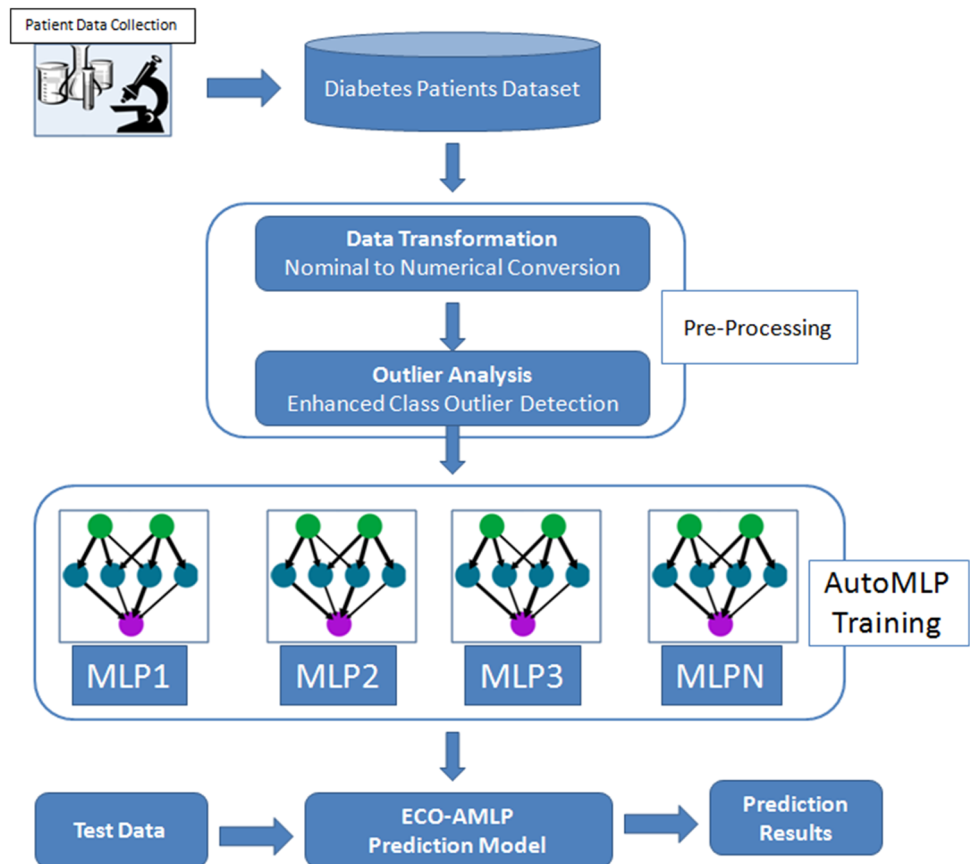
4.1 Data preprocessing

Data preprocessing involves steps to clean and improve the quality of data that can result in a better training of model. First step performed in the framework is the transformation of nominal attributes into numeric. Each record in the dataset contains the information about a patient and a class label. The class label of the record is “Yes” for sufferer of diabetes and “No” for healthy. These nominal values are mapped to 0 and 1, respectively. The dataset is then divided

into training, validation and test sets. The resultant dataset is then subjected to detection of outliers using the Auto-ECODB algorithm (described below).

Outliers are data instances that deviate in behavior from other records in datasets. They can be defined as exceptions or rare cases. Conventional techniques detect outliers based on the whole dataset. Such techniques do not consider the class label while detecting outliers. On the other hand, class-based outlier mining techniques detect outliers in the dataset with respect to class label. One such technique is ECODB that detects outliers based on enhanced class outlier factor (ECOF) that ranks the data records for their degree of being a class outlier. ECOF considers the following factors to rank any record as an outlier:

Fig. 2 Auto-MeDiSine framework for prediction of diabetes from patients dataset



- probability of class label of the record S compared to its K nearest neighbors
- deviation of the record S from records of the same class
- distance between the record S and its K nearest neighbors

All records are ranked for ECOF among which the N records with highest rank are eliminated. A given record is labeled as class outlier that produces the least K -distance from its K nearest neighbors, its deviation from the respective records of the same class is the greatest, and it has different class label of its K nearest neighbors' class. The mathematical expressions of ECOF [25] for any record (S) are given in Eq. 1.

