PLOS ONE

RESEARCH ARTICLE

Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2

Praveen Talari[1], Bharathiraja N[2], Gaganpreet Kaur[2], Hani Alshahrani[3], Mana Saleh Al Reshan[4,5], Adel Sulaiman[3], Asadullah Shaikh[4]*

1 Department of Computer Science and Engineering, Vignana Bharathi Institute of Technology, Hyderabad, India, **2** Chitkara University Institute of Engineering and Technology, Chitkara University Punjab, Rajpura, India, **3** Department of Computer Science, College of Computer Science and Information Systems, Najran University, Najran, Saudi Arabia, **4** Department of Information Systems, College of Computer Science and Information Systems, Najran University, Najran, Saudi Arabia, **5** Scientific and Engineering Research Centre, Najran University, Najran, Saudi Arabia

* asshaikh@nu.edu.sa

Citation: Talari P, N B, Kaur G, Alshahrani H, Al Reshan MS, Sulaiman A, et al. (2024) Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2. PLoS ONE 19(1): e0292100. https://doi.org/10.1371/journal.pone.0292100

Editor: Suja A. Alex, St Xavier's Catholic College of Engineering, INDIA

Received: June 26, 2023

Accepted: September 12, 2023

Published: January 18, 2024

Data Availability Statement: All relevant data are within the manuscript.

Funding: The authors express their gratitude to the Deanship of Scientific Research and the Center of Scientific and Engineering Research Centre at Najran University for providing financial support for their research through the Research Centers Funding program grant code (NU/RCP/SERC/12/16).

Competing interests: The authors have declared that no competing interests exist.

Abstract

Diabetes prediction is an ongoing study topic in which medical specialists are attempting to forecast the condition with greater precision. Diabetes typically stays lethargic, and on the off chance that patients are determined to have another illness, like harm to the kidney vessels, issues with the retina of the eye, or a heart issue, it can cause metabolic problems and various complexities in the body. Various worldwide learning procedures, including casting a ballot, supporting, and sacking, have been applied in this review. The Engineered Minority Oversampling Procedure (Destroyed), along with the K-overlay cross-approval approach, was utilized to achieve class evening out and approve the discoveries. Pima Indian Diabetes (PID) dataset is accumulated from the UCI Machine Learning (UCI ML) store for this review, and this dataset was picked. A highlighted engineering technique was used to calculate the influence of lifestyle factors. A two-phase classification model has been developed to predict insulin resistance using the Sequential Minimal Optimisation (SMO) and SMOTE approaches together. The SMOTE technique is used to preprocess data in the model's first phase, while SMO classes are used in the second phase. All other categorization techniques were outperformed by bagging decision trees in terms of Misclassification Error rate, Accuracy, Specificity, Precision, Recall, F1 measures, and ROC curve. The model was created using a combined SMOTE and SMO strategy, which achieved 99.07% correction with 0.1 ms of runtime. The suggested system's result is to enhance the classifier's performance in spotting illness early.

1. Introduction

One of the most common diseases that threaten a person's well-being and life was insulin resistance, and also its frequency across the globe was increasing rapidly. Hyperglycemia appears to

be a chronic health problem associated with diabetes mellitus. Pressure, heart disease, kidney problems, blindness, and many other important problems can result from chronic high blood sugar because it can cause persistent damage and the functioning of many tissues and organs [1–3]. The patient's quality of life would deteriorate and he would die earlier as a result of those consequences. Type-1, type-2, and many additional types of clinical diagnostics could be distinguished by WHO classifications [4]. Type 1 diabetes was caused by a blockage of pancreatic production and maybe a complete absence of insulin in the body. The most common type of insulin, type 2 diabetes mellitus, has been caused by inadequate compensatory mechanisms for insulin sensitivity and insulin secretion [5].

The International Diabetes Federation (IDF) recently released information indicating that there were 425 million adult diabetics globally in 2017 compared to 151 million in 2000, especially type 2 diabetes accounting for roughly 90% of cases [6–8]. By 2040, there will be 642 million diabetics worldwide, or one in every ten. To better prevent diabetes & lower diabetes prevalence, diabetes mellitus has thus emerged as a critical worldwide health problem that necessitates prompt diagnosis & treatment [9, 10]. Data mining software enables precise judgment for the diagnosis and treatment of illnesses by knowledge extraction & pattern concealed by illnesses from a significant volume of diagnostic medical information. Predicting and collecting data on diabetes mellitus had become a challenging & important study due to the increasing complexity and scope of medical information [11]. Several weight foundation classifications were combined using the supervised learning approach to create an ensemble of classifiers that works better than a single one [12].

A hybrid insulin forecasting model has been created using random forest (RF) & severe gradients enhancing to enhance the performance of the classifier & diagnostic accuracy [13]. These 2 ensemble learning techniques had been used in several regression or categorization research investigations and also have produced accurate predictions, proving the effectiveness of RF & XGBoost as classifiers design techniques. This section deals with the small number of works that are very closely connected. To predict health, numerous scientific papers had made utilize the Pima Indians Diabetes Collection. Weka tools and algorithms for machine learning have been employed [14].

2. Related works

Researchers' techniques could be divided into four categories: deep learning methods, knowledge discovery, hybrid approach, genetics, or machine learning methods. Insulin diagnosis using deep learning on ECG signals [15]. They particularly used neural networks with convolution and extended selective memory, and subsequently, vector support machines were used to extract features. They discovered a very good precision of 95.7% as an outcome. Data mining methods were used [16] to accurately estimate the probability of a person developing type 2 diabetes in 95.42% of cases [17]. The adjustment was made by empirically selecting the first seed spot as the actual size. By performing 100 tests and choosing the lowest number of the "group error function", the start capital point was discovered. A data mining process called categorization and predictions first depends on the skills data to create a system, which is then used to test information to generate predictions [18]. The discoveries of applying a few characterization calculations to disease data for the finding of long-haul sickness are incredibly reassuring [19].

The creation of a unique classification technique that can hasten and streamline the process of chronic disease diagnosis is necessary. A great deal of medical information is created and modified each day in this age of information expansion [20]. Electronic health records which would include the patient's medical records, findings, prescriptions data, pharmacist data,

client insurance details, and social networking entries like blogs and tweets, were examples of health information [21]. An efficient flow processing process that can manage and evaluate the huge quantities of health records was needed [22]. A main edge screening and order technique was utilized to make the methodology. Apriori and hereditary calculations were utilized to recognize the most huge and solid attributes. Random forest and LS-SVM orders were utilized to evaluate their exhibition [23, 24]. In addition, wavelet transformations were used to differentiate between regular and outliers. The efficiency of LS-SVM using the a priori method was highest, with a reliability of 94.31%, according to a study of the results based on evaluation measurements [25].

The proposed technique could be exceptionally valuable in the early discovery of knee joint sicknesses so that individuals can get clinical therapy at an early age [26]. The creators have partitioned these techniques into three classifications: supervised, semi-supervised, and unsupervised variable selection [27]. In addition, several challenges and obstacles to understanding gene expression information were discussed. Some fundamental challenges raised included lowering the data dimensionality with tens of thousands of characteristics, handling inaccurately labeled & highly unbalanced information, identifying the relevance & repetition among genes, & extrapolating pertinent biological information from gene expressions. A comparative study of feature engineering found that the results of the classification of moderate training and uncontrolled procedures were just as promising as controlled selection [28]. The Naïve Bayes' classifier predictive accuracy has greatly improved with the proposed strategy. The method used was very simple but effective and would undoubtedly make it easier for doctors and health professionals to identify type 2 diabetes [29, 30]. It was correctly determined what would be the perfect limit with the included symmetric uncertainty. Using symmetric uncertainty, the minimum spanning tree was built. Focusing on expectation execution or the extent of change utilized, the results of the proposed calculation were contrasted and those of different calculations like Quick, FCBF, Relief, and CFS, and it was found that ModifiedFAST was the best technique out of every one of them [31].

