Multimedia Systems (2022) 28:1289–1307
https://doi.org/10.1007/s00530-021-00817-2

SPECIAL ISSUE PAPER

Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus

Chandrashekhar Azad[1] · Bharat Bhushan[2] · Rohit Sharma[3] · Achyut Shankar[4] · Krishna Kant Singh[5] · Aditya Khamparia[6]

Received: 5 September 2020 / Accepted: 29 May 2021 / Published online: 6 June 2021

Abstract

Diabetes mellitus is a well-known chronic disease that diminishes the insulin producing capability of the human body. This results in high blood sugar level which might lead to various complications such as eye damage, nerve damage, cardiovascular damage, kidney damage and stroke. Although diabetes has attracted huge research attention, the overall performance of such medical disease classification using machine learning techniques is relatively low, majorly due to existence of class imbalance and missing values in the data. In this paper, we propose a novel *P*rediction **M**odel using **S**ynthetic Minority Oversampling Technique, **G**enetic Algorithm and **D**ecision Tree (PMSGD) for Classification of Diabetes Mellitus on Pima Indians Diabetes Database (PIDD) dataset. The framework of the proposed PMSGD prediction model is composed of four different layers. The first layer is the pre-processing layer which is responsible for handling missing values, detection of outlier and oversampling the minority class. In the second layer, the most significant features are selected using correlation and genetic algorithm. In the third layer, the proposed model is trained, and its effectiveness is evaluated in the fourth layer in terms of classification accuracy (CA), classification error (CE), precision, recall (sensitivity), measure (FM), and Area_Under_ROC (AUROC). The proposed PMSGD algorithm clearly outperforms its counterparts and achieves a remarkable accuracy of 82.1256%. The best outcome achieved by the proposed system in terms of CA, CE, precision, sensitivity, FM and AUROC is 82.1256%, 17.8744%, 0.8070%, 0.8598, 0.8326 and 0.8511, respectively. The obtained simulation results show the effectiveness and superiority of our proposed PMSGD model and their by reduced error rate to help in decision-making process.

Keywords Decision tree · Genetic algorithm · SMOTE · Data classification · Healthcare · Machine learning

✉ Rohit Sharma
rohitapece@gmail.com

Chandrashekhar Azad
csazad.ca@nitjsr.ac.in

Bharat Bhushan
bharat_bhushan1989@yahoo.com

Achyut Shankar
ashankar2711@gmail.com

Krishna Kant Singh
krishnaiitr2011@gmail.com

Aditya Khamparia
aditya.khamparia88@gmail.com

[1] Department of Computer Applications, National Institute of Technology, Jamshedpur, India

[2] Department of Computer Science and Engineering, School of Engineering and Technology, Sharda University, Greater Noida, India

[3] Department of Electronics & Communication Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, NCR Campus, Delhi- NCR Campus, Ghaziabad, India

[4] Department of CSE, ASET, Amity University, Noida 201301, India

[5] Department of Computer Science & Engineering, Jain (Deemed-to-be University), Bengaluru, India

[6] School of Computer Science and Engineering, Lovely Professional University, Punjab, India

1 Introduction

Pancreas is an essential organ of human body, majorly because it produces insulin that helps in the metabolism of protein, fat and sugar for daily life energy. Insulin deficiency results in increased blood sugar concentration and drives out the redundant sugar via urine. This results in a disease called 'diabetes mellitus' that has symptoms like increased hunger, increased thirst, hypertension, frequent urination, stroke, high blood sugar, dyslipidaemia, cardiovascular damage and kidney damage [1, 2]. Lack of exercise and obesity is the premiere cause of diabetes as it depends on weight–height ration, diet style, and hereditary factors. Diabetes is the most serious long-term illness situation that has globally impacted lots of people in both developing as well as developed countries. World Health Organization (WHO) reported diabetes as the highest contributing non-communal disease (NCD) deaths across the globe [3]. According to a report, 20 million people including children and adults suffered from diabetes in USA during 2007 [4]. Another report suggested that, by 2030, more than four-fifth of the diabetic patients across the globe will be from developing countries [5].

The huge amount of data including treatment data, electronic medical records, and patient diagnosis information are generated in healthcare industry. This can be used to extract knowledge that mitigates cost and supports efficient decision-making. The advancements in the medical field have made significant strides in the development of antibiotics, vaccinations and sterilization that enabled industrial disruption and caused a cascading effect on the associated doctors as well as patients. Owing to the recent advancements in intelligent analysis methods [7], employing intelligence for medical diagnosis has emerged as an unrivalled hot issue [7]. To this end, machine learning (ML) algorithms have gained much significance due to its strength of managing voluminous data and making efficient predictions in computationally intensive manner [8, 9]. ML can serve as a solution for mitigating the cost involved in healthcare management and also enable the establishment of better doctor–patient relationship. However, numerous clinical issues exist such as requirement of quick, reliable and accurate decision models. This issue needs to be addressed for accurate disease diagnosis. The existence of huge amount of unstructured data in healthcare makes it difficult to categorize and quantify a conversation between a provider and the patient. Furthermore, performance of the classification models degrades, as the majority of medical datasets contain incomplete, redundant, irrelevant and noisy information [10]. Around one of every seven US grown-ups currently suffer from diabetes, as indicated by the Centers for Disease Control and Prevention. If that remains the case, it is estimated that by 2050, one of every three individuals will suffer from diabetes. In this regard, we utilize machine learning to assist us in early prediction of diabetes. This work presents a prediction system for diabetes disease that also addresses the problems of data imbalance and curse of dimensionality in the diabetes datasets. The significance of information quality (particularly in clinical information), has driven towards an ever-expanding development in information pre-processing strategies. The recent studies fail to identify the right approach that solves the issues of efficiency as well as the ease of implementation [11–15].

In this paper, we proposed a novel *P*rediction **M**odel using **S**ynthetic Minority Oversampling Technique, **G**enetic Algorithm and **D**ecision Tree (PMSGD) for Classification of Diabetes Mellitus on Pima Indians Diabetes Database (PIDD) dataset. The proposed PMSGD model is comprised of four different layers. The first layer is the pre-processing layer that is responsible for missing values treatment, outlier detection and its handling [16, 17]. The class imbalance problem is solved by oversampling the minority class using synthetic minority oversampling technique (SMOTE) that yields high-quality training datasets [18, 19]. The second layer is responsible for feature selection that eliminates the insignificant features to generate high-quality datasets using correlation and genetic algorithm (GA). This reduced dimension of the dataset lowers the training complexity and also solves the issues of over fitting. The third layer relies on decision tree (DT) for predicting the diabetic patients' records [20]. The fourth layer is the performance evaluation layer in which the implementation of our prediction model on the Pima Indian Diabetes (PID) dataset yields adequate confirmation that the proposed prediction model outperforms the existing models in terms of various performance metrics including classification accuracy (CA), classification error (CE), precision, recall (sensitivity), F_Measure (FM), and area under receiver operating characteristic (AUROC).

The major contribution of this proposed approach is as follows:

1. Pre-processing of the dataset is performed for the following: (a) checking and handling missing values, (b) outliers' detection and handling, and (c) production of high-quality training datasets by oversampling the minority class (solves the class imbalance problem). As most of the existing artificial intelligence approaches neglect the minority class, these are prone to inconsistent results. This is the major issues in dealing with the imbalanced data sets.

2. Feature selection is employed to remove the insignificant features using correlation and GA from the PID dataset to reproduce the high-quality dataset. Owing to this reduction in the dimension of the dataset, the training complexity is reduced thereby resolving the issues of over fitting.

The remainder of the paper is organized as follows. Section 2 presents the related work. Section 3 presents the detailed description of the materials and methods employed. It explores the operations involved in GA, DT and SMOTE along with outlining the framework of the proposed prediction model. Section 4 presents the experimental discussion and analysis. It provides the statistical description of dataset, visualization of attribute values and relative performance measures. Finally, the paper concludes itself in Sect. 5 highlighting few open research trends in the related field.

2 Related work

Typically, several ML techniques are employed to capitulate diagnostic or prognostic models by learning from a sample of observed cases for diagnosis of diabetes or prediction of new diabetic cases. Such models can sometimes outperform the expert predictions and can serve as an appropriate model to guide physicians' decisions. Further, the dataset used in this work have been utilized in huge number of studies that had approached the desired task differently and achieved varied results. Some of the most influential works in this field are reviewed in the section below.

Barakat et al. [21] used support vector machine (SVM) for the diagnosis of diabetes and incorporated an interpretation module that converts the SVM's black box model into an intelligible SVM representation. The purpose of these rules extracted from it is to work as a second opinion for the diagnosis of diabetes and as a tool for predicting diabetes by identifying high-risk people. The significance of the proposed model lies in its simplicity, understandability, and validity. The obtained results show that the proposed model achieves high-quality precision in diagnosis and prediction. In another work, Ganji et al. [22] proposed FCS-ANTMINER for the diagnosis of diabetes in which set of fuzzy rules are extracted using an ant colony-based classification system. The proposed model uses artificial ants to explore state space and progressively generate fuzzy rules. The authors estimated the parameters in such a way that the cooperation and competition between the ants to discover more precise rules is balanced. The proposed scheme achieves high accuracy and accurately identifies diabetes. Karegowda et al. [23] proposed to integrate GA and back propagation network (BPN) for diagnosis of diabetes. The proposed scheme relies on estimating the optimal network connection weights of BPN with the help of GA. Similarly, Aslam et al. [24] classified diabetes using genetic programming (GP)-based model that performs feature selection using GP, *F*-score selection, and *t* test. Further, KNN and SVM classifiers are employed to test the GP generated classification features. Similarly, Han et al. [25] proposed a hybrid model that utilized SVM to screen diabetes mellitus. The work employed an ensemble learning module dedicated to generate transparent rules using SVM's black box to solve the imbalance problem.