$$\text{ECOF}(S) = K \times \text{PCL}(S, K) - \text{norm}(\text{Deviation}(S)) + \text{norm}(\text{KDist}(S)) \quad (1)$$

where $\text{PCL}(S, K)$ is probability of the class label of record (S) among class labels of its K nearest neighbors. $\text{Deviation}(S)$ is deviation of the record (S) from records of the same class. It is calculated as sum of the distances between the record (S) and other records. $\text{KDist}(S)$ is the sum of distances between record (S) and its K nearest neighbors. ECOF is applied on the normalized values of $\text{Deviation}(S)$ and $\text{KDist}(S)$, and their range of values is [0–1].

$$\text{norm}(\text{Deviation}(S)) = \frac{\text{Deviation}(S) - \text{MinDev}}{\text{MaxDev} - \text{MinDev}} \quad (2)$$

$$\text{norm}(\text{KDist}(S)) = \frac{\text{KDist}(S) - \text{MinKDist}}{\text{MaxKDist} - \text{MinKDist}} \quad (3)$$

Here, MaxDev and MinDev represent the highest and lowest deviation value for top N outlier instances. MinKDist and MaxKDist are the lowest and highest KDist value for top N outlier instances. Calculation of top N outlier instances and working of ECODB algorithm are described [44] here:

1. Compute $\text{PCL}(S, K)$ for all records in given dataset.
2. Maintain a list of top N instances with least $\text{PCL}(S, K)$ value.
3. Compute $\text{KDist}(S)$ and $\text{Deviation}(S)$ for each record in the list of top N records.
4. Using the values in point 3, maintain MaxKDist , MinKDist , MaxDev and MinDev values.
5. Compute ECOF value for all instances in the top N list according to Eq. 1.
6. Resort the top N list in ascending order according to their ECOF value.

We designed a wrapper, named as AutoECODB, that performs automatic optimization of parameters that are involved in the implementation of ECODB. AutoECODB optimizes the following parameters:

- The number of neighbors (K) to be considered to calculate probability
- The number of top class outliers (N) to be eliminated from dataset
- The measure types (numerical, mixed, nominal)
- The numerical measures (Euclidean distance, cosine-based similarity, etc.).

The algorithm chooses the best values of these parameters in order to maximize the performance of overall system.

4.2 Training the prediction model

Auto-MeDiSine builds upon the strengths of AutoMLP and ensemble methods. The topology of network while designing ANNs is of utmost significance. ANN, like a human brain, comprises a network of interconnected neurons where each connection has an associated weight with it. These weights are adjusted based on learning experience of algorithm. The network topology for ANNs has to be adjusted before training the algorithm that includes the number of hidden layers and hidden units in them, learning rate (training parameter that controls the size of weight and bias changes) and number of epochs (number of iterations over training set). Parameter optimization is an old problem of ANNs [43] which requires human intervention to choose the best suitable parameters for the network. However, AutoMLP works on a mechanism to optimize the parameters involved in structure of ANN. Working of AutoMLP is briefly described here. The experimental setup is shown in Fig. 2.

AutoMLP introduced in [12] is a type of multilayered feedforward neural network which is auto-tunable, i.e., it adjusts the learning rate and number of hidden units is automated. AutoMLP combines ideas from genetic algorithm and stochastic optimization. It trains a small ensemble of MLP networks in parallel using different numbers of hidden units and learning rate. It optimizes using gradient-based optimization techniques. The error rate is determined on a validation set after a small fixed number of epochs followed by replacing worst performer networks with best ones. This way the networks have different numbers of hidden units and learning rates. Learning rates and hidden unit numbers are drawn according to probability distributions derived from successful rates and sizes.

5 Experimental setup and results

In order to measure the performance of the proposed framework, a series of experiments are carried out on PIDD. Experiments are performed using various combinations of preprocessing methods and classifiers as follows:

1. AutoMLP with varying preprocessing techniques

2. AutoECODB with varying classifiers
3. Varying classifiers with varying preprocessing techniques.

Our search for the best predictive model is based on the hypothesis that the performance of ANN-based methods can be improved as it is largely dependent on their structure and parameters. The proposed method, i.e., Auto-MeDiSine, should be able to perform better as it involves the auto-tuning of structure and parameters to their best combination. The results of experiments demonstrated that Auto-MeDiSine (AutoECODB as preprocessing technique, followed by AutoMLP as classifier) produced the best results on a given dataset. The complete configuration of best performing combination is provided in Sect. 5.2.