SVM showed the most promising results for diagnostic decision-making when [32] assessed the various DM & ML systems for the diagnosis of diabetes. The outcomes obtained on a dataset obtained demonstrated the potential of smart SVM for diabetic identification. According to the experimental results, the classifications be able to reach 94% average accuracy, 93% specificity rates, and the corresponding production of 94%. The heuristics and research questions are where evolutionary operators excel in the field of machine learning [33]. When applied to real-world issues that reflect the natural selection method, these methods often provide more specific solutions. The first stage of their suggested three-phase technique had been an attribute selection procedure that was carried out by maintaining an organized list of characteristics that have been maintained in diminishing rank ordering [34–39]. New characteristics were generated in the second stage of applying the method of selecting additional characteristics from each subtype of the characteristics of the original database [40–42]. The tests were performed in the last step using a neighboring k-nearest & SVM classifier [43]. The effectiveness of scalable algorithms has been evaluated as part of the PIMA, and preliminary reports indicate that the proposed methods performed better than other options.

3. Proposed system

3.1 Dataset and K-fold cross-validation

Despite being a non-transmittable sickness, type 2 diabetes has of late achieved the place of a scourge quiet executioner. This perspective on sickness is the consequence of two key elements. In the first place, paying little heed to mature gathering, district, or orientation, a

sluggish yet outstanding ascent in disease predominance has been seen. As far as numerous risks implied, the underlying asymptomatic stage, different short and long-haul results that represent a significant well-being risk, and related co-morbidities, the illness elements are likewise very muddled. Except for some certain risk factors including a family background of diabetes, ethnic inclination, maturing, and others, most of its risk factors are a way of life decisions like deficient actual work, absence of activity, high weight file (BMI), unfortunate food, and smoking. For our exploration, we utilized the Pima Indian Diabetes (PID) dataset, which was acquired from the UCI AI Vault and came from the Public Organization of Diabetes and Stomach Related and Kidney Infections (NIDDK). PID is collected from the UCI ML repository of this research, & the selection of this dataset was made because the majority of individuals nowadays have similar lifestyles, relying mostly on packaged foods & engaging in less physical exercise. People were genetically predisposed to long-term survival on low-carbohydrate diets. But in the last few days, the PIMA group has suddenly changed its regular diet to packaged foods, which has been accompanied by a reduction in physical activity. With a k-value of ten, the K-fold cross-validation approach was applied.

A resampling strategy called cross-validation is utilized to evaluate machine learning models on a little information test. The cycle contains a solitary boundary, k, that assigns the number of gatherings that ought to be made from a given information test. Thus, the cycle is much of the time alluded to as k-overlap cross-approval. At the point when a specific number for k is chosen, it could be filled in for k in the model's reference, for example, when k = 10 is utilized to allude to cross-approval by a 10-crease factor. In applied AI, cross-validation is generally used to measure how well a machine-learning model performs on undeveloped information. That is, to utilize a small example to survey how the model will perform when used to produce expectations on information that was not used during the model's preparation. Ten partitions of equal size were created by randomly dividing the complete data set. One in ten partitions was kept to test the system, and the remaining 10 partitions—minus one—were also used for training examples. Each of the 10 divisions has only been used once as a test dataset in the entire procedure, which has been performed ten times. The summation function was used to combine the results from all repetitions. To match the effectiveness of the training and test data sets, the issue of classification and approval has been reduced within the data set. The advantage of this strategy was that it eliminated the data bias required to develop machine learning models to produce accurate results. The search was conducted using the HP Z60 computer. The system has an Intel XEON 2.4GHz processor and an NVIDIA Quadro-K2200 GPU, according to its technical specs. The system RAM and screen RAM were both 4GB. The Linux system is installed along with Windows Pro 64-bit, and the machine has a 1TB storage capacity.

3.2 ML models and ensembling

To strengthen the analytical capacities of the different frameworks created and address real-world issues, machine learning models are consciously merged. Similar behavior is seen in machine learning predictions. Models work with inputs and generate results. The result is a forecast based on the pattern that the models identify throughout the training phase. For a certain collection of data, no single method will always produce the ideal forecast. It is difficult to create a model with high accuracy using machine learning algorithms since they have constraints. We can increase overall accuracy by creating and combining numerous models. The foundation of bagging is giving an iterative learning process access to the training data. Each model uses a slightly different subset of the training data set to learn the error made by the

prior model. Bagging lessens overfitting and variation. Her random forest algorithm is one instance of such a method. In bagging, many models are often of the same kind of learning and have been developed from many subsamples of the training data set. On the other hand, various variations of a similar type have indeed been constructed using the boost technique, where every other system learns to be correct to the prediction flaws of a preceding model in the chain.

To maximize the different ML/statistical metrics for a better Type 2 diabetes mellitus (T2DM) disease prognosis, several models of different types were created and their predictions were incorporated using the voting technique. Type 2 diabetes mellitus (T2DM) has been reported to be more common among children and adolescents during the past 20 years, particularly in those who are members of underrepresented racial and ethnic groups. Even though T2DM is as yet a generally remarkable condition in youngsters, this pattern, which corresponds with the ascents in pediatric heftiness recurrence and seriousness, has raised serious worries. Since youth T2DM seems, by all accounts, to be a forceful sickness with quickly advancing cell decline, a high therapy disappointment rate, and speed improvement of confusion, it changes not just from type 1 diabetes in kids, from which it can some of the time be hard to recognize yet in addition from T2DM in grown-ups.

The training packages were created by randomly replicating the original data. After developing various training data sets, several models were applied to the ensemble structured sampling procedure. The final projection is done after having cumulated all the results of the students. To combat the problem of overflow, it is advantageous to reduce modeling differences during learning. Initialization, Concurrent Training, and Aggregation are the three basic procedures employed in the bagging approach. Multiple base models are separately and concurrently trained on various subsets of the training data while using bagging (also known as Bootstrap aggregating), a sort of ensemble learning. Using bootstrap sampling, which involves selecting data points at random and replacing them, each subset is produced. Using majority voting, the Bagging classifier aggregates the all-base model's predictions to arrive at its conclusion. Bagging regression is the process of making the final prediction in a regression analysis by averaging the results from all of the base models.