Hayashi et al. [26] proposed to combine rule extraction algorithm and sampling selection technique to achieve interpretable and accurate classification rules for PID data set. Similarly, Li et al. [27] proposed a probabilistic fuzzy-based classification framework that overcomes the fuzzy uncertainties and stochastic uncertainties. The work achieved better classification performance and effectiveness on lower back pain diagnosis and PID data. Cheruku et al. [28] proposed RST-BatMiner, a hybrid decision support system that relies on eliminating the redundant features from the data set using a rough set theory (RST)-based Quick-Reduct scheme. Further, the proposed fitness function is minimized using bat optimization algorithm (BOA) to generate fuzzy rules. In another work, Sharma et al. [29] proposed a novel guided stochastic gradient descent (GSGD) approach that employs greedy selection scheme to overcome the issues of inconsistency in a dataset. The proposed scheme achieved enhanced CA and convergence as compared to its counterparts. Wang et al. [30] proposed a prediction algorithm to classify diabetes mellitus by balancing data class distribution using oversampling technique. In this work, the missing data values are compensated using Naïve Bayes algorithm and the predictions are generated using random forest (RF) classifiers. The proposed work achieves CA of 87.10% on PID dataset.

In another work, Ontiveros et al. [31] proposed a shadowed Type-2 fuzzy inference system (FIS) to mitigate the computational cost and provide better approximation. Similarly, Zhang et al. [32] proposed a fuzzy partition classifier, aimed to achieve enhanced classification performance in diabetes diagnosis exploiting its strong interpretability and uncertainty handling capability. The proposed scheme employs fuzzy clustering to partition the training data set into several subsets, and use fuzzy weighted algorithm for final prediction of each classifier. The obtained results confirm that it provides enhanced interpretability and classification performance. Similarly, Das et al. [33] proposed a medical disease classification approach that generates membership values using linguistic neuro-fuzzification (LNF) process and extracts the significantly contributing features using feature extraction algorithms. Authors validated the proposed model using eight benchmark datasets.

Nnamoko et al. [34] proposed a selective data pre-processing scheme aimed to achieve even distribution among the artificially generated subsets. In this work, authors identified outliers, performed oversampling and used SMOTE to balance the training data. Similarly, Ameena et al. [35] aimed to predict and detect diabetes using Pima Indian women dataset. The work focuses on finding the accuracy of existing prediction models for diabetes analysis using various ML techniques such as RF, SVM, DT and logical regression. In another work, Tan et al. [36] presented a case study demonstrating high burden of cardio-metabolic risk among Asian youth having Type 2 diabetes. Further, the study highlighted glomerular hyperfiltration as a strong Type 2 diabetes predictor. The following conclusions can be drawn from the above discussed literature reviews.

- The predictive accuracy of diagnosis of diabetes remains a challenging problem and requires further investigation.
- Missing values problem, curse of dimensionality, outlier's detection and class imbalance issues are common phenomenon in medical dataset that directly or indirectly affect the outcome of the classification system.
- Due to class imbalance in the diabetes dataset (like PIDD), CA is alone inadequate to determine the efficiency of the system.

3 Related terminologies

Improvement in the healthcare industries can significantly contribute towards the economic development of the nation because a healthier individual is capable of carrying out workplace tasks more efficiently as compared to any unhealthy person. The use of technology such as ML plays a major role in developing healthcare infrastructures as it can aid in the treatment, diagnosis and prevention of various health conditions. ML along with techniques of data mining such as classification [37, 38], clustering [39], regression [40, 41] and feature selection [42–44] are the main tools for developing an efficient healthcare system. ML operates on a basic principle—if you input garbage, you'll get garbage. In this work, garbage refers to noise, outlier, and class imbalance in the dataset. Prediction using class imbalanced dataset is prejudiced in favor of the common class or majority class. The dataset used in this work is imbalanced and therefore there is a need to oversample the dataset. For this purpose, SMOTE is used to produce class-balanced data. As every individual feature in not required for training a system, the proposed prediction model considers only the most significant features. The proposed model uses the concept of correlation and GA for feature selection. This helps to address the issues related to training complexity, performance and curse of dimensionality in the prediction system. Finally, DT is employed to achieve the main objective of prediction. The proposed model employs GA, DT and SMOTE as explored in the subsections below.

3.1 Genetic algorithm (GA)

GA is a searching scheme based on natural genetic mechanism and natural reduction [45]. Based on the concept of "survival of the fittest", GA makes use of random genetic operators to eliminate the poorer, and generate new promising solutions. The novel unknown area of the search space is found by constantly utilizing the information related to the best solutions. This movement of GA towards the best direction makes it similar to tabu searching and simulated annealing algorithm [46, 47]. Therefore, GA can also be considered as a directed random searching approach. Equation 1 presents the formal definition of GA.

$$\mathrm{GA} = \{P(0), N, g, s, l, p, f, t\}, \tag{1}$$

where $P(0) = (x_1(0), x_2(0), \dots, x_N(0)) \in I^N$, denotes the initial population; N denotes the initial population size; g denotes the genetic operators; s denotes the reduction strategy; l denotes the length of string (chromosome); f denotes the fitness function $[f : I \rightarrow R^+]$; and t represents a termination law $[t : I^N \rightarrow \{0, 1\}]$.

Abundancy of redundant and irrelevant features in the modern medical dataset lowers the efficacy of the existing data mining techniques leading to uninterpretable results. This is known as Hughes phenomenon [48]. However, appropriate attribute selection might yield interpretable and accurate results. This highlights the need for pre-processing phase in data mining. To overcome the issues of Hughes phenomenon, data reduction in the proposed model is done via attribute subset selection [49]. In the proposed architecture, the attribute selection is done using CFS-GA. CFS (correlation-based feature selection) is an attribute selection scheme that obtains final feature subset by heuristic evaluation for a single feature in every category label. Equation 2 represents the assessment method of CFS.

$$A_s = \frac{m \times \overline{\mathrm{MCD}}_{\mathrm{al}}}{\sqrt{m + m(m-1) + \overline{\mathrm{MCD}}_{\mathrm{aa}}}}, \tag{2}$$

where A_s represents the evaluation of an attribute subset s with m items, $\overline{\mathrm{MCD}}_{\mathrm{aa}}$ represents the mean correlation degree between various attributes and $\overline{\mathrm{MCD}}_{\mathrm{al}}$ represents the mean correlation degree between category label and the attributes. Higher evaluation value is produced by bigger $\overline{\mathrm{MCD}}_{\mathrm{al}}$ or smaller $\overline{\mathrm{MCD}}_{\mathrm{aa}}$. The correlation degree can be estimated by information gain as shown in Eqs. (3) and (4).

$$H(C) = -\sum_{c \in C} p(c) \times \log_2(p(c)), \tag{3}$$

$$H(C|D) = -\sum_{d \in D} p(d) \sum_{c \in C} p(c|d) \times \log_2(p(c|d)), \tag{4}$$

where c is any possible value of the category attribute C. $H(C)$ and $H(C|D)$ represents the entropy of C and entropy of C under the condition D respectively. Therefore, the entropy reduction of attribute C can be estimated as

$$\mathrm{ER}_C = H(C) - H(C|D). \tag{5}$$

As ER_C represents the amount of information provided by attribute C to attribute D, higher value of ER_C reflects a higher correlation degree between these attributes. For an effective comparison among attributes, normalization of information gain to [0, 1] is necessary as these tend to select attributes possessing higher values. Comparison effect among C and D can be estimated as

$$U_{CD} = 2.0 \times \frac{H(C) - H(C|D)}{H(C) + H(D)}. \tag{6}$$

Even though the algorithm shows better performance in terms of dimension reduction, it does not achieve a global optimum result. Considering its global search capability, GA is a wrapping scheme for dimension reduction. The proposed scheme combines CFS and GA to make a hybrid CFS-GA algorithm that operates in four parts: *coding scheme* in which every entity is encoded using binary codes; *selection operator* that employs roulette wheel method; *crossover operator* that produces new individuals by swapping the cross points; and *mutation operator* that uses bit mutation in binary encoding. Description of the proposed hybrid CFS-GA algorithm is presented as Algorithm 1.

Algorithm 1: CFS-GA Algorithm

```
Input:
Initial amount of population ($P$);
Iteration number of population ($g$);
Off springs in $t^{th}$ generation $C(t)$;
Parents in $t^{th}$ generation $P(t)$;
Mutation rate $R_{mutation}$;
Crossover rate $R_{crossover}$;
Binary coded encoding records of the dataset;
Selection operator;
Output:
Selected attributes
CFS-GA Algorithm:
START
$t \leftarrow 0$
Initialize $g$, $R_{mutation}$ and $R_{crossover}$;
Initialize $P(t)$;
Evaluate fitness of $P(t)$;
While (not terminating criteria) do
{
    Generate $C(t)$ from $P(t)$ through crossover operation;
    Generate $C'(t)$ from $C(t)$ through mutation operation;
    Evaluate fitness of $C'(t)$;
    Select $P(t+1)$ from $P(t)$ and $C'(t)$;
    $t \leftarrow t+1$;
}
Return selected attributes
END
```

Importance of GA is outlined as follows.

- It is beneficial to use GA as it helps the right approach to come from the best of previous solutions. GA improves the candidate solution over time. GA's theory is to unify different solutions to derive the best genes (features) from every generation and generate better solutions in subsequent generations.
- Data sets with multiple characteristics can be controlled by GA.
- These may not require particular domain knowledge for computation.

3.2 Decision tree

In contrast to other classification techniques, DT is a white box model and also an active learning scheme [50]. DT comprise of several leaf nodes, some internal nodes and a single root node. A decision tree is shown in the Fig. 1, with its root at the top. In the figure square shape shows condition or interior node, in view of which the tree parts into different branches or edges. The end of the branch or edge that does not split any longer is the decision or leaf and is shown using oval shape. Every leaf node possesses a class label and is connected to the root node via internal nodes. The starting node of a DT is the root node and the path from this node to the leaf nodes yields the classification rules. System operators can use these rules as guidelines to assess and monitor real-time voltage stability.