In all experiments, PIDD is divided into three sets: training, validation and test with 70%, 15% and 15% records in each set, respectively. The performance is evaluated on the test sets having the same parameters as those tuned on validation set. The testing data remains unseen throughout training and preprocessing process. The number of records in training, validation and test sets is 528, 115 and 115, respectively, for PIDD. As shown in Fig. 2, the original dataset is first transformed. Data transformation is performed to transform the nominal attribute to numerical. After transformation, the training set is subjected to preprocessing methods, after which the training of learner is performed. The records comprising attributes such as plasma glucose concentration, BMI and diabetes pedigree function are fed as input to the respective classifier. The trained model is then applied on test data to measure the performance metrics.

5.1 Performance metrics

Performance metrics used during training determine the performance of classifier. We made use of the following metrics: accuracy, precision and recall. The metrics are briefly described below.

- True positives (TP) represent the number of actual diabetic patients correctly predicted.
- True negatives (TN) represent the non-diabetic patients predicted as non-diabetic.
- False positives (FP) represent the non-diabetic patients predicted diabetic.
- False negatives (FN) represent the actual diabetic patients predicted as non-diabetic.
- Positives (P) represent all actual diabetic patients.
- Negatives (N) represent all actual non-diabetic patients.

Using the above-stated variables, the evaluation metrics can be defined as follows:

Accuracy is the percentage of patients that are correctly diagnosed by classifier (diabetic or non-diabetic).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}.$$

Precision/specificity represents the correctness of diabetic diagnosis, i.e., percentage of patients labeled as diabetic are actually diabetic (exactness)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

Recall/Sensitivity represents the completeness of coverage, i.e., percentage of actual diabetic patients correctly diagnosed by our framework

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

The average precision calculated per class is weighted mean precision (WMP)

The average recall calculated per class is weighted mean recall (WMR).

5.2 The configuration of proposed method: Auto-MeDiSine

The training set, after transformation, is subjected to AutoECODB to remove the outliers. AutoECODB tunes the parameters by systematically optimizing the parameters using error reduction with gradient descent. The best performance is achieved with top 10 outliers from the training set using correlation similarity considering the 12 nearest neighbors. In the next step, the records comprising attributes such as plasma glucose concentration, BMI and diabetes pedigree function are fed as input and constitute the input layer. The numbers of attributes for PIDD are 7. These inputs are weighted and then passed from input layer to ensembles of AutoMLP that apply nonlinear activation function to the weighted input. The training parameters are as follows:

- Training cycles: The training cycles used during training the neural network.
- Number of generations: The number of generations for training.
- Number of ensemble MLPs: The number of MLPs per ensemble.

Experiments are performed by varying the number of these three parameters. The best results were obtained using 4 MLPs and 10 generations. After 10 training cycles, worst MLPs are replaced with the best ones. The proposed network topology consisted of one hidden layer with 160 nodes in them for the best MLP selected after training process. Weights were adjusted using sigmoid activation

function. The performance of the proposed framework is evaluated on validation and test set after preprocessing and training stages.

5.3 Experiments using AutoMLP with varying preprocessing techniques

In this set of experiments, the classifier (i.e., ensemble of AutoMLP) is kept fixed in terms of parameters and structure, while preprocessing techniques are varied to perform a comparison among them. The results demonstrate that the AutoECODB with ensemble of AutoMLPs performs better than other methods. The accuracy, WMR and WMP by all methods are presented in Table 3. Different feature selection, attribute weight generation, normalization and sampling techniques are compared with AutoECODB. We have particularly focused on other outlier-based methods, including simple distance-based outliers and outlier detection using principal component analysis (PCA). The performance of AutoECODB with different numbers of nearest neighbors and outliers is illustrated in Fig. 3. The graph shows that the best performance is achieved using 12 nearest neighbors and 10 outliers.

5.4 Experiments using AutoECODB with varying classifiers

In this set of experiments, the preprocessing technique is kept fixed as AutoECODB and classifiers are varied. The results

shown in Table 4 demonstrate that the proposed combination, i.e., Auto-MeDiSine, performs better than other methods. As compared to accuracy at 88.70% of the proposed method, the highest accuracy achieved using other methods is 81.74% with SVM. The lowest accuracy is 74.78% using KNN. Results of other classifiers as reported in the literature are also improved in this research as AutoECODB proved to be a better option for preprocessing. For example, in [33], accuracy of SVM is 78%. Similarly, the literature reports an accuracy of 78.17% [4] using DTs, while using AutoECODB with DT produced an accuracy of 79.13%. Comparison is also performed between other flavors of ANN and ensemble of AutoMLP. Table 4 shows that the performance of Auto-MeDiSine is better as compared to other classifiers. Results show that performance of AutoMLP is much better than other neural network-based classifiers. The reason is that the performance of ANNs is highly dependent on the parameters and structure of network and AutoMLP is able to tune itself to better structure in terms of parameters.