Bagging, also known as Bootstrap Aggregation, is a machine-learning ensemble technique. For each ensemble member, a bootstrap sample of the training dataset is created, a decision tree model is trained on each sample, and the predictions are then directly merged using a statistic like the average of the predictions. The bootstrap methodology effectively accomplishes the desired outcome of rendering each sample within the dataset highly distinct, or at the very least, reasonably diverse, to construct an ensemble. Subsequently, each data sample is subjected to a decision tree fitting process. Due to the variances present within the training dataset, each tree will exhibit some degree of dissimilarity. Typically, the decision tree is configured to possess a greater depth or to forego pruning, which may augment the specialization of each tree to the training dataset and, as a result, the discrepancies among the trees. The "diversity" of the ensemble will be increased by differences in the trees, which will result in ensemble members with lower correlations in their predictions or prediction errors. It is widely acknowledged that groups with skilled and varied members—those who are skilled in a variety of ways or make a variety of mistakes—perform better. The bagging tree approach employs several methods, such as the structure of bagging trees, random forests, additional trees, etc. The bagging product's algorithm (Algorithm 1) & equation is given as follows:

$$B_{bag} = \sum_{x=1}^{n} B_x(I) \tag{1}$$

```
Algorithm 1: Bagging Procedure
Input: Lifestyle Dataset
Output: Prediction of T2DM
Start
Step 1: Dataset import
Step 2: Dataset preprocessing
{
Step 2.1: Integration of data
Step 2.2: Transformation of data
Step 2.3: Cleaning of data
}
Step 3: Train the dataset with 70% (X and Y) axis
Step 4: Test the dataset with a 30% (X and Y) axis
Step 5: Ensemble Learning Methods Algorithms
mn = Ensemble Learning (EL) algorithms for x ranges from 0 to 8 do
{
EL = mn[x];
EL.fit ();
EL.predict ();
printf (Performance measures);
}
Step 6: Framework Deployment
End
```

AdaBoost, XGBoost, RF, Gradient Boost, and other boosting techniques were only ever a few example. The basic enhancing algorithm is explained in Algorithm 2

Algorithm 2: Boosting procedure

Step 1: Estimator Fitness E^{x}

Step 2: Weal estimators for x in [1, C] // x : no. of iterations

Step 3: $\text{Loss}^{y} = \sum_{y=1}^{n} (J_y - E^x(I_y))^2$ // loss in x^{th} iterations (2)

Step 4: Gradient $-\frac{\partial Lx}{\partial l_y} = -\frac{2}{n} * \left((J_y - E^x(I_y)) Jx \right)$ (3)

Step 5: Weak estimator fitness: $H^x on\left(I, \frac{\partial L}{\partial l}\right)$ (4) // ρ changes the step size

Step 6: Forecast: $E^m(I) = E^x(I) + \rho * H^x(I) = E^1 + \rho * \sum_{x=1}^{m} H^x(I)$ (5)

3.3 Data pre-processing

The procedure used on the dataset before processing is called pre-processing. Pre-processing typically alters the raw data, which can improve the processing's capacity to classify data. Pre-processing usually modifies the raw data, which may improve the processing's ability to classify the data. It also functions the functions of standardization, integration, extraction, aggregation, & discretization of properties. Datasets with many cases are believed to have been evaluated using the method we suggested. During the course of data mining and analysis, a crucial stage known as data preprocessing is undertaken, wherein unprocessed data is transformed into a structured format that can be comprehended and analyzed by computers and machine learning algorithms as shown in Fig 1. The "features" that makeup data sets can be used to describe or convey them. By size, location, age, time, color, and other factors, for example. Features, sometimes referred to as attributes, variables, fields, and characteristics, are represented in datasets as columns.

There are usually more occurrences of negative than positive classes in the data set. Moreover, the fact that this group was larger was encouraging because it mirrored the disease category. Furthermore, since there would not be enough data for the classifier to form the minority class, giving it unbalanced data would direct it towards the majority class. This bias would have a greater impact on the classifier, resulting in better results for the majority class and worse results for the minority class. The issue of data imbalance seems to be very

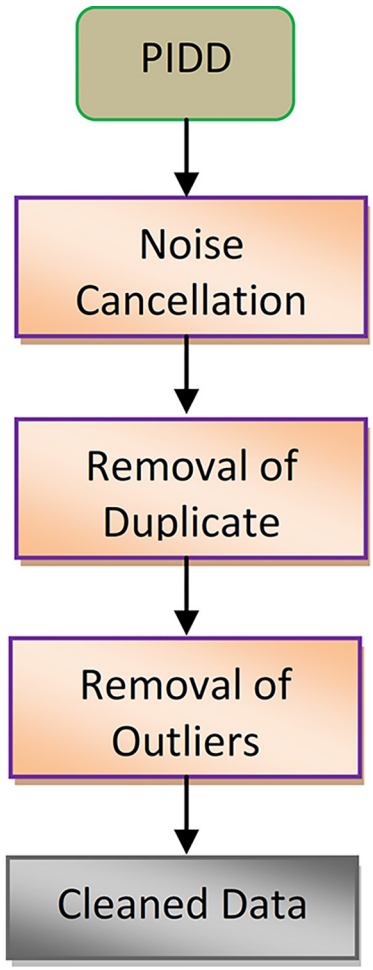


Fig 1. Steps for traditional pre-processing of data.

https://doi.org/10.1371/journal.pone.0292100.g001

prevalent. This could be rectified using resampling techniques. These solutions, however, address the issue of the minority class by also decreasing an instance for the class of the majority or raising the examples of the minority class by using repeating groups. But this leads to reduce chances of higher precision and data loss. The SMOTE approach was utilized to analyze the negative occurrences of the SMOTE dataset, which dominated the positive occurrences. Many medical academics use SMOTE because it has become the most promising of the various oversampling techniques that are now available.

3.4 SMOTE and Sequential Minimal Optimization (SMO)

The Sequential small Optimisation (SMO) algorithm is developed by extending the concept of the decomposition technique to its maximum potential and optimizing a small subset of only two points at each iteration. The efficacy of this approach is attributed to the analytical resolution of the optimization problem for two data points, thereby eliminating the need for an iterative quadratic programming optimizer as a component of the method. To predict diabetes, this investigation integrates the synthetic minority oversampling technique (SMOTE) and sequential minimal optimization (SMO) algorithms. The SMOTE method is used in the first phase of this suggested two-phase classification model to pre-process the data, and the SMO

classifier is used in the second phase. SMO receives the pre-processing output to improve the classifier's performance. The Sequential Minimal Optimisation (or SMO) learning technique for SVMs is a recent development. The utilization of an analytical quadratic programming (QP) phase instead of the numerical quadratic programming (QP) inner loop employed by earlier support vector machine (SVM) learning algorithms is a distinguishing feature of Sequential Minimal Optimization (SMO). The SVM QP problem can be expeditiously solved through SMO, which is a simple approach that does not necessitate any supplementary matrix storage or numerical QP optimization stages. By leveraging Osuna's theorem to ensure convergence, SMO decomposes the primary QP problem into smaller QP subproblems. In contrast to the earlier techniques, SMO decides to resolve the lowest optimization challenge at each stage.

SMOTE is an oversampling method that has been proposed to overcome the unbalanced class issue of the dataset. Linking minority class values to parallel lines and adding false points to those lines improves classifier effectiveness. As with conventional oversampling, SMOTE generates new examples by synthesizing and recreating minority class data. It differs from the conventional approach in that it considers the minority class instance at its nearest vector in feature space as opposed to data space. Two methods can be used to generate the new synthesis parameters: the oversample rate method and the nearest number of neighbors method. SMO is the name of the SVM classifier's modified design. John Plat developed the SMO technique in late 1998 to address the quadratic programming problem that had been brought up during the SVM training process. Without employing additional matrix storage or QP numeric calculation, it resolves the QP problem. To guarantee convergence, SMO breaks the QP issue down into several sub-issues and chooses the smallest. Compared to other classifiers, this statistical one may be more computationally efficient and have a lower average error rate.