In this work, we use C4.5 DT algorithms that make use of information gain ratio for attribute selection. The employed C4.5 algorithm solves the over-fitted problem and is capable of effectively handling continuous attributes [50, 51]. The computation procedure of C4.5 algorithm can be described in five steps as detailed below.

1) The initial information entropy for the dataset S is calculated as

$$\text{Entropy}(S) = -\sum_{a=1}^{m} p_a \times \log_2\left(p_a\right), \tag{7}$$

where m represents the total number of classes and p_a represents the percentage of class a sample among these. This can result in two cases.

Case 1: If distinct class labels are assigned to all the data, $\left[p_a = \frac{1}{m}\right]$, then Entropy $(S) = \log_2 m$ (highest).

Case 2: If same class label is assigned to all the data, $[p_a = m = 1]$, then $\text{Entropy}(S) =$ zero (lowest).

(2) Partition S into two attribute partitions (S_{left} and S_{right}). The split entropy for every subset S is calculated as

$$\text{Entropy}_A(S) = \frac{\left|S_{\text{left}}\right|}{|S|} \times \text{Entropy}\left(S_{\text{left}}\right) + \frac{\left|S_{\text{right}}\right|}{|S|} \times \text{Entropy}\left(S_{\text{right}}\right), \tag{8}$$

where A is an attribute of S. $|S|$, $\left|S_{\text{left}}\right|$ and $\left|S_{\text{right}}\right|$ represents the number of samples in S, S_{left} and S_{right} respectively.

(3) Information gain of attribute A is obtained as

$$\text{Information}_{\text{gain}} = \text{Entropy}(S) - \text{Entropy}_A(S) \tag{9}$$

Higher value of $\text{Information}_{\text{gain}}$ denotes more entropy reduction resulting in a better attribute.

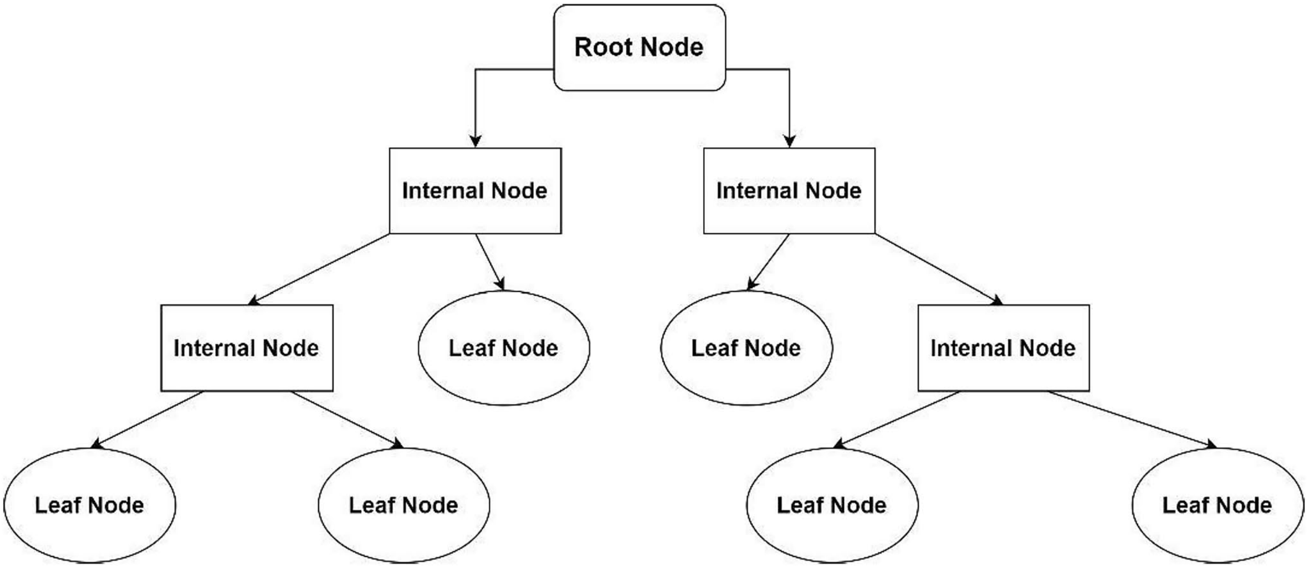

Fig. 1 A typical decision tree architecture

(4) To normalize the information gain and avoid overfitted problems, C4.5 algorithm introduces a split information value estimated as

$$\text{Split}_{\text{info}}(A) = -\sum_{a=1}^{k} \frac{|S_a|}{|S|} \times \log_2\left[\frac{|S_a|}{|S|}\right]. \tag{10}$$

(5) For every node of a DT, information gain ratio is calculated as

$$\text{IG}_{\text{ratio}}(A) = \frac{\text{Information}_{\text{gain}}}{\text{Split}_{\text{info}}(A)}, \tag{11}$$

where $\text{IG}_{\text{ratio}}(A)$ represents the information gain ratio of attribute A. The attribute having higher value of IG_{ratio} is selected. This process is recursively executed to split S into several better subsets. The DT learning algorithm is presented as Algorithm 2.

Algorithm 2: DT Learning Algorithm

```
Input:
Subset of classified instances, S;
Output:
Decision Tree
DT Learning Algorithm:
procedure BUILD_DT
    repeat
        $maxInformation_{gain} \leftarrow 0$
        $Split_{info}(A) \leftarrow Null$
        $e \leftarrow Entropy_A(S)$
        for all Attributes $a$ in $S$ do
            $gain \leftarrow Information_{gain}(a, e)$
            if $(gain > maxInformation_{gain})$ then
                $maxInformation_{gain} \leftarrow gain$
                $Split_{info}(A) \leftarrow a$
            end if
        end for
        Partition $(S, Split_{info}(A))$
    until all partitions processed
end procedure
```

Importance of DT:

- understandable classification rules are generated from the training data;
- constructs the fastest tree;
- only necessary features are needed before all information is classified;
- finding leaf nodes allows the pruning of test results, decreasing the number of tests;
- whole dataset is scanned to build tree.

3.3 SMOTE

Chawla et al. [52] introduced an oversampling technique named SMOTE that utilize neighbouring information to create new artificial instances in contrast to other existing methods that relies on random oversampling of instances. SMOTE replicates and randomly increases the minority class thereby effectively balancing the class distribution. It relies on synthesizing new minority instances from existing ones and use linear interpolation to generate virtual training records. Pseudocode of SMOTE algorithm is presented as Algorithm 3.

Algorithm 3: Pseudocode of SMOTE algorithm

```
Input:
Number of nearest neighbours, $N_{NN}$;
Amount of SMOTE, $S$ %;
Number of Minority Class Samples (MCS), $N_{MCS}$;
Output:
$\left[\frac{S}{100} \times N_{MCS}\right]$ synthetic MCSs
SMOTE Algorithm:
START
    if ($S < 100$) then
        Randomize the $S$ MCSs
        $N_{MCS} = \frac{S}{100} \times N_{MCS}$
        S = 100
    end-if
$S = (int)\frac{S}{100}$ /*Amount of SMOTE converted to integral multiples of 100*/
$N_{NN}$ ← Number of nearest neighbours
$N_A$ ← Number of attributes
$Sample$ [ ][ ] /*original array for MCS*/
$index$ ← 0 /*counts the number of synthetically generated samples*/
$Synthetic$ [ ][ ] /*new array for synthetic samples*/
for $a \leftarrow 1\ to\ N_{NCS}$
    Compute $N_{NN}$ nearest neighbours for a
    update the $Array\ [n][n]$ with the indices of these neighbours
    Populate ($S,\ a,\ Array\ [n][n]$)
end for
while $S \neq 0$
Choose a random number $N_{ran}$ between $N_{NN}\ and$ 1.
for $attribute \leftarrow 1\ to\ A$
        Compute difference between original minority instance and neighbour
        Compute $gap \leftarrow RangeRandom\ (0, 1)$
    end-for
$index++$
$S = S - 1$
end-while
return
END
```

The main reason behind using SMOTE is enumerated as below:

- SMOTE is used to solve the class imbalance problem in classification;
- independent on underlying classifier;
- can be easily implemented.

3.4 Proposed PMSGD model

In the previous section, we discussed about the various concepts of ML that are used to solve the aforementioned problems associated with the existing diabetes prediction system. The general architecture of our proposed prediction model can be divided into four layers namely pre-processing layer, dimensionality reduction layer, training layer and performance evaluation layer. The functionality of these layers is discussed in the subsections below.

3.4.1 Pre-processing layer

In this layer, pre-processing of the dataset is performed for the following: (1) checking and handling missing values, (2) outliers' detection and handling, and (3) production of high-quality training datasets by oversampling the minority class (solves the class imbalance problem). As most of the existing artificial intelligence approaches neglect the minority class, these are prone to inconsistent results. This is the major issues in dealing with the imbalanced data sets. Therefore, the most significant output is success on the minority class.

3.4.2 Dimensionality reduction layer

The performance of machine learning algorithm depends on input variables. In case of more number of input variables, the performance of ML algorithms degrade. This may have a dramatic effect on the output of ML algorithms that fit on data with many input characteristics. In this layer, feature selection is employed to remove the insignificant features using correlation and GA from the PID dataset to reproduce the high-quality dataset. Owing to this reduction in the dimension of the dataset, the training complexity is reduced thereby resolving the issues of overfitting. The simulation in this layer reveals the four most significant features of individuals with diabetes namely glucose, BMI, diabetes pedigree function and age.

3.4.3 Training layer

In this layer, the DT-based prediction model is trained using different split of training data set. The training dataset is comprised of array of features and associated class labels. Through iterative process of C 4.5 DT algorithm, the proposed prediction model is trained that can be further used to predict the output for new inputs.