5.5 Experiments using varying preprocessing techniques with varying classifiers

This section presents the results of using different combinations of preprocessing techniques and classifiers. The results shown in Table 5 demonstrate that the proposed combination, i.e., Auto-MeDiSine, performs better than other methods. Experiments are performed with a limited

Table 3 Comparison of AutoECODB with other preprocessing techniques used on PIDD

Preprocessing methods	Accuracy	WMR	WMP
<i>Feature selection</i>			
Principal component analysis [60]	65.04	62.59	62.77
Fast correlation-based filter (FCBF) [62]	82.61	74.00	82.31
Select by recursive feature elimination with SVM [21]	82.61	74.00	82.31
Select by feature quantile filter [26]	82.61	78.27	79.31
<i>Attribute weight generation</i>			
Weight by maximum relevance [16]	82.61	74.00	82.31
Weight by correlation-based weak association [22]	82.61	74.00	82.31
<i>Normalization</i>			
Normalization (Z-transform) [5]	29.57	50.00	14.78
<i>Sampling</i>			
Bootstrap sampling [29]	81.74	75.09	79.08
Stratified sampling [27]	81.74	75.09	79.08
<i>Outlier methods</i>			
Simple outliers using distance [10]	81.74	75.09	79.08
Stratified outlier using distance [10]	82.4	77.1	81.1
Outlier using PCA [41]	81.74	75.09	79.08
Auto-MeDiSine	88.7	88.56	85.83

Accuracy, WMP and WMR are recorded in percentage
 Bold values show the highest reported/measured results

Fig. 3 Performance comparison of ECODB with different numbers of nearest neighbors and outliers

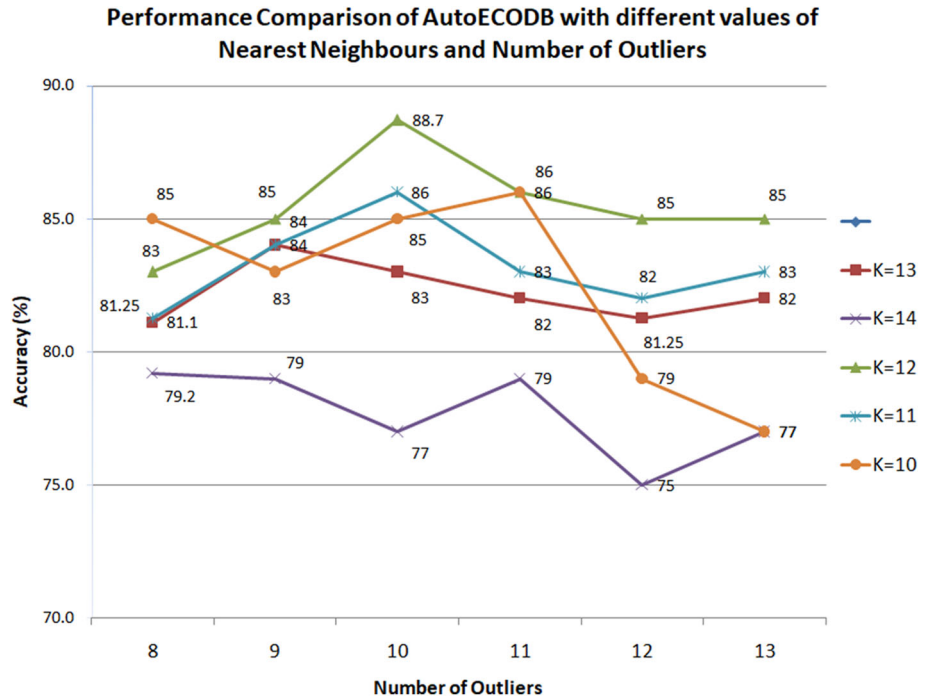


Table 4 Comparison of Auto-MeDiSine with state-of-the-art classification techniques on PIDD