3.5 Proposed architecture

One approach that is more effective than traditional QP solvers for solving the SVM training issue is sequential minimal optimization (SMO). SMO divides the training challenge into smaller problems that may be resolved analytically using heuristics. The working set selection heuristics' underlying assumptions heavily influence how effectively they perform. Usually, it significantly shortens the training period. One of the most widely used oversampling techniques for addressing the issue of class imbalance is known as SMOTE (synthetic minority oversampling technique). This method aims to balance the distribution of classes by randomly increasing the number of minority class samples and duplicating them. SMOTE generates new minority instances by combining existing minority instances. To create virtual training records for the minority class, SMOTE employs linear interpolation. Specifically, for each example in the minority class, one or more of its k-nearest neighbors are randomly selected to serve as synthetic training records. Once the oversampling process is complete, the data is reconstructed and can be subjected to various classification models. A proposed method combines SMOTE and SMO algorithms, which preprocess unbalanced data before the SMO grader improves its performance. The treatment strategy can be looked at in Fig 2.

Our proposed approach has been validated through the examination of testing data from the Pima Indians Diabetes Database (PIDD). Performance and accountability reporting (PAR) is a process that involves the collection and retention of data that measures an organization's achievements, efficiency, and adherence to budgetary constraints, while also comparing actual results with previously established goals. The effectiveness of PAR was assessed using a variety of performance indicators, including accuracy, Recall, accuracy, and F measurement. The SMOTE and SMO algorithms were used in an integrated manner to attain an accuracy rate of

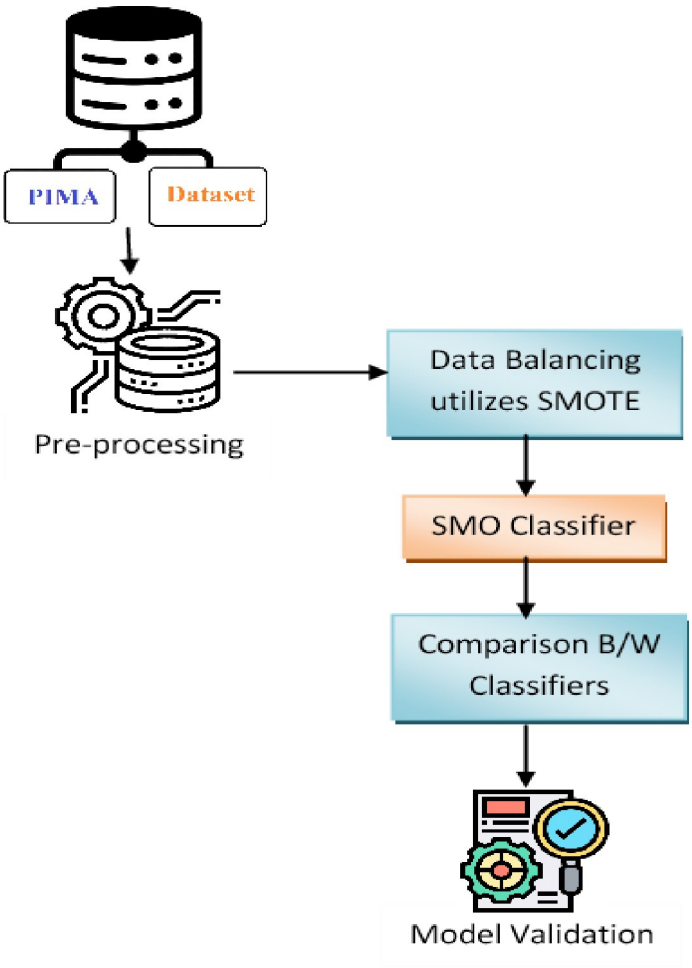


Fig 2. Proposed architecture for diagnosing diabetes.

https://doi.org/10.1371/journal.pone.0292100.g002

99.07%. After that, it had to be compared to Packing, boosting, and Voting. The model we propose may help a doctor make better choices based on the characteristics that are extracted. Many scientists have executed ML on PIDD using various features. Table 1 lists some of the earlier research with the suggested approaches and accuracy rates.

4. Results

Experimental design results utilizing the ML/EL approach of the T2DM assumption based on the lifestyle predictors were investigated and reported. The HP Z60 computer used as part of the search was used. Technical specifications for the equipment include an Intel XEON 2.4GHz processor and an NVIDIA Quadro-K2200 GPU. Both screen RAM & the system

Table 1. A comparative analysis of earlier studies.

Reference number	Applied Methods	Accuracy Rate (%)
[36]	a hybrid method of cuckoo and firefly searching	82
[37]	Feedforward NN	84
[18]	NB algorithm	79.64
[24]	LDA, MWSVM	89.97

https://doi.org/10.1371/journal.pone.0292100.t001

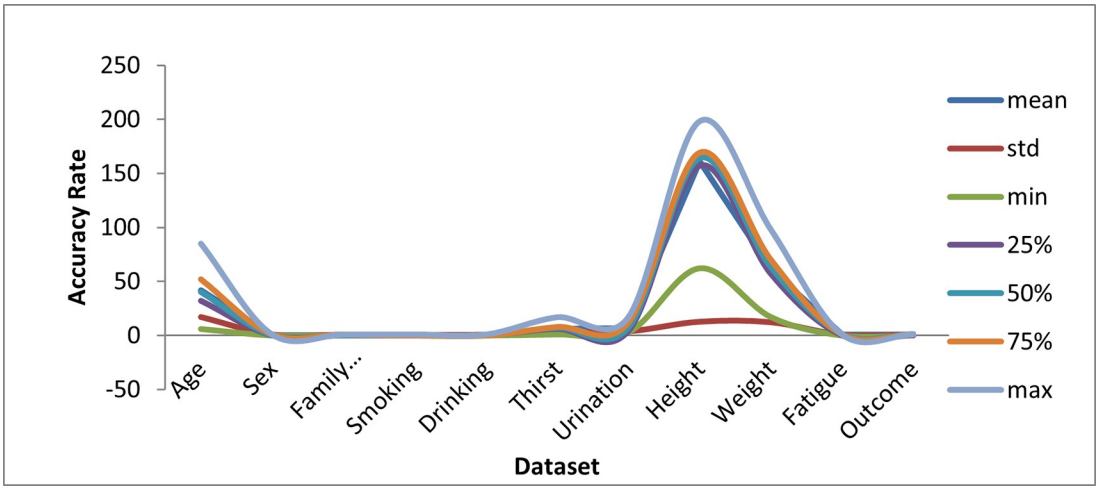


Fig 3. Description of parameters used in the dataset.

https://doi.org/10.1371/journal.pone.0292100.g003

RAM were 4GB every. Windows pro-64 bit is the Linux distribution loaded, as well as the storage capacity of the machine, is 1TB. Fig 3 shows basic descriptive statistics of lifestyle characteristics as well as their measurements, such as averages, ICD, min, max, etc.

$$CCA = \frac{kaverage(corr_{fc})}{\sqrt{k + k(k-1)}average(corr_{ff})} \tag{6}$$

The correlation matrix is a formal representation of the correlation between variables. It presents the correlation between all possible pairings of values in a matrix format. As depicted in Fig 4, a correlation matrix is utilized to summarize a large dataset, identify patterns, and make informed decisions based on the results. This matrix enables us to determine the degree of correlation between variables and visualize the outcomes. The correlation matrix is a table with rows and columns that display the variables, and each cell in the matrix contains the correlation coefficient. It is commonly used in conjunction with other types of statistical analysis and is particularly useful in regression techniques such as simple linear regression, multiple linear regression, and lasso regression models.