The models training phase starts from a series of pre-processed training data using the gain ratio concept. Each training set sample consists of an n-dimensional vector in which the sample is set feature values and the class in which the sample belongs. In the training process, select a node that most efficiently divides the set of samples into subsets enriched in one class or another is chosen for each node of the tree. The gain ratio is the partitioning criterion. To make the decision, the attribute with the highest gain ratio is picked. The procedure then recurses on the divided sub lists.

3.4.4 Performance evaluation layer

This layer is used to measure the effectiveness of the model. The performance of the proposed prediction model on the PIDD dataset is evaluated on different metrics such as CA, CE, precision, sensitivity, FM and AUROC.

Figure 2 depicts the framework of the proposed prediction model.

4 Experiment and analysis

The Experiment and analysis section provide the details of dataset used, experimental environment, statistical study of dataset, and the results of the prediction model on different split of datasets. This section also states that the significance of the proposed prediction with the help of comparative study.

4.1 Dataset

This dataset originated from National Diabetes and Digestive and Kidney Diseases Institute. The dataset's purpose is to predict whether a patient has diabetes or not, based on some diagnostic measures used in the dataset. Various restrictions have been imposed on choosing such instances from a database. In particular, all patients considered in this dataset are females of Pima Indian Diabetes dataset (PIDD) heritage who are at least 21 years old. The Training and Testing set is taken from the UCI Repository site (https://www.archive.ics.uci.edu/) [53–56]. The PIDD dataset is composed of 768 samples, with 268 diabetic and 500 non-diabetic samples and. This contains eight numerically valued features and a class number, where the value '0' diabetes negative and the value '1' means diabetes positive. Table 1 presents the statistical description of the dataset attributes and the visualization of the attribute values with respect to various other attributes are depicted in Fig. 3. Visualization of data helps to curate information in such a way that it is easy to identify patterns and outliers. A successful visualization eliminates the noise from the information and shows the useful details. In the proposed scheme, pre-processing phase is capable of handling the noise and outliers [57–61].

4.2 Experimental environment and simulation parameters

Experiments are performed on a PC with Intel(R) Core (TM) i7 7th generation and 8 GB memory, running on Windows 10. For simulation, weka ML library and java 1.8 is used. To get the uniformity in the results, the proposed algorithm is executed ten times with all the variations of the dataset and the best outcomes are recorded. Three different types of simulation strategies have been performed on the proposed PMSGD model using PIDD dataset. These simulation strategies are as follows: (1) with and without oversampling (2) with and without feature selection (3) with and without feature selection and oversampling.

4.3 Performance measures

The confusion matrix describes the classifier's performance by contrasting the real classes and those projected classes. The confusion matrix for binary classification is composed of quadrants as shown in Table 2. True positive (TP) is a measure in which the model predicts the positive class as positive. False positive (FP) is a measure in which the model predicts the positive class as negative. True negative (TN) is a measure in which the model predicts the negative class as negative. False negative (FN) is a measure in which the model predicts the positive class as negative.

The performance indicators such as CA, CE, precision, sensitivity, FM and AUROC are quantified in accordance with the confusion matric. CA is defined as the proportion of correctly classified tuples and the total tuples. CE is the proportion of incorrectly classified tuples and the total number of tuples. Precision is the proportion of TP and the predicted positive tuples. Sensitivity is the proportion of TP and positive samples. AUROC curve gives the area under recall and false positive rate. It tells how much the model is fit for recognizing classes. Higher the value, better the model is at classifying 0 s as 0 s and 1 s as 1 s. By example, the higher the value, the better the model is to distinguish between patients with disease and no diseases. The calculation of these indicators is as below.

$$\mathrm{Classificationaccuracy(CA)} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \quad (12)$$

$$\mathrm{Classificationerror(CE)} = \frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \quad (13)$$

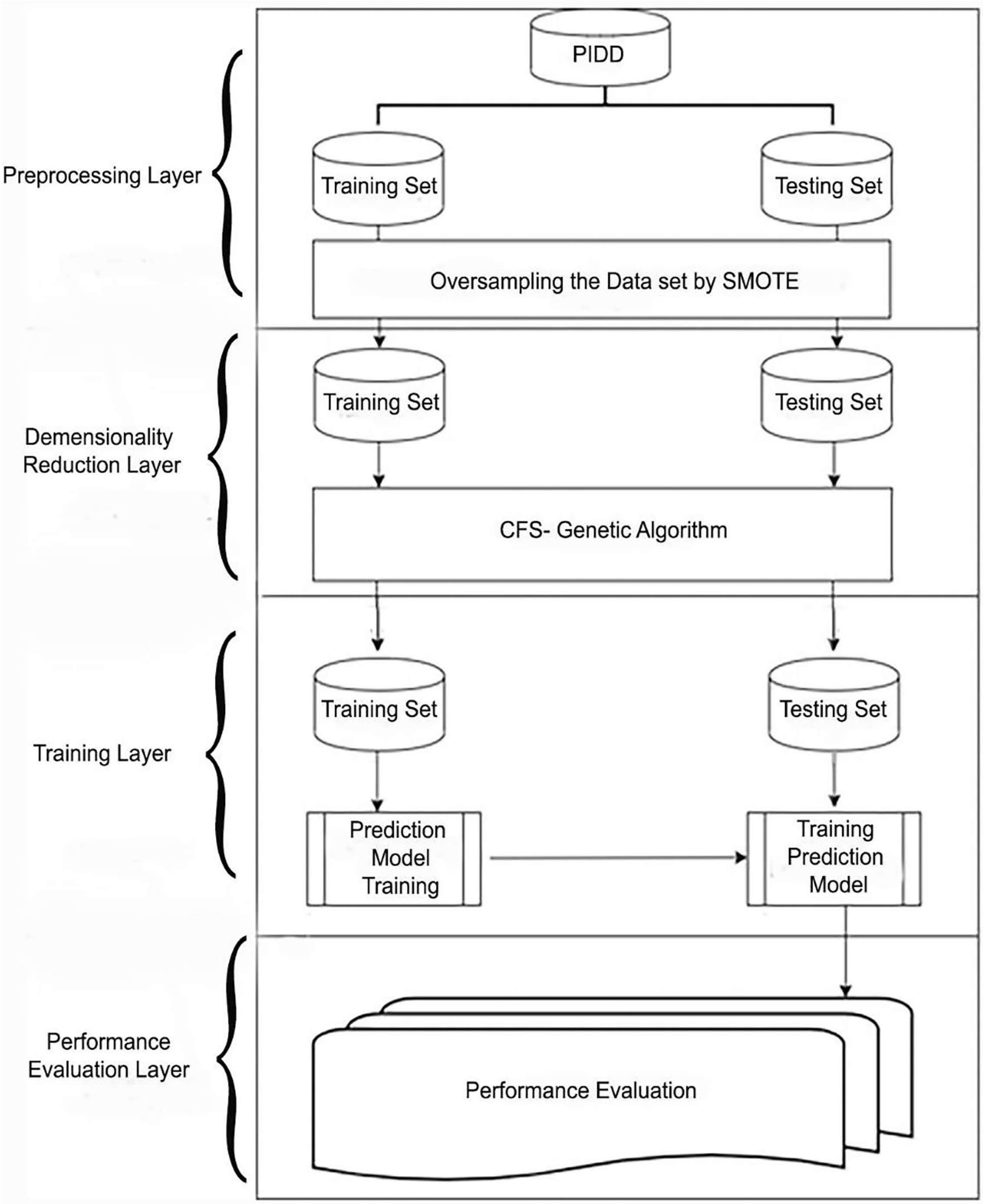

Fig. 2 The framework of proposed PMSGD model

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{14}$$

$$Recall(sensitivity) = \frac{TP}{TP + FN}, \tag{15}$$

$$F_{\text{Measure}} = \frac{2 \times \text{recall} \ \times \text{precision}}{\text{recall} + \text{precision}}. \tag{16}$$

Table 1 Dataset attribute statistical description

ID	Attribute	Type	Statistics			
			Minimum	Maximum	Mean	Standard deviation
1	Pregnancies	Numeric	0	17	3.845	3.37
2	Glucose		0	199	120.895	31.973
3	Blood pressure		0	122	69.105	19.356
4	Skin thickness		0	99	20.536	15.952
5	Insulin		0	846	79.799	115.244
6	BMI		0	67.1	31.993	7.884
7	Diabetes pedigree function		0.078	2.42	0.472	0.331
8	Age		21	81	33.241	11.76
9	Outcome	Nominal	Positive 268; Negative 500			