Classification technique	Accuracy	WMR	WMP
KNN	74.78	71.86	70.31
Naïve Bayes (NB)	80.87	77.03	77.03
Decision tree (DT)	79.13	68.12	78.58
Rule induction (RI)	79.13	71.53	75.63
Linear regression (LR)	82.61	73.15	83.55
SVM	81.74	75.53	81.43
Bagging	78.26	70.92	74.24
AdaBoost	79.13	74.95	74.95
Boosting	79.13	74.95	74.95
Stacking	73.91	68.68	68.68
<i>Different architectures of artificial neural network</i>			
Artificial neural net (ANN)	78.26	71.77	74.04
Perceptron (P'tron)	70.43	52.56	60.78
Multilayer perceptron (MLP)	81.74	75.94	78.65
Voted perceptron (V P'tron)	72.17	58.91	65.50
RBF network	80.87	74.47	77.67
Proposed technique	88.70	88.56	85.83
Auto-MeDiSine			

Accuracy, weighted mean precision and weighted mean recall are presented in percentage (results with these classifiers are measured by authors)

Bold values show the highest reported/measured results

number of preprocessing techniques and classifiers as the possible combinations are very high. ANN-based and ensemble-based classifiers are proved to be the best performing.

5.6 Comparison of Auto-MeDiSine with state-of-the-art results

A comparison was performed between the proposed Auto-MeDiSine and existing best reported methods of diabetes prediction on PIDD as shown in Table 6. Our experimentation and literature reported ANN to produce the highest accuracies ranging from 81 to 82%. Table 6 details the performance of evaluations along with the preprocessing techniques and prediction technique. The proposed technique outperformed other techniques presented in the literature as clearly evident from the results reported.

Furthermore, in order to validate the generality of Auto-MeDiSine, experiments are performed on BioStat dataset as well, in a similar manner as described for PIDD. Results, as shown in Fig. 4, demonstrate the generality of Auto-MeDiSine as it is capable of producing good results on different datasets. The performance of AdaBoost is measured to be closer to that of Auto-MeDiSine.

6 Conclusions

We present a novel framework Auto-MeDiSine to predict diabetes, performing experiments on public dataset of patients' named as Pima Indian Diabetes Dataset (PIDD). The paper summarizes the reported studies on PIDD and other private datasets and presents a number of experiments performed using Auto-MeDiSine to show that the proposed technique provides promising results. Instead of relying on complex feature selection or extraction tasks,

Table 5 Performance of different classifiers combined with different preprocessing methods

Classifier	Preprocessing	Acc	Classifier	Preprocessing	Acc
KNN	PCA	69.57	Bagging	PCA	74.91
	FCBF	73.91		FCBF	79.87
	FE with SVM	73.91		FE with SVM	79.61
	Weight by MR	73.91		Weight by MR	80.87
	Bootstrap sampling	69.57		Bootstrap sampling	80.61
SVM	PCA	79.13	Stacking	PCA	76.52
	FCBF	80.87		FCBF	80
	FE with SVM	80.87		FE with SVM	80.87
	Weight by MR	80.87		Weight by MR	80.00
	Bootstrap sampling	80.87		Bootstrap sampling	78.26
DT	PCA	76.52	MLP	PCA	73.91
	FCBF	79		FCBF	80
	FE with SVM	77.2		FE with SVM	80.87
	Weight by MR	78.2		Weight by MR	80.00
	Bootstrap sampling	78.3		Bootstrap sampling	82.3

Accuracy is recorded in percentage

Bold values show the highest reported/measured results

Table 6 Comparison of Auto-MeDiSine with state-of-the-art techniques on PIDD

Year	Preprocessing methods	Prediction methods	Accuracy
2010 [47]	Data formatting/normalization	CVNN (complex-valued neural network)	81
2011 [3]	None	CVNN-based CAR model	81.28
2012 [20]	Normalization, discretization and feature selection	Naïve Bayes network	72.3
2013 [13]	None	Neuro-fuzzy classifier	82.32
2014 [35]	SMOTE	Metamodel of 5 classifiers	77.0
2014 [30]	None	Neuro-fuzzy inference system	80
2015 [49]	None	C4.5	81.27
2017	Proposed: Auto-MeDiSine		88.70

Accuracy is recorded in percentage

Bold values show the highest reported/measured results

Auto-MeDiSine makes use of an auto-tunable outlier detection-based technique (AutoECODB) as a preprocessing step to detect and remove outliers in the dataset. The experiments show that the AutoECODB performs better as compared to other normalization, attribute weight generation and feature selection techniques.