4.1 Confusion matrix

The assessment of the efficacy of ML/EL models in detecting labeling errors/errors in T2DM disease prediction is conducted through the utilization of the confusion matrix presented in Table 2. This matrix evaluates the veracity of the outcomes by comparing them with the expected values across four key components, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Fig 5 displays the confounding matrices used by ensemble learning classifiers to assess their performance in predicting T2DM illness. The several EL/ML classifiers were evaluated utilizing confounding matrices using ML/statistical metrics such as ROC curve, accuracy, recall, f1 score, specificity, mistake classification rate, etc.

4.2 Performance evaluation

Table 3 shows the compilation/summation of results obtained utilizing a 10-prong cross-validation technique for multiple ML/EL models. The results of various measurement measures,

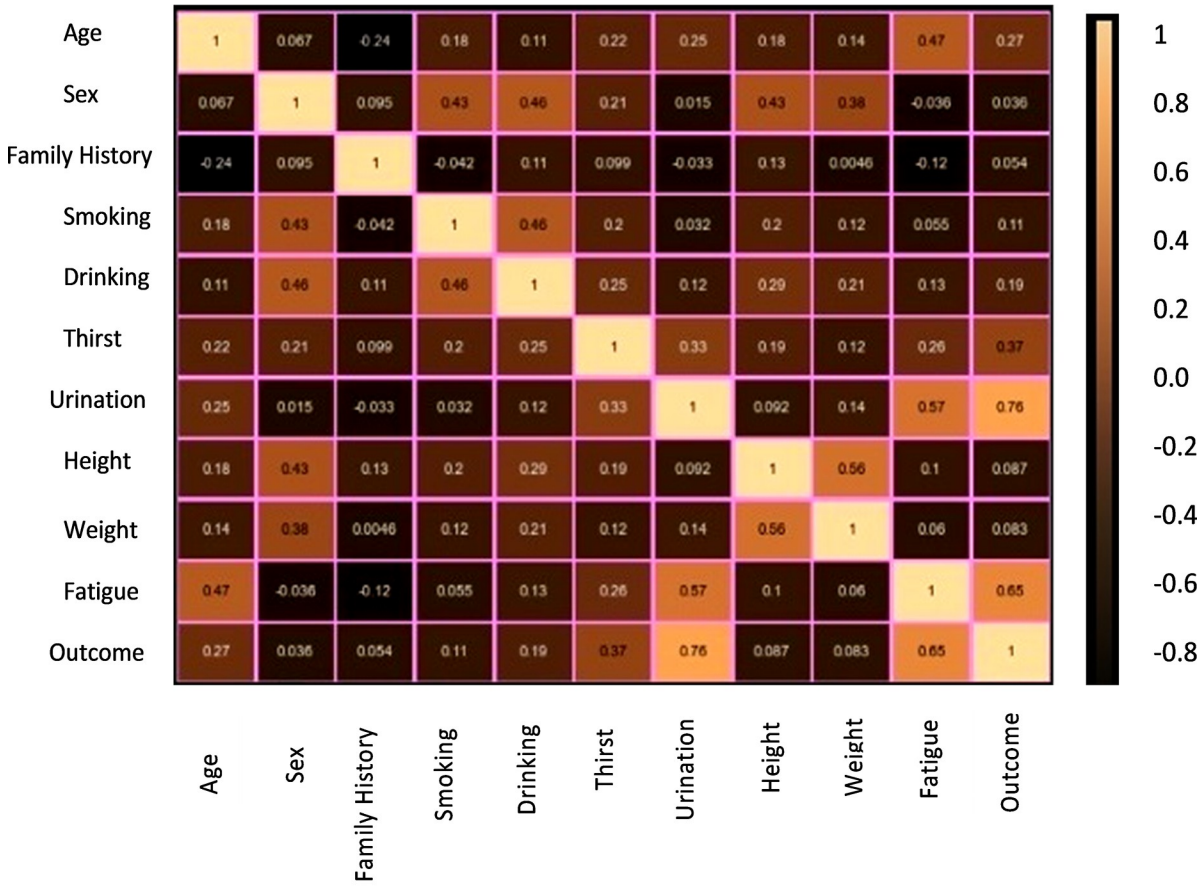


Fig 4. Correlation coefficient matrix.

https://doi.org/10.1371/journal.pone.0292100.g004

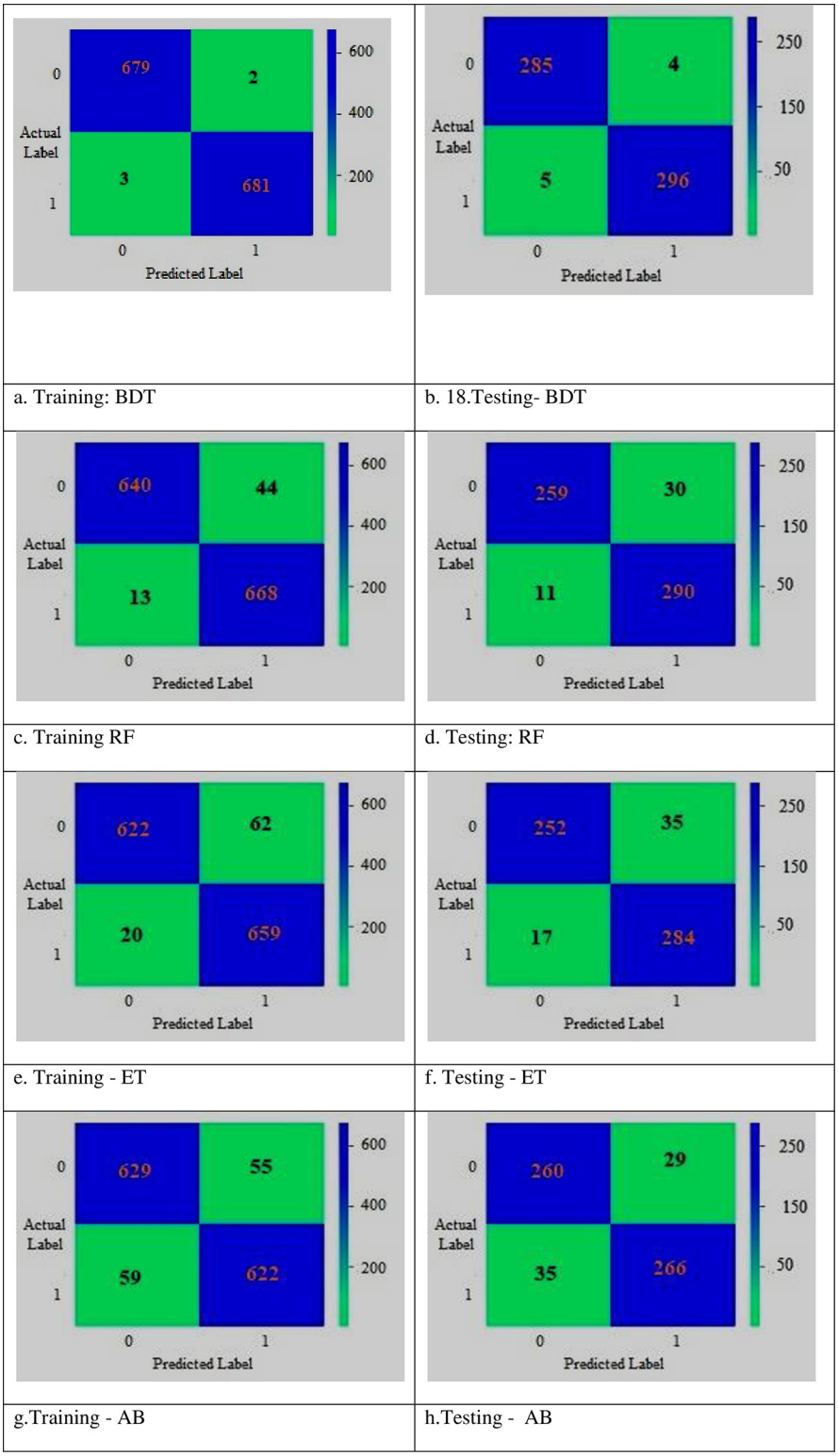
including test accuracy, training accuracy, Kappa, breach classification rate, and operating time, are described. A boosting decision tree (BDT) method outperformed all other models, with a 99.14% testing accuracy rate, followed by the ET, RF, SGB, AB, & Voting classifiers, which reached 98.45%, 93.63%, & 91.41%, respectively. Fig 5(F) and 5(G) shows the ET Testing Confusion Matrix of Uncertainty in AB Instruction. Fig 5(H) Matrix for AB Testing Trouble, both 89.69% & 89.51% were acceptable. But when it comes to BDT received the lowest score at 0.86% & Vote scored the best at 10.49%. BDT had the highest efficiency ratio in the sigma data analysis (98.17%), whereas ET had the lowest rate (84.60%). Additionally, when algorithms are being executed, AB requires a minimal runtime of 0.0330 s as well as the Vote technique has a max running time of 0.0990sec.