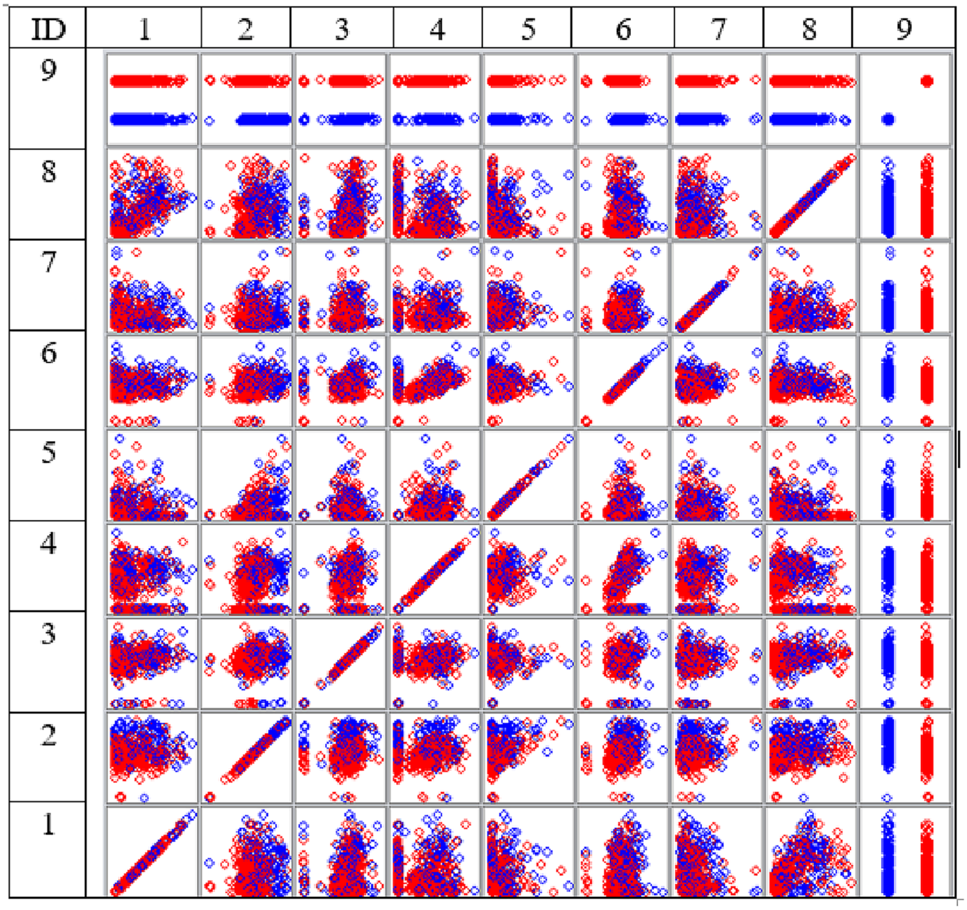

Fig. 3 Visualization of attribute values

Table 2 Confusion matrix

	Predicted	
	True	False
Actual		
True	TP	FN
False	FP	TN

4.4 Results and discussions

The split test methodology is implemented as a technique for planning and validating the results for the training and test dataset. The main motivation behind the selection of split test methodology is as follows:

Table 3 Features selected using GA on PIDD

Attribute	Type	Statistics			
		Minimum	Maximum	Mean	Standard deviation
Glucose	Numeric	0	199	120.895	31.973
BMI		0	67.1	31.993	7.884
Diabetes pedigree function		0.078	2.42	0.472	0.331
Age		21	81	33.241	11.76
Outcome	Nominal			Positive 268; Negative 500	

Table 4 Features selected using GA on PIDD + SM

Attribute	Type	Statistics			
		Minimum	Maximum	Mean	Standard Deviation
Pregnancies	Numeric	0	17	4.084	3.349
Glucose		0	199	126.123	32.443
Insulin		0	846	84.894	121.337
BMI		0	67.1	32.765	7.522
Diabetes pedigree function		0.078	2.42	0.494	0.332
Age		21	81	34.2	11.43
Outcome	Nominal			Positive 536; Negative 500	

- The problem with training and testing on same data is that you will only know the output of the model on the datasets, but have no idea of how the algorithm would perform on data in which the model was not trained.
- The problem with multiple split tests is that few instances of data might never be used for training. This leads to distorted results that do not give a clear indication of the algorithm's accuracy.
- Cross-validation is unbiased estimation of the efficiency of the methods on unknown data. If randomness is used by the method itself, it will lead to different results for the same training data each time a different random number of seed (start of the pseudo-randomness sequence) was trained. Cross-validation does not compensate for the uncertainty in the results of the algorithm.

In the split test methodology, we tested the considered model for different percentage splits such as 60–40%, 65–35%, 70–30%, 75–25% and 80–20%. Here, the first part represents the size of training set and the second part represents the size of the testing set. The method is simulated ten times for each split and the best five outcomes are recorded for each data set. The four varieties of datasets are used for performing training and testing namely PIDD, PIDD + SM, PIDD + GA and PIDD + SM + GA. PIDD is the PIMA Indian Diabetes Dataset, PIDD + SM is the oversampled data set using SMOTE. PIDD + GA is the datasets with features selected using correlation and GA. PIDD + SM + GA is the dataset that is over sampled using SMOTE and features selected using correlation and GA. The features selected using GA on PIDD are shown in Table 3.

Table 5 Parameter configuration

Parameter	PIDD + GA	PIDD + SM + GA
Crossover probability	0.6	0.7
Mutation probability	0.03	0.03
Population size	20	20
Maximum generations	50	60

The features selected using GA on oversampled dataset using SMOTE is shown in Table 4.

Table 5 shows the parameter configuration for best-selected features using GA.

In ML, the CA is frequently used as the performance measure for diabetes research. Because of the class imbalance in the diabetes dataset (like PIDD), CA is alone inadequate to determine the efficiency of the system. CA alone is inadequate for evaluating efficiency as stated in the related work. To assess and equate the proposed prediction model, the following three simulation scenarios are performed.

- Classification using C 4.5 Decision Tree classifier with PIDD, PIDD + GA, PIDD + SM and PIDD + SM + GA.

- Evaluation of the trained model against a series of metrics such as CA, CE, precision, sensitivity, FM and AUROC.
- Outcome of the proposed prediction model is compared with other standard existing systems in terms of the CA, CE, precision, sensitivity, FM and AUROC.

Table 6 60–40 Training–testing result

Key_Dataset	CA	CE	Precision	Sensitivity	FM	AUROC
PIDD	76.6234	23.3766	0.7500	0.5000	0.6000	0.7754
PIDD	75.9740	24.0260	0.6977	0.5556	0.6186	0.7188
PIDD	75.3247	24.6753	0.7353	0.4630	0.5682	0.7291
PIDD	75.2443	24.7557	0.6867	0.5327	0.6000	0.7907
PIDD	74.5928	25.4072	0.8085	0.3551	0.4935	0.6551
PIDD + GA	76.9481	23.0519	0.6609	0.7037	0.6816	0.8036
PIDD + GA	76.2215	23.7785	0.7742	0.4486	0.5680	0.8196
PIDD + GA	75.9740	24.0260	0.6977	0.5556	0.6186	0.7188
PIDD + GA	75.5700	24.4300	0.7286	0.4766	0.5763	0.7374
PIDD + GA	75.3247	24.6753	0.6778	0.5648	0.6162	0.7273
PIDD + SM	78.5024	21.4976	0.7854	0.8037	0.7945	0.8230
PIDD + SM	75.4217	24.5783	0.7511	0.7860	0.7682	0.7789
PIDD + SM	74.8792	25.1208	0.7670	0.7383	0.7524	0.7952
PIDD + SM	74.3961	25.6039	0.7432	0.7710	0.7569	0.8011
PIDD + SM	73.9759	26.0241	0.7610	0.7256	0.7429	0.7790
PIDD + SM + GA	77.5362	22.4638	0.7642	0.8178	0.7901	0.8230
PIDD + SM + GA	76.8116	23.1884	0.8172	0.7103	0.7600	0.8278
PIDD + SM + GA	74.1546	25.8454	0.7421	0.7664	0.7540	0.7942
PIDD + SM + GA	73.9759	26.0241	0.7610	0.7256	0.7429	0.7790
PIDD + SM + GA	73.9130	26.0870	0.7172	0.8178	0.7642	0.7941

Table 7 65–35 Training–testing result

Key_Dataset	CA	CE	Precision	Sensitivity	FM	AUROC
PIDD	79.5539	20.4461	0.7910	0.5638	0.6584	0.7619
PIDD	75.4647	24.5353	0.7692	0.4255	0.5479	0.8013
PIDD	75.4647	24.5353	0.6250	0.7447	0.6796	0.7688
PIDD	75.0929	24.9071	0.6753	0.5532	0.6082	0.7272
PIDD	73.6059	26.3941	0.6386	0.5638	0.5989	0.7340
PIDD + GA	79.5539	20.4461	0.7910	0.5638	0.6584	0.7632
PIDD + GA	75.8364	24.1636	0.7843	0.4255	0.5517	0.7995
PIDD + GA	75.4647	24.5353	0.6250	0.7447	0.6796	0.7688
PIDD + GA	75.4647	24.5353	0.6628	0.6064	0.6333	0.7510
PIDD + GA	75.0929	24.9071	0.7368	0.4468	0.5563	0.6867
PIDD + SM	78.7879	21.2121	0.7581	0.8670	0.8089	0.8068
PIDD + SM	78.7293	21.2707	0.7696	0.8396	0.8031	0.8258
PIDD + SM	77.1350	22.8650	0.7778	0.7819	0.7798	0.7705
PIDD + SM	75.9669	24.0331	0.7688	0.7647	0.7668	0.8359
PIDD + SM	74.3802	25.6198	0.7387	0.7819	0.7597	0.8274
PIDD + SM + GA	78.7293	21.2707	0.7350	0.9198	0.8171	0.8078
PIDD + SM + GA	78.2369	21.7631	0.7535	0.8617	0.8040	0.8027
PIDD + SM + GA	75.4821	24.5179	0.7565	0.7766	0.7664	0.7601
PIDD + SM + GA	75.4144	24.5856	0.7753	0.7380	0.7562	0.8012
PIDD + SM + GA	73.5537	26.4463	0.7644	0.7074	0.7348	0.7803