Our detailed literature review indicated that ANNs achieve the best results in terms of accuracy. One of the major issues with ANN architectures is around the optimization of parameters, such as the number and composition of hidden layers, the learning rate, epochs and other

aspects of network topology. These parameters are to be decided before training the ANN. AutoMLP, an auto-tunable ensemble of multilayer perceptrons (MLPs), addresses this issue by enabling automated optimization of the above parameters. Auto-MeDiSine is compared with other state-of-the-art methods reported in the literature where Auto-MeDiSine realized an accuracy of 88.7%. Furthermore, in order to validate the generality of framework, Auto-MeDiSine is tested on another publicly available BioStat Diabetes Dataset where it outperforms the existing reported results, realizing the accuracy of 97.1.

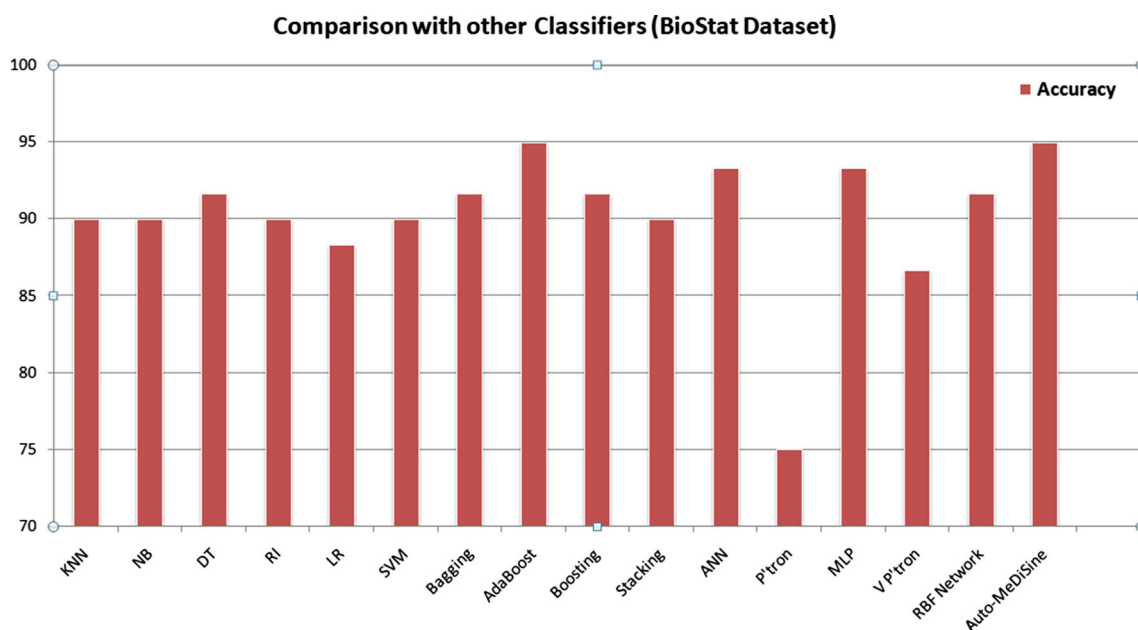


Fig. 4 Comparison of Auto-MeDiSine with other classification techniques on BioStat dataset