Other statistical/ML parameters for test data are shown in Table 4 and include accuracy, memorization, specific alarm rate, actual alarm rate, negative predictive value, F1 measurement, etc. However, BDT attained a desirable performance rate of 99.32%, 98.95%, 98.98%,

Table 2. Confusion matrix.

	Values Predicted		
		No	**Yes**
Values Actual	No	TN	TP
	Yes	FN	FP

https://doi.org/10.1371/journal.pone.0292100.t002



a. Training: BDT | b. 18.Testing- BDT

c. Training RF | d. Testing: RF

e. Training - ET | f. Testing - ET

g.Training - AB | h.Testing - AB

Fig 5. Confusion matrix. a. Training: BDT, b. 18.Testing- BDT, c. Training RF, d. Testing: RF, e. Training–ET, f. Testing–ET, g.Training–AB, h.Testing—AB.

https://doi.org/10.1371/journal.pone.0292100.g005

Table 3. Performance measure.

Methods	Accuracy (training)	Accuracy (testing)	Kappa	MCR (%)	RT (s)
WFS boost Algorithm	99.71	99.16	98.19	0.87	0.0740
Random Forest	96.12	93.66	87.69	6.38	0.0642
Extra Trees	94.13	91.42	84.72	8.63	0.0600
AdaBoost	91.92	89.73	86.72	10.32	0.0330
Stochastic	99.57	98.47	91.12	1.57	0.0342

https://doi.org/10.1371/journal.pone.0292100.t003

1.04%, 0.67%, 99.29%, and 99.15% in regards to precision, memory, specificity, FPR, FNR, and F1-score. Additionally, the voting technique had the poorest accuracy and efficiency of 86.86%. The ET technique had the highest recall rate, coming in at 88.95%. BDT had the highest negative predictive value (NPV) of 99.29%, whereas ET & RF obtained a minimum NPV value of 87.71% and a maximum NPV rate of 90.17%.

BDT's NPV was the highest at 99.29%, while ET & RF obtained a minimum NPV value of 87.71% and a maximum NPV ratio of 90.17%. Irrelevant or inappropriate functionality can interfere with the way a model works. The training time is reduced and accuracy is increased with careful feature selection. Embedded, filtered, wrapping, embedded, and hybrid methods are a few of the feature selection methods used in DL paradigms. The selection of characteristics in this work was performed using information gain and correlation techniques. It has been discovered that, except "Gender," practically all of the chosen features have made a significant contribution to the prediction of T2DM showcasing using the Bagging Decision Tree classifier in Fig 6. Urine, obesity, hunger, fatigue, family history, smoking, alcohol, height, and age are the features that are ranked in order of importance in terms of outcomes. Although the sex factor does not influence the outcome class, it has a strong correlation with the independent variables and is an important lifestyle factor.

4.3 Discussion

The attributes that are prioritized based on how they affect results are urine, obesity, hunger, exhaustion, family history, drinking, smoking, height, and age. While the sex factor does not affect the outcome class, it has a significant link with the independent factors and affects lifestyle choices. the collection/summarization of findings from several ML/EL models using a 10-prong cross-validation approach. The findings from several measuring techniques, including test accuracy, training accuracy, Kappa, breach categorization rate, and operation duration, are reported. The ET, RF, SGB, AB, and Voting classifiers came in second with testing accuracy rates of 98.45%, 93.63%, and 91.41%, respectively, while the boosting decision tree (BDT) approach fared the best of all the models. The employment of the confusion matrix is a common practice in the validation of the efficacy of machine learning and ensemble learning

Table 4. Classification performance measurements of the test dataset.

Method	Results (%)						
	Precision	Recall	Specificity	FPR	FNR	F1-Score	NPV
BDT	98.97	99.33	98.96	1.07	0.68	99.32	99.32
RF	96.95	91.15	96.63	3.40	8.88	93.97	90.18
ET	94.93	88.97	94.35	5.68	11.05	91.88	87.72
AdaBoost	88.87	90.70	88.66	12.02	9.28	89.81	90.53
Stochastic Gradient Boosting	98.97	97.98	98.95	1.07	2.11	98.51	97.92
Voting	86.85	91.15	87.11	12.95	7.66	90.02	92.25

https://doi.org/10.1371/journal.pone.0292100.t004

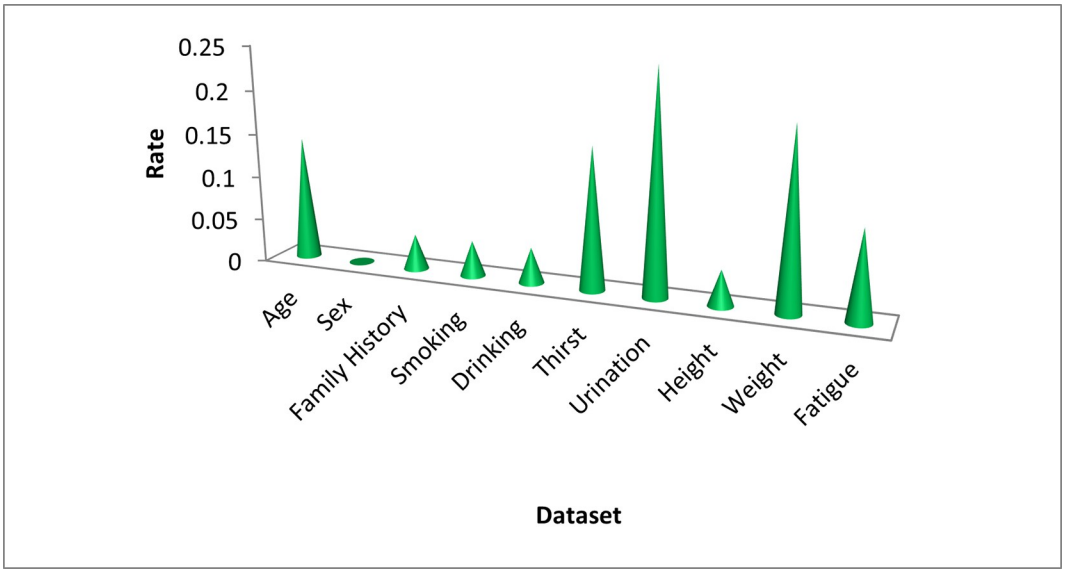


Fig 6. Feature importance towards prediction of T2DM.