Table 8 70–30 Training–testing result

Key_Dataset	CA	CE	Precision	Sensitivity	FM	AUROC
PIDD	77.0563	22.9437	0.6458	0.7654	0.7006	0.7865
PIDD	77.0563	22.9437	0.7414	0.5309	0.6187	0.7635
PIDD	75.3247	24.6753	0.6818	0.5556	0.6122	0.6790
PIDD	74.8918	25.1082	0.7091	0.4815	0.5735	0.6874
PIDD	73.5931	26.4069	0.6471	0.5432	0.5906	0.7519
PIDD + GA	77.9221	22.0779	0.7419	0.5679	0.6434	0.7776
PIDD + GA	76.5217	23.4783	0.7321	0.5125	0.6029	0.7765
PIDD + GA	75.3247	24.6753	0.6875	0.5432	0.6069	0.7341
PIDD + GA	74.8918	25.1082	0.6353	0.6667	0.6506	0.7762
PIDD + GA	74.8918	25.1082	0.7091	0.4815	0.5735	0.6874
PIDD + SM	78.4566	21.5434	0.7765	0.8199	0.7976	0.8270
PIDD + SM	77.4920	22.5080	0.7974	0.7578	0.7771	0.8053
PIDD + SM	76.8489	23.1511	0.7799	0.7702	0.7750	0.8251
PIDD + SM	76.8489	23.1511	0.7602	0.8075	0.7831	0.7719
PIDD + SM	75.8842	24.1158	0.7443	0.8137	0.7774	0.7679
PIDD + SM + GA	79.4212	20.5788	0.7513	0.9006	0.8192	0.8150
PIDD + SM + GA	78.4566	21.5434	0.7765	0.8199	0.7976	0.8270
PIDD + SM + GA	76.8489	23.1511	0.7799	0.7702	0.7750	0.8251
PIDD + SM + GA	76.2058	23.7942	0.7486	0.8137	0.7798	0.7774
PIDD + SM + GA	74.9196	25.0804	0.7515	0.7702	0.7607	0.7652

Table 9 75–25 Training–testing result

Key_Dataset	CA	CE	Precision	Sensitivity	FM	AUROC
PIDD	77.6042	22.3958	0.7143	0.5970	0.6504	0.7427
PIDD	76.5625	23.4375	0.7115	0.5522	0.6218	0.7756
PIDD	76.5625	23.4375	0.6833	0.6119	0.6457	0.7630
PIDD	76.5625	23.4375	0.7200	0.5373	0.6154	0.7273
PIDD	74.4792	25.5208	0.6552	0.5672	0.6080	0.8044
PIDD + GA	77.6042	22.3958	0.7143	0.5970	0.6504	0.7427
PIDD + GA	77.6042	22.3958	0.7222	0.5821	0.6446	0.8236
PIDD + GA	77.0833	22.9167	0.7347	0.5373	0.6207	0.7167
PIDD + GA	76.5625	23.4375	0.7037	0.5672	0.6281	0.7279
PIDD + GA	76.5625	23.4375	0.7200	0.5373	0.6154	0.7273
PIDD + SM	79.5367	20.4633	0.7956	0.8134	0.8044	0.8359
PIDD + SM	77.6062	22.3938	0.7405	0.8731	0.8014	0.7993
PIDD + SM	75.6757	24.3243	0.7752	0.7463	0.7605	0.7959
PIDD + SM	75.6757	24.3243	0.7178	0.8731	0.7879	0.7549
PIDD + SM	74.1313	25.8687	0.6959	0.8881	0.7803	0.7856
PIDD + SM + GA	79.9228	20.0772	0.8060	0.8060	0.8060	0.8473
PIDD + SM + GA	78.3784	21.6216	0.7438	0.8881	0.8095	0.8100
PIDD + SM + GA	76.4479	23.5521	0.7704	0.7761	0.7732	0.7942
PIDD + SM + GA	74.9035	25.0965	0.7066	0.8806	0.7841	0.7756
PIDD + SM + GA	74.5174	25.4826	0.7833	0.7015	0.7402	0.8025

Tables 6, 7, 8, 9 and 10 shows the simulation results of the top 5 outcomes of the proposed model on PIDD, PIDD + GA, PIDD + SM, PIDD + SM + GA datasets. The model is simulated ten times for each dataset and the top 5 outcomes are recorded. In each iteration, the dataset is randomised that may lead to change in its performance.

Table 6 depicts the simulation outcomes of PMSGD model in which 60% tuples are considered as training set and the remaining 40% is considered as a testing set. The

Table 10 80–20 Training–testing result

Key_Dataset	CA	CE	Precision	Sensitivity	FM	AUROC
PIDD	77.2727	22.7273	0.8065	0.4630	0.5882	0.8129
PIDD	77.2727	22.7273	0.7111	0.5926	0.6465	0.8050
PIDD	75.1634	24.8366	0.8261	0.3585	0.5000	0.7179
PIDD	74.6753	25.3247	0.6667	0.5556	0.6061	0.7766
PIDD	74.0260	25.9740	0.6591	0.5370	0.5918	0.7668
PIDD + GA	77.9221	22.0779	0.7273	0.5926	0.6531	0.8222
PIDD + GA	76.6234	23.3766	0.8750	0.3889	0.5385	0.7202
PIDD + GA	75.9740	24.0260	0.8148	0.4074	0.5432	0.7805
PIDD + GA	75.3247	24.6753	0.9444	0.3148	0.4722	0.7181
PIDD + GA	75.1634	24.8366	0.8261	0.3585	0.5000	0.7179
PIDD + SM	82.1256	17.8744	0.8070	0.8598	0.8326	0.8511
PIDD + SM	78.2609	21.7391	0.7870	0.7944	0.7907	0.8179
PIDD + SM	77.7778	22.2222	0.7440	0.8692	0.8017	0.7950
PIDD + SM	76.9231	23.0769	0.7344	0.8704	0.7966	0.8097
PIDD + SM	75.8454	24.1546	0.7664	0.7664	0.7664	0.7780
PIDD + SM + GA	80.1932	19.8068	0.7797	0.8598	0.8178	0.8490
PIDD + SM + GA	77.7778	22.2222	0.7798	0.7944	0.7870	0.8319
PIDD + SM + GA	77.7778	22.2222	0.7480	0.8598	0.8000	0.8161
PIDD + SM + GA	76.9231	23.0769	0.7273	0.8889	0.8000	0.8297
PIDD + SM + GA	76.8116	23.1884	0.7565	0.8131	0.7838	0.7862

following observations are noted in this simulation strategy with respect to accuracy.

- The best outcome is observed on PIDD + SM. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 78.5024, 21.4976, 0.7854, 0.8037, 0.7945 and 0.8230, respectively.
- The second-best outcome is observed on PIDD + SM + GA. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 77.5362, 22.4638, 0.7642, 0.8178, 0.7901 and 0.8230, respectively.
- The third best outcome is observed on PIDD + GA. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 76.9481, 23.0519, 0.6609, 0.7037, 0.6816 and 0.8036, respectively.
- The fourth best outcome is observed on PIDD. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 76.6234, 23.3766, 0.7500, 0.5000, 0.6000 and 0.7754, respectively.

  Table 7 depicts the simulation outcomes of PMSGD model in which 65% tuples are considered as training set and the remaining 35% is considered as a testing set. The following observations are noted in this simulation strategy with respect to accuracy.

- The best outcome is observed on PIDD. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 79.5539, 20.4461, 0.7910, 0.5638, 0.6584 and 0.7619, respectively.
- The second-best outcome is observed on PIDD + GA. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 79.5539, 20.4461, 0.7910, 0.5638, 0.6584 and 0.7632, respectively.
- The third best outcome is observed on PIDD + SM. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 78.7879, 21.2121, 0.7581, 0.8670, 0.8089 and 0.8068, respectively.
- The fourth best outcome is observed on PIDD + SM + GA. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 78.7293, 21.2707, 0.7350, 0.8171 and 0.8078, respectively.

  Table 8 depicts the simulation outcomes of PMSGD model in which 70% tuples are considered as training set and the remaining 30% is considered as a testing set. The following observations are noted in this simulation strategy with respect to accuracy.

- The best outcome is observed on PIDD + SM + GA. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 79.4212, 20.5788, 0.7513, 0.9006, 0.8192 and 0.8150, respectively.

- The second-best outcome is observed on PIDD + GA. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 77.9221, 22.0779, 0.7419, 0.5679, 0.6434 and 0.7776, respectively.
- The third best outcome is observed on PIDD + SM. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 77.4920, 22.5080, 0.7974, 0.7578, 0.7771 and 0.8053, respectively.
- The fourth best outcome is observed on PIDD. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 77.0563, 22.9437, 0.6458, 0.7654, 0.7006 and 0.7865, respectively.

Table 9 depicts the simulation outcomes of PMSGD model in which 75% tuples are considered as training set and the remaining 25% is considered as a testing set. The following observations are noted in this simulation strategy with respect to accuracy.

- The best outcome is observed on PIDD + SM + GA. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 79.9228, 20.0772, 0.8060, 0.8060, 0.8060 and 0.8473, respectively.
- The second-best outcome is observed on PIDD + SM. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 79.5367, 20.4633, 0.7956, 0.8134, 0.8044 and 0.8359, respectively.
- The third best outcome is observed on PIDD + GA. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 77.6042, 22.3958, 0.7143, 0.5970, 0.6504 and 0.7427, respectively.
- The fourth best outcome is observed on PIDD. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 77.6042, 22.3958, 0.7143, 0.5970, 0.6504 and 0.7427.

Table 10 depicts the simulation outcomes of PMSGD model in which 80% tuples are considered as training set and the remaining 20% is considered as a testing set. The following observations are noted in this simulation strategy with respect to accuracy.

- The best outcome is observed on PIDD + SM. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 82.1256, 17.8744, 0.8070, 0.8598, 0.8326 and 0.8511, respectively.
- The second-best outcome is observed on PIDD + SM + GA. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 80.1932, 19.8068, 0.7797, 0.8598, 0.8178 and 0.8490, respectively.
- The third best outcome is observed on PIDD + GA. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 77.9221, 22.0779, 0.7273, 0.5926, 0.6531 and 0.8222, respectively.
- The fourth best outcome is observed on PIDD. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 77.2727, 22.7273, 0.8065, 0.4630, 0.5882 and 0.8129, respectively.

