Healthcare Analytics 2 (2022) 100118

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators ☆

Victor Chang [a,*], Meghana Ashok Ganatra [b], Karl Hall [b], Lewis Golightly [a], Qianwen Ariel Xu [a]

[a] Department of Operations and Information Management, Aston Business School, Aston University, Birmingham, UK
[b] School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK

ARTICLE INFO

Keywords:
Diabetes
Diabetic analysis
Machine learning
Healthcare analytics

ABSTRACT

Breakthroughs in healthcare analytics can help both the doctor and the patient. Analytics in healthcare can help spot and diagnose diseases early on. Therefore, they can also be used to improve healthcare quality and patient outcomes. Machine learning models can be used to find patterns in data and generate predictions based on these patterns. They are employed in healthcare applications for disease diagnosis, prognosis, and treatment. With the development of new algorithms and other technological innovations, these models have become more effective than ever at delivering patient treatment. The primary objective of this research is to apply different machine learning algorithms to predict the diagnosis of diabetes. Furthermore, these models are compared to determine the most effective model in this regard by evaluating their accuracy of prediction, alongside other performance metrics such as precision, recall and F1 score. Of the models investigated, Random Forest significantly outperformed the others, achieving an accuracy of 82.26%.

1. Introduction

As one of the most common diseases in the world, diabetes mellitus affects 37.3 million Americans in 2019 or 11.3% of the country's population. The proportion of diabetes-related deaths in the US is estimated to be between 11.5% and 12%, and significantly higher among obese people at 19.4% [1]. There are two main types of diabetes, Type 1 and Type 2, both of which share many similarities and differences in their underlying causes and recommended management strategies. While accounting for around 8% of diabetes patients, Type 1 is much more uncommon and is primarily caused by genetics, although not the only factor. While taking this into account, Type 2 diabetes receives more attention due to its prevalence. The exact biological mechanisms that allow for Type 2 diabetes to manifest are more unclear, but it is the case that genetic, environmental and lifestyle variables are all contributing factors in a variety of ways. Diabetes is not fully curable at this time, but medications such as insulin can be used to manage the symptoms carefully. The effective management of diabetes is crucial to prevent additional complications, such as eye, foot, and mouth problems, kidney disease and even certain cancers. Like