https://doi.org/10.1371/journal.pone.0292100.g006

models in detecting labeling errors and errors in the prediction of Type 2 Diabetes Mellitus (T2DM). This matrix facilitates the comparison of the actual outcomes with the anticipated values through the utilization of four key components, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The confounding matrices employed by ensemble learning classifiers to test their efficacy in predicting T2DM disease are shown in Fig 5. ROC curve, accuracy, recall, f1 score, specificity, error classification rate, and other ML/statistical metrics were used to evaluate the various EL/ML classifiers. The recall rate for the ET method was the highest, coming in at 88.95%. With a negative predictive value (NPV) of 99.29%, BDT had the greatest NPV, whereas ET & RF had a minimum NPV of 87.71% and a maximum NPV of 90.17%.

5. Conclusions and future work

Diabetes affects millions of people worldwide and is becoming worse. T2DM greatly improved its understanding of biological and lifestyle factors by using ML/EL methods. The framework was developed after a careful analysis of the customer lifestyle data. To analyze the data and summarize the relevant insights, exploratory data analysis (EDA) is used. This study aims to impart a foundational understanding of the data under consideration, encompassing its distribution, null values, and other pertinent characteristics. The research methodology employed a diverse range of global learning techniques, such as voting, boosting, and bagging. To ensure class balance and validate the data, the Synthetic Minority Oversampling Technique (SMOTE) was utilized in conjunction with the K-fold cross-validation approach. The Pima Indian Diabetes (PID) dataset was sourced from the UCI machine learning (UCI ML) repository for this investigation.

The EDA stage involved patching in null values, and finding and deleting outliers since class balance is a big problem and affects the prediction model by raising the level of quality control for a set of data. CCA was utilized to choose the ideal medley of lifestyle components. Finally, 10 cross-validations were used in conjunction with several set-based machine-learning approaches to predict sickness. A combined SMOTE and SMO techniques were used to build the model, resulting in a 99.07% correction rate with 0.1 ms of runtime. The ET, RF, SGB, AB,

and Voting classifiers came in second with testing accuracy rates of 98.45%, 93.63%, and 91.41%, respectively, while the boosting decision tree (BDT) approach fared the best of all the models. The accuracy, sensitivity measurement, and specificity of the proposed system were some of the assessment criteria that were used to determine its efficacy. To our knowledge, this is the first instance in which a framework has generated predicted results that are appreciably better than those of earlier research projects. Metric analysis can be used to show that the intended work's outcomes are accurate and useful. Future studies may be conducted to increase classification accuracy using the picture dataset by combining feature selection techniques with deep learning algorithms.

Author Contributions

Conceptualization: Praveen Talari.

Data curation: Praveen Talari, Bharathiraja N, Gaganpreet Kaur.

Formal analysis: Praveen Talari, Bharathiraja N, Gaganpreet Kaur.

Investigation: Praveen Talari, Bharathiraja N, Gaganpreet Kaur.

Methodology: Hani Alshahrani, Mana Saleh Al Reshan, Adel Sulaiman, Asadullah Shaikh.

Project administration: Hani Alshahrani, Mana Saleh Al Reshan, Adel Sulaiman, Asadullah Shaikh.

Resources: Hani Alshahrani, Mana Saleh Al Reshan, Adel Sulaiman, Asadullah Shaikh.

Validation: Hani Alshahrani, Mana Saleh Al Reshan, Adel Sulaiman.

Writing – original draft: Praveen Talari, Bharathiraja N, Gaganpreet Kaur, Asadullah Shaikh.

Writing – review & editing: Hani Alshahrani, Mana Saleh Al Reshan, Adel Sulaiman.

References

1. Sneha N., & Gangil T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*, 6(1), 1–19.
2. Tigga N. P., & Garg S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, 167, 706–716.
3. Maleki N., Zeinali Y., & Niaki S. T. A. (2021). A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. *Expert Systems with Applications*, 164, 113981.
4. Haq A. U., Li J. P., Khan J., Memon M. H., Nazir S., Ahmad S.,. . . & Ali A. (2020). Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data. *Sensors*, 20 (9), 2649. https://doi.org/10.3390/s20092649 PMID: 32384737
5. Carter J. A., Long C. S., Smith B. P., Smith T. L., & Donati G. L. (2019). Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes. *Expert Systems with Applications*, 115, 245–255.
6. Maniruzzaman M., Rahman M., Al-MehediHasan M., Suri H. S., Abedin M., El-Baz A., & Suri J. S. (2018). Accurate diabetes risk stratification using machine learning: role of missing value and outliers. *Journal of medical systems*, 42(5), 1–17. https://doi.org/10.1007/s10916-018-0940-7 PMID: 29637403
7. Al-Behadili H. N. K., & Ku-Mahamud K. R. (2021). A fuzzy unordered rule using greedy hill climbing feature selection method: An application to diabetes classification. *Journal of Information and Communication Technology*, 20(3), 391–422.
8. Ijaz M. F., Alfian G., Syafrudin M., & Rhee J. (2018). Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over-sampling technique (SMOTE), and random forest. *Applied Sciences*, 8(8), 1325.
9. Azad C., Bhushan B., Sharma R., Shankar A., Singh K. K., & Khamparia A. (2022). Prediction model using SMOTE, genetic algorithm, and decision tree (PMSGD) for classification of diabetes mellitus. *Multimedia Systems*, 28(4), 1289–1307.

10. Maniruzzaman M., Rahman M., Ahammed B., & Abedin M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health information science and systems*, 8(1), 1–14.

11. Samant P., & Agarwal R. (2018). Machine learning techniques for medical diagnosis of diabetes using iris images. *Computer methods and programs in biomedicine*, 157, 121–128. https://doi.org/10.1016/j.cmpb.2018.01.004 PMID: 29477420

12. Chatrati S. P., Hossain G., Goyal A., Bhan A., Bhattacharya S., Gaurav D., & Tiwari S. M. (2020). Smart home health monitoring system for predicting type 2 diabetes and hypertension. *Journal of King Saud University-Computer and Information Sciences*.