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Ahmed M, Afzal H, Siddiqi I, Khan B (2017) Mcs: multiple classifier system to predict the churners in the telecom industry. In: SAI Intelligent Systems Conference 2017, London, UK
- Ahmed M, Rasool AG, Afzal H, Siddiqi I (2017) Improving handwriting based gender classification using ensemble classifiers. *Expert Syst Appl* 85(1):158–168
- Aibinu AM, Salami MJE, Shafie AA (2011) A novel signal diagnosis technique using pseudo complex-valued autoregressive technique. *Expert Syst Appl* 38(8):9063–9069
- Al Jarullah AA (2011) Decision tree discovery for the diagnosis of type ii diabetes. In: 2011 International conference on innovations in information technology (IIT). IEEE, pp 303–307
- Al Shalabi L, Shaaban Z (2006) Normalization as a preprocessing engine for data mining and the approach of preference matrix. In: International conference on dependability of computer systems, 2006. DepCos-RELCOMEX'06. IEEE, pp 207–214
- Ali R, Siddiqi MH, Idris M, Kang BH, Lee S (2014) Prediction of diabetes mellitus based on boosting ensemble modeling. In: International conference on ubiquitous computing and ambient intelligence. Springer, pp 25–28
- Anbarasi M, Anupriya E, Iyengar N (2010) Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *Int J Eng Sci Technol* 2(10):5370–5376
- Apolloni B, Avanzini G, Cesa-Bianchi N, Ronchini G (1990) Diagnosis of epilepsy via backpropagation. In: Proceedings of the 1990 international joint conference on neural networks, vol 2, pp 571–574
- Baudat G, Anouar F (2000) Generalized discriminant analysis using a kernel approach. *Neural Comput* 12(10):2385–2404
- Bay SD, Schwabacher M (2003) Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 29–38
- Bounds DG, Lloyd PJ, Mathew B, Waddell G (1988) A multi-layer perceptron network for the diagnosis of low back pain. In: IEEE international conference on neural networks 1988. IEEE, pp 481–489
- Breuel T, Shafait F (2010) Automlpl: simple, effective, fully automated learning rate and size adjustment. In: The learning workshop. Utah
- Daho MEH, Settouti N, Lazouni MEA, Chikh MA (2013) Recognition of diabetes disease using a new hybrid learning algorithm for nefclass. In: 2013 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA). IEEE, pp 239–243
- DeGroff CG, Bhatikar S, Hertzberg J, Shandas R, Valdes-Cruz L, Mahajan RL (2001) Artificial neural network-based method of screening heart murmurs in children. *Circulation* 103(22):2711–2716
- Delen D, Walker G, Kadam A (2005) Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med* 34(2):113–127
- Ding C, Peng H (2005) Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol* 3(02):185–205
- Dreiseitl S, Ohno-Machado L (2002) Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform* 35(5):352–359
- Farhanah S, Jafan B, Ali DM (2005) Diabetes mellitus forecast using artificial neural networks (ann). In: Asian conference on sensors and the international conference on new techniques in pharmaceutical and medical research proceedings (IEEE), pp 135–138
- Floyd CE, Lo JY, Yun AJ, Sullivan DC, Kornguth PJ (1994) Prediction of breast cancer malignancy using an artificial neural network. *Cancer* 74(11):2944–2948