Table 11 Performance evaluation with other existing methods

Techniques	CA	CE	Precision	Sensitivity	FM	AUROC
Naïve Bayes [62]	76.3000	23.700	0.759	0.7630	0.7600	0.8190
SVM [62]	65.1000	24.9000	0.424	0.6510	0.5130	0.5000
Decision Tree [62]	73.8200	26.1800	0.735	0.7380	0.7360	0.7510
RepTree [62]	74.8400	25.1600	0.6700	0.5300	0.5900	0.7600
KStar [62]	68.2300	31.7700	0.5800	0.3300	0.4200	0.6800
OneR [63]	70.8300	29.1700	0.6100	0.4600	0.5200	0.6500
PART [63]	74.3500	25.6500	0.7000	0.4700	0.5600	0.7700
SMO [63]	72.1400	27.8600	0.7800	0.2800	0.4100	0.6200
BayesNet [63]	73.8300	26.1700	0.6400	0.5700	0.6000	0.8100
PMSGD + (PIDD)	79.5539	20.4461	0.7910	0.5638	0.6584	0.7619
PMSGD + (PIDD + GA)	79.5539	20.4461	0.7910	0.5638	0.6584	0.7632
PMSGD + (PIDD + SM)	82.1256	17.8744	0.8070	0.8598	0.8326	0.8511
PMSGD + (PIDD + SM + GA)	80.1932	19.8068	0.7797	0.8598	0.8178	0.8490

4.5 Performance evaluation with existing systems

The proposed method is compared on the basis of CA, CE, precision, sensitivity, FM, and AUROC. It is worth to mention that the proposed model yields superior results in comparison to the various existing schemes as shown in the Table 11. The best outcome is observed on the PIDD + SM data set. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 82.1256, 17.8744, 0.8070, 0.8598, 0.8326 and 0.8511, respectively. The second-best outcome is observed on PIDD + SM + GA data set. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 80.1932, 19.8068, 0.7797, 0.8598, 0.8178 and 0.8490, respectively. The third best outcome is observed on both PIDD and PIDD + GA dataset. The outcome of the considered performance indicators namely CA, CE, precision, sensitivity, FM and AUROC are 79.5539, 20.4461, 0.7910, 0.5638, 0.6584 and 0.7619, respectively.

The proposed PMSGD model addresses the issues of missing values, outlier detection and its handling in preprocessing. The dataset used in this work suffers from class imbalance problem. The proposed model solves this problem by oversampling the minority class using SMOTE that yields high-quality training datasets. Important attributes in the datasets are selected via feature selection method to eliminate the insignificant features to generate high-quality datasets using Correlation and GA. This reduced dimension of the dataset lowers the training complexity and solves the issues of over fitting. Further the processed data is used to predict whether the testing instance is suffering from diabetes or not. The remarkable observation observed in this experimentation is that the proposed PMSGD model outperforms other techniques as given in [45, 46]. The best outcome achieved by the proposed system in terms of CA, CE, precision, sensitivity, FM and AUROC is 82.1256%, 17.8744%, 0.8070%, 0.8598, 0.8326 and 0.8511, respectively.

5 Conclusion

In this paper, a novel PMSGD prediction model is proposed for diabetes disease classification that also addresses the problems of data imbalance, curse of dimensionality and missing data values in the diabetes datasets. The difficulty of dealing with imbalanced data sets is that most AI approaches neglect the minority class thereby leading to inconsistent results. In this regard, the proposed model uses SMOTE to oversample the minority class in its pre-processing stage whereas makes use of correlation and GA to extract significant features. Through simulation of feature selection, it is observed that Glucose, BMI, Diabetes Pedigree Function and Age are the significant features of individuals in the PIDD. On the basis of the outcome of feature selection, the training and testing sets are formed. The training set is used to train the proposed PMSGD prediction model and testing set is used to test its efficacy. The proposed model outperforms the existing models in terms of various metrics such as CA, CE, precision, sensitivity, FM, and AUROC. The best outcome achieved by the proposed system in terms of CA, CE, precision, sensitivity, FM and AUROC is 82.1256%, 17.8744%, 0.8070%, 0.8598, 0.8326 and 0.8511, respectively. In future work, the proposed model can be tested for automatic diabetes analysis and prediction with high precision. Testing its applicability to diagnose other diseases can serve as another research direction. Also, pruning the rule sets of the proposed PMSGD model can be an interesting future research work. Furthermore, implementation of various nature-inspired algorithms such as PSO, ACO, grass hopper optimization, grey wolf optimization, Jaya algorithm or fruit fly optimization may be investigated so as to increase the accuracy and reduce the dimensionality of the dataset and consequently mitigate the time complexity.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Amutha, A., Mohan, V.: Diabetes complications in childhood and adolescent onset type 2 diabetes—a review. J. Diabetes Complicat. **30**(5), 951–957 (2016). https://doi.org/10.1016/j.jdiacomp.2016.02.009
2. Domingueti, C.P., Dusse, L.M., Carvalho, M.D., Sousa, L.P., Gomes, K.B., Fernandes, A.P.: Diabetes mellitus: the linkage between oxidative stress, inflammation, hypercoagulability and vascular complications. J. Diabetes Complicat. **30**(4), 738–745 (2016). https://doi.org/10.1016/j.jdiacomp.2015.12.018
3. World health organization statistics on diabetes. http://www.who.int/mediacentre/factsheets/fs312/en/. Accessed 02 Mar 2020
4. Pham, H.N., Triantaphyllou, E.: Prediction of diabetes by employing a new data mining approach which balances fitting and generalization. Comput. Inf. Sci. Stud. Comput. Intell. (2008). https://doi.org/10.1007/978-3-540-79187-4_2
5. Wild, S., Roglic, G., Green, A., Sicree, R., King, H.: Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care **27**(5), 1047–1053 (2004). https://doi.org/10.2337/diacare.27.5.1047
6. Wang, X., Bi, D., Wang, S.: Fault recognition with labeled multi-category support vector machine. In: Third international conference on natural computation (ICNC 2007). https://doi.org/10.1109/icnc.2007.382(2007)
7. Zhang, B., Wei, Z., Ren, J., Cheng, Y., Zheng, Z.: An empirical study on predicting blood pressure using classification and regression trees. IEEE Access **6**, 21758–21768 (2018). https://doi.org/10.1109/access.2017.2787980