13. Nguyen B. P., Pham H. N., Tran H., Nghiem N., Nguyen Q. H., Do T. T.,. . . & Simpson C. R. (2019). Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Computer methods and programs in biomedicine*, 182, 105055. https://doi.org/10.1016/j.cmpb.2019.105055 PMID: 31505379

14. Pradeepa K., Bharathiraja N., Meenakshi D., Hariharan S., Kathiravan M., & Kumar V. (2022, December). Artificial Neural Networks in Healthcare for Augmented Reality. In 2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP) (pp. 1–5). IEEE.

15. Larabi-Marie-Sainte S., Aburahmah L., Almohaini R., & Saba T. (2019). Current techniques for diabetes prediction: review and case study. *Applied Sciences*, 9(21), 4604.

16. Chen P., & Pan C. (2018). The diabetes classification model is based on boosting algorithms. *BMC Bioinformatics*, 19(1), 1–9.

17. Manikandan K. (2019). Diagnosis of diabetes diseases using an optimized fuzzy rule set by grey wolf optimization. *Pattern Recognition Letters*, 125, 432–438.

18. López B., Torrent-Fontbona F., Viñas R., & Fernández-Real J. M. (2018). Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction. *Artificial intelligence in medicine*, 85, 43–49. https://doi.org/10.1016/j.artmed.2017.09.005 PMID: 28943335

19. Zou Q., Qu K., Luo Y., Yin D., Ju Y., & Tang H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*, 9, 515. https://doi.org/10.3389/fgene.2018.00515 PMID: 30459809

20. Chen R. C., Dewi C., Huang S. W., & Caraka R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7(1), 1–26.

21. Naz H., & Ahuja S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders*, 19(1), 391–403. https://doi.org/10.1007/s40200-020-00520-5 PMID: 32550190

22. Marappan R., Vardhini P. H., Kaur G., Murugesan S., Kathiravan M., Bharathiraja N., & Venkatesan R. (2023). Efficient evolutionary modeling in solving maximization of lifetime of wireless sensor healthcare networks. Soft Computing, 1–15.

23. Gadekallu T. R., Khare N., Bhattacharya S., Singh S., Maddikunta P. K. R., & Srivastava G. (2020). Deep neural networks to predict diabetic retinopathy. *Journal of Ambient Intelligence and Humanized Computing*, 1–14.

24. Kaur H., & Kumari V. (2020). Predictive modeling and analytics for diabetes using a machine learning approach. *Applied computing and informatics*.

25. Tao Z., Huiling L., Wenwen W., & Xia Y. (2019). GA-SVM-based feature selection and parameter optimization in hospitalization expense modeling. *Applied soft computing*, 75, 323–332.

26. Gárate-Escamila A. K., El Hassani A. H., & Andrès E. (2020). Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*, 19, 100330.

27. Mishra S., Tripathy H. K., Mallick P. K., Bhoi A. K., & Barsocchi P. (2020). EAGA-MLP—an enhanced and adaptive hybrid classification model for diabetes diagnosis. *Sensors*, 20(14), 4036. https://doi.org/10.3390/s20144036 PMID: 32698547

28. Yahyaoui A., Jamil A., Rasheed J., & Yesiltepe M. (2019, November). A decision support system for diabetes prediction using machine learning and deep learning techniques. In *2019 1st International Informatics and Software Engineering Conference (UBMYK)* (pp. 1–4). IEEE.

29. Cui S., Wang D., Wang Y., Yu P. W., & Jin Y. (2018). An improved support vector machine-based diabetic readmission prediction. *Computer methods and programs in biomedicine*, 166, 123–135. https://doi.org/10.1016/j.cmpb.2018.10.012 PMID: 30415712

30. Khourdifi Y., & Bahaj M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International Journal of Intelligent Engineering and Systems*, 12(1), 242–252.

31. Bashir S., Khan Z. S., Khan F. H., Anjum A., & Bashir K. (2019, January). Improving heart disease prediction using feature selection approaches. In *2019 16th international bhurban conference on applied sciences and technology (IBCAST)* (pp. 619–623). IEEE.

32. Alam T. M., Iqbal M. A., Ali Y., Wahab A., Ijaz S., Baig T. I.,. . . & Abbas Z. (2019). A model for early prediction of diabetes. *Informatics in Medicine Unlocked*, 16, 100204.

33. Mishra S., Mallick P. K., Tripathy H. K., Bhoi A. K., & González-Briones A. (2020). Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier. *Applied Sciences*, 10(22), 8137.

34. Islam M. M., Ferdousi R., Rahman S., & Bushra H. Y. (2020). Likelihood prediction of diabetes at an early stage using data mining techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis* (pp. 113–125). Springer, Singapore.

35. Doumatey A. P., Adeyemo A., Zhou J., Lei L., Adebamowo S. N., Adebamowo C., & Rotimi C. N. (2020). Gut microbiome profiles are associated with type 2 diabetes in urban Africans. *Frontiers in cellular and infection microbiology*, 63. https://doi.org/10.3389/fcimb.2020.00063 PMID: 32158702

36. Kannadasan K., Edla D. R.,& Kuppili V. (2019). Type 2 diabetes data classification using stacked autoencoders in deep neural networks. *Clinical Epidemiology and Global Health*, 7(4), 530–535.

37. Sharma M., & Kaur P. (2021). A comprehensive analysis of nature-inspired meta-heuristic techniques for feature selection problems. *Archives of Computational Methods in Engineering*, 28(3), 1103–1127.

38. Thiruneelakandan A., Kaur Gaganpreet, Vadnala Geetha, Bharathiraja N, Pradeepa K, Mervin Retnadhas, Measurement of oxygen content in water with purity through soft sensor model, *Measurement: Sensors*, vol. 24, 2022, 100589, ISSN 2665-9174, https://doi.org/10.1016/j.measen.2022.100589.

39. Kaur G., Prabha C., Chhabra D., Kaur N., Veeramanickam M. R. M. and Gill S. K., "A Systematic Approach to Machine Learning for Cancer Classification," 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, 2022, pp. 134–138, https://doi.org/10.1109/IC3I56241.2022.10072474

40. Kaur B., Kaur G. (2023). Heart Disease Prediction Using Modified Machine Learning Algorithm. In: Gupta D., Khanna A., Bhattacharyya S., Hassanien A.E., Anand S., Jaiswal A. (eds) International Conference on Innovative Computing and Communications. Lecture Notes in Networks and Systems, vol 473. Springer, Singapore. https://doi.org/10.1007/978-981-19-2821-5_16.

41. Kaur H., Kaur G., and Pannu H. S. Novel similarity measure-based random forest for fingerprint recognition using dual-tree complex wavelet transform and ring projection. Mod. Phys. Lett. B, vol. 34, no. 02, p. 2050022, Jan. 2020, https://doi.org/10.1142/S0217984920500220

42. Lilhore U.K., Simaiya S., Pandey H., Gautam V., Garg A., Ghosh P. (2022). Breast Cancer Detection in the IoT Cloud-based Healthcare Environment Using Fuzzy Cluster Segmentation and SVM Classifier. In: Hu YC., Tiwari S., Trivedi M.C., Mishra K.K. (eds) Ambient Communications and Computer Systems. Lecture Notes in Networks and Systems, vol 356. Springer, Singapore. https://doi.org/10.1007/978-981-16-7952-0_16.

43. Trivedi N. K., Gautam V., Sharma H., Anand A., and Agarwal S., "Diabetes Prediction using Different Machine Learning Techniques," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 2173–2177, https://doi.org/10.1109/ICACITE53722.2022.9823640