20. Guo Y, Bai G, Hu Y (2012) Using bayes network for prediction of type-2 diabetes. In: 2012 international conference for internet technology and secured transactions. IEEE, pp 471–472
21. Gysels E, Reneveu P, Celka P (2005) Svm-based recursive feature elimination to compare phase synchronization computed from broadband and narrowband eeg signals in brain-computer interfaces. *Signal Process* 85(11):2178–2189
22. Hall MA (2000) Correlation-based feature selection of discrete and numeric class machine learning (Working paper 00/08). University of Waikato, Hamilton, New Zealand
23. Han J, Rodriguez JC, Beheshti M (2008) Diabetes data analysis and prediction model discovery using rapidminer. In: 2008 Second international conference on future generation communication and networking, vol 3. IEEE, pp 96–99
24. Han L, Luo S, Yu J, Pan L, Chen S (2015) Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. *IEEE J Biomed Health Inform* 19(2):728–734
25. Hewahi NM, Saad MK (2007) Class outliers mining: distance-based approach. *Int J Intell Technol* 2(1):55–68
26. Hilger F, Molau S, Ney H et al (2002) Quantile based histogram equalization for online applications. In: INTERSPEECH
27. Imbens GW, Lancaster T (1996) Efficient estimation and stratified sampling. *J Econom* 74(2):289–318
28. Jahangir M, Afzal H, Ahmed M, Khurshid K, Nawaz R (2017) An expert system for diabetes prediction using auto tuned multi-layer perceptron. In: Intelligent systems conference (IntelliSys) 2017. IEEE, pp 722–728
29. Johns MV (1988) Importance sampling for bootstrap confidence intervals. *J Am Stat Assoc* 83(403):709–714
30. Kalaiselvi C, Nasira G (2014) A new approach for diagnosis of diabetes and prediction of cancer using anfis. In: 2014 World congress on computing and communication technologies (WCCCT). IEEE, pp 188–190
31. Kayaer K, Yıldırım T (2003) Medical diagnosis on pima indian diabetes using general regression neural networks. In: Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP), pp 181–184
32. Kharya S (2012) Using data mining techniques for diagnosis and prognosis of cancer disease. arXiv preprint [arXiv:1205.1923](https://arxiv.org/abs/1205.1923)
33. Kumari VA, Chitra R (2013) Classification of diabetes disease using support vector machine. *Int J Eng Res Appl* 3(2):1797–1801
34. Li L (2014) Diagnosis of diabetes using a weight-adjusted voting approach. In: 2014 IEEE international conference on bioinformatics and bioengineering (BIBE). IEEE, pp 320–324
35. Nnamoko NA, Arshad FN, England D, Vora J (2014) Meta-classification model for diabetes onset forecast: a proof of concept. In: 2014 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, pp 50–56
36. Ohno-Machado L, Musen MA (1997) Sequential versus standard neural networks for pattern recognition: an example using the domain of coronary heart disease. *Comput Biol Med* 27(4):267–281
37. Park J, Edington DW (2001) A sequential neural network model for diabetes prediction. *Artif Intell Med* 23(3):277–293
38. PObi S, Hall LO (2006) Predicting juvenile diabetes from clinical test results. In: The 2006 IEEE international joint conference on neural network proceedings. IEEE, pp 2159–2165
39. Polat K, Güneş S, Arslan A (2008) A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine. *Expert Syst Appl* 34(1):482–487
40. Polikar R (2006) Ensemble based systems in decision making. *IEEE Circuits Syst Mag* 6(3):21–45
41. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904–909
42. Raicharoen T, Lursinsap C (2002) Critical support vector machine without kernel function. In: Proceedings of the 9th international conference on neural information processing, 2002. ICONIP'02, vol 5. IEEE, pp 2532–2536
43. Rashid SF, Shafait F, Breuel TM (2012) Scanning neural network for text line recognition. In: 2012 10th IAPR international workshop on document analysis systems (DAS). IEEE, pp 105–109
44. Saad MK, Hewahi NM (2009) A comparative study of outlier mining and class outlier mining. *Comput Sci Lett* 1(1)
45. Sabariah MMK, Hanifa SA, Sa'adah MS (2014) Early detection of type ii diabetes mellitus with random forest and classification and regression tree (cart). In: 2014 International conference of advanced informatics: concept, theory and application (ICAICTA). IEEE, pp 238–242
46. Saha S, Raghava G (2006) Prediction of continuous b-cell epitopes in an antigen using recurrent neural network. *Proteins Struct Funct Bioinform* 65(1):40–48
47. Salami M, Shafie A, Aibinu A (2010) Application of modeling techniques to diabetes diagnosis. In: IEEE EMBS conference on biomedical engineering & sciences
48. Sathyadevi G (2011) Application of cart algorithm in hepatitis disease diagnosis. In: 2011 International conference on recent trends in information technology (ICRTIT). IEEE, pp 1283–1287
49. Saxena K, Sharma R et al (2015) Diabetes mellitus prediction system evaluation using c4. 5 rules and partial tree. In: 2015 4th international conference on reliability, infocom technologies and optimization (ICRITO) (trends and future directions). IEEE, pp 1–6
50. Shanker MS (1996) Using neural networks to predict the onset of diabetes mellitus. *J Chem Inf Comput Sci* 36(1):35–41
51. Specht DF (1991) A general regression neural network. *IEEE Trans Neural Netw* 2(6):568–576
52. Srinivas K, Rani BK, Govrdhan A (2010) Applications of data mining techniques in healthcare and prediction of heart attacks. *Int J Comput Sci Eng (IJCSE)* 2(02):250–255
53. Sumathy M, Thirugnanam M, Kumar P, Jishnujit T, Kumar KR (2010) Diagnosis of diabetes mellitus based on risk factors. *Int J Comput Appl* 10(4):1–4
54. Suykens JA, Vandewalle J (1999) Least squares support vector machine classifiers. *Neural Process Lett* 9(3):293–300
55. Tafa Z, Pervetica N, Karahoda B (2015) An intelligent system for diabetes prediction. In: 2015 4th Mediterranean conference on embedded computing (MECO). IEEE, pp 378–382
56. Temurtas H, Yumusak N, Temurtas F (2009) A comparative study on diabetes disease diagnosis using neural networks. *Expert Syst Appl* 36(4):8610–8615
57. Venkatesan P, Anitha S (2006) Application of a radial basis function neural network for diagnosis of diabetes mellitus. *Curr Sci* 91(9):1195–1199
58. Wang MH, Lee CS, Li HC, Ko WM (2007) Ontology-based fuzzy inference agent for diabetes classification. In: NAFIPS 2007–2007 annual meeting of the north American fuzzy information processing society. IEEE, pp 79–83
59. Wettayaprasit W, Sangket U (2006) Linguistic knowledge extraction from neural networks using maximum weight and frequency data representation. In: 2006 IEEE conference on cybernetics and intelligent systems. IEEE, pp 1–6
60. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. *Chemom Intell Lab Syst* 2(1–3):37–52

-
61. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, Hua L (2012) Data mining in healthcare and biomedicine: a survey of the literature. *J Med Syst* 36(4):2431–2448
 62. Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of the 20th

international conference on machine learning (ICML-03), pp 856–863

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.