8. Tejedor, M., Woldaregay, A.Z., Godtliebsen, F.: Reinforcement learning application in diabetes blood glucose control: A systematic review. Artif. Intell. Med. **104**, 101836 (2020). https://doi.org/10.1016/j.artmed.2020.101836
9. Pramanik, P.K., Solanki, A., Debnath, A., Nayyar, A., El-Sappagh, S., Kwak, K.: Advancing modern healthcare with nanotechnology, nanobiosensors, and internet of nano things: taxonomies, applications, architecture, and challenges. IEEE Access **8**, 65230–65266 (2020). https://doi.org/10.1109/access.2020.2984269
10. Nielsen, K.B., Lautrup, M.L., Andersen, J.K., Savarimuthu, T.R., Grauslund, J.: Deep learning-based algorithms in screening of diabetic retinopathy: a systematic review of diagnostic performance. Ophthalmology Retina **3**(4), 294–304 (2019). https://doi.org/10.1016/j.oret.2018.10.014
11. Remeseiro, B., Bolon-Canedo, V.: A review of feature selection methods in medical applications. Comput. Biol. Med. **112**, 103375 (2019). https://doi.org/10.1016/j.compbiomed.2019.103375
12. Santos, B.S., Steiner, M.T., Fenerich, A.T., Lima, R.H.: Data mining and machine learning techniques applied to public health problems: a bibliometric analysis from 2009 to 2018. Comput. Ind. Eng. **138**, 106120 (2019). https://doi.org/10.1016/j.cie.2019.106120
13. Rendón, E., Alejo, R., Castorena, C., Isidro-Ortega, F.J., Granda-Gutiérrez, E.E.: Data sampling methods to deal with the big data multi-class imbalance problem. Appl. Sci. **10**(4), 1276 (2020). https://doi.org/10.3390/app10041276
14. Kumar, A., Krishnamurthi, R., Nayyar, A., Sharma, K., Grover, V., Hossain, E.: A novel smart healthcare design, simulation, and implementation using healthcare 4.0 processes. IEEE Access **8**, 118433–118471 (2020). https://doi.org/10.1109/access.2020.3004790
15. Thabtah, F., Hammoud, S., Kamalov, F., Gonsalves, A.: Data imbalance in classification: experimental evaluation. Inf. Sci. **513**, 429–441 (2020). https://doi.org/10.1016/j.ins.2019.11.004
16. Hu, T., Sung, S.Y.: Detecting pattern-based outliers. Pattern Recognit. Lett. **24**(16), 3059–3068 (2003). https://doi.org/10.1016/s0167-8655(03)00165-x
17. Maniruzzaman, M., Rahman, M.J., Al-Mehedihasan, M., Suri, H.S., Abedin, M.M., El-Baz, A., Suri, J.S.: Accurate diabetes risk stratification using machine learning: role of missing value and outliers. J. Med. Syst. (2018). https://doi.org/10.1007/s10916-018-0940-7
18. Ijaz, M., Alfian, G., Syafrudin, M., Rhee, J.: Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest. Appl. Sci. **8**(8), 1325 (2018). https://doi.org/10.3390/app8081325
19. Shuja, M., Mittal, S., Zaman, M.: Effective prediction of type II diabetes mellitus using data mining classifiers and SMOTE. Adv. Comput. Intell. Syst. Algorithms Intell. Syst. (2020). https://doi.org/10.1007/978-981-15-0222-4_17
20. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., Tang, H.: Predicting diabetes mellitus with machine learning techniques. Front. Genet. (2018). https://doi.org/10.3389/fgene.2018.00515
21. Barakat, N., Bradley, A.P., Barakat, M.N.: intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Trans. Inf Technol. Biomed. **14**(4), 1114–1120 (2010). https://doi.org/10.1109/titb.2009.2039485
22. Ganji, M.F., Abadeh, M.S.: A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. Expert Syst. Appl. **38**(12), 14650–14659 (2011). https://doi.org/10.1016/j.eswa.2011.05.018
23. Karegowda, A.G., Manjunath, A., Jayaram, M.: Application of genetic algorithm optimized neural network connection weights for medical diagnosis of PIMA Indians diabetes. Int. J. Soft Comput. **2**(2), 15–23 (2011). https://doi.org/10.5121/ijsc.2011.2202
24. Aslam, M.W., Zhu, Z., Nandi, A.K.: Feature generation using genetic programming with comparative partner selection for diabetes classification. Expert Syst. Appl. **40**(13), 5402–5412 (2013). https://doi.org/10.1016/j.eswa.2013.04.003
25. Han, L., Luo, S., Yu, J., Pan, L., Chen, S.: 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 (2015). https://doi.org/10.1109/jbhi.2014.2325615
26. Hayashi, Y., Yukita, S.: Rule extraction using recursive-rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. Inform. Med. Unlocked **2**, 92–104 (2016). https://doi.org/10.1016/j.imu.2016.02.001
27. Li, H., Wang, Y., Zhang, G.: Probabilistic fuzzy classification for stochastic data. IEEE Trans. Fuzzy Syst. **25**(6), 1391–1402 (2017). https://doi.org/10.1109/tfuzz.2017.2687402
28. Cheruku, R., Edla, D.R., Kuppili, V., Dharavath, R.: RST-Bat-Miner: a fuzzy rule miner integrating rough set feature selection and Bat optimization for detection of diabetes disease. Appl. Soft Comput. **67**, 764–780 (2018). https://doi.org/10.1016/j.asoc.2017.06.032
29. Sharma, A.: Guided stochastic gradient descent algorithm for inconsistent datasets. Appl. Soft Comput. **73**, 1068–1080 (2018). https://doi.org/10.1016/j.asoc.2018.09.038
30. Wang, Q., Cao, W., Guo, J., Ren, J., Cheng, Y., Davis, D.N.: DMP_MI: an effective diabetes mellitus classification algorithm on imbalanced data with missing values. IEEE Access **7**, 102232–102238 (2019). https://doi.org/10.1109/access.2019.2929866
31. Ontiveros-Robles, E., Melin, P.: A hybrid design of shadowed type-2 fuzzy inference systems applied in diagnosis problems. Eng. Appl. Artif. Intell. **86**, 43–55 (2019). https://doi.org/10.1016/j.engappai.2019.08.017
32. Zhang, X., Jiang, Y., Hu, W., Wang, S.: A parallel ensemble fuzzy classifier for diabetes diagnosis. J. Med. Imaging Health Inform. **10**(3), 544–551 (2020). https://doi.org/10.1166/jmihi.2020.2972
33. Das, H., Naik, B., Behera, H.: Medical disease analysis using neuro-fuzzy with feature extraction model for classification. Inform. Med. Unlocked **18**, 100288 (2020). https://doi.org/10.1016/j.imu.2019.100288
34. Nnamoko, N., Korkontzelos, I.: Efficient treatment of outliers and class imbalance for diabetes prediction. Artif. Intell. Med. **104**, 101815 (2020). https://doi.org/10.1016/j.artmed.2020.101815
35. Ameena, R.R., Ashadevi, B.: Predictive analysis of diabetic women patients using R. Syst. Simul. Model. Cloud Comput. Big Data Appl. (2020). https://doi.org/10.1016/b978-0-12-819779-0.00006-x
36. Tan, F.H., Hor, C.P., Lim, S.L., Tong, C.V., Hong, J.Y., Zain, F.M., Yeow, T.P.: Traditional and emerging cardiometabolic risk profiling among Asian youth with type 2 diabetes: a case-control study. Obes. Med. **18**, 100206 (2020). https://doi.org/10.1016/j.obmed.2020.100206
37. American Diabetes Association: Classification and diagnosis of diabetes: standards of medical care in diabetes—2020. Diabetes Care **43**(Supplement 1), S14–S31 (2020). https://doi.org/10.2337/dc20-s002
38. Heslinga, F.G., Pluim, J.P., Houben, A., Schram, M.T., Henry, R.M., Stehouwer, C.D., Veta, M.: Direct classification of type 2 diabetes from retinal fundus images in a population-based sample from The Maastricht Study. Med. Imaging 2020 Comput. Aided Diagn. (2020). https://doi.org/10.1117/12.2549574
39. Albahli, S.: Type 2 machine learning: an effective hybrid prediction model for early type 2 diabetes detection. J. Med. Imaging Health Inform. **10**(5), 1069–1075 (2020). https://doi.org/10.1166/jmihi.2020.3000

40. Zhu, C., Idemudia, C.U., Feng, W.: Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Inform. Med. Unlocked **17**, 100179 (2019). https://doi.org/10.1016/j.imu.2019.100179
41. Alshamlan, H., Taleb, H. B., Sahow, A. A.: A gene prediction function for type 2 diabetes mellitus using logistic regression. In: 2020 11th International conference on information and communication systems (ICICS). https://doi.org/10.1109/icics49469.2020.239549 (2020)
42. Lukmanto, R.B., Suharjito, N.A., Akbar, H.: Early detection of diabetes mellitus using feature selection and fuzzy support vector machine. Proc. Comput. Sci. **157**, 46–54 (2019). https://doi.org/10.1016/j.procs.2019.08.140
43. Tripathi, D., Manoj, I., Prasanth, G.R., Neeraja, K., Varma, M.K., Reddy, B.R.: Survey on classification and feature selection approaches for disease diagnosis. Emerg. Res. Data Eng. Syst. Comput. Commun. Adv. Intell. Syst. Comput. (2020). https://doi.org/10.1007/978-981-15-0135-7_52
44. Dzulkalnine, M.F., Sallehuddin, R.: Missing data imputation with fuzzy feature selection for diabetes dataset. SN Appl. Sci. (2019). https://doi.org/10.1007/s42452-019-0383-x
45. Zhou, M., Sun, S.D.: GA principle and application. National Defense industry press, Beijing (1999)
46. Mantawy, A., Abdel-Magid, Y., Selim, S.: Integrating genetic algorithms, tabu search, and simulated annealing for the unit commitment problem. IEEE Trans. Power Syst. **14**(3), 829–836 (1999). https://doi.org/10.1109/59.780892
47. Han, X., Dong, Y., Yue, L., Xu, Q.: State transition simulated annealing algorithm for discrete-continuous optimization problems. IEEE Access **7**, 44391–44403 (2019). https://doi.org/10.1109/access.2019.2908961
48. Hughes, G.: On the mean accuracy of statistical pattern recognizers. IEEE Trans. Inf. Theory **14**(1), 55–63 (1968). https://doi.org/10.1109/tit.1968.1054102
49. Abdel-Aal, R.: GMDH-based feature ranking and selection for improved classification of medical data. J. Biomed. Inform. **38**(6), 456–468 (2005). https://doi.org/10.1016/j.jbi.2005.03.003
50. Zaki, M.J., Meira, W., Jr.: Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press (2014)
51. Sun, K., Likhate, S., Vittal, V., Kolluri, V.S., Mandal, S.: An online dynamic security assessment scheme using phasor measurements and decision trees. IEEE Trans. Power Syst. **22**(4), 1935–1943 (2007). https://doi.org/10.1109/tpwrs.2007.908476
52. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002). https://doi.org/10.1613/jair.953
53. Kaur, P., Kaur, R.: Comparative analysis of classification techniques for diagnosis of diabetes. Adv. Intell. Syst. Comput. Adv. Bioinform. Multimedia Electron. Circuits Signals (2019). https://doi.org/10.1007/978-981-15-0339-9_17
54. Hemeida, A.M., Hassan, S.A., Mohamed, A.A.A., Alkhalaf, S., Mahmoud, M.M., Senjyu, T., El-Din, A.B.: Nature-inspired algorithms for feed-forward neural network classifiers: a survey of one decade of research. Ain Shams Eng. J. (2020). https://doi.org/10.1016/j.asej.2020.01.007
55. Hasan, M.K., Alam, M.A., Das, D., Hossain, E., Hasan, M.: Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access **8**, 76516–76531 (2020). https://doi.org/10.1109/ACCESS.2020.2989857
56. Tama, B.A., Rhee, K.: Tree-based classifier ensembles for early detection method of diabetes: an exploratory study. Artif. Intell. Rev. **51**(3), 355–370 (2017). https://doi.org/10.1007/s10462-017-9565-3
57. Rehman, A., Naz, S., Razzak, I.: Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. Multimedia Syst. **21**, 1–33 (2021)
58. Hossain, M.S., Muhammad, G., Alamri, A.: Smart healthcare monitoring: a voice pathology detection paradigm for smart cities. Multimedia Syst. **25**(5), 565–575 (2019)
59. Li, J., Zhang, B., Lu, G., You, J., Zhang, D.: Body surface feature-based multi-modal learning for diabetes mellitus detection. Inf. Sci. **472**, 1–14 (2019)
60. Tama, B.A., Rhee, K.H.: Tree-based classifier ensembles for early detection method of diabetes: an exploratory study. Artif. Intell. Rev. **51**(3), 355–370 (2019)
61. Islam, M.M., Rahman, M.J., Roy, D.C., Maniruzzaman, M.: Automated detection and classification of diabetes disease based on Bangladesh demographic and health survey data, 2011 using machine learning approach. Diabetes Metab. Syndr. **14**(3), 217–219 (2020)
62. Sisodia, D., Sisodia, D.S.: Prediction of diabetes using classification algorithms. Proc. Comput. Sci. **132**, 1578–1585 (2018). https://doi.org/10.1016/j.procs.2018.05.122
63. Larabi-Marie-Sainte, A., Almohaini, R., Saba, T.: Current techniques for diabetes prediction: review and case study. Appl. Sci. **9**(21), 4604 (2019). https://doi.org/10.3390/app9214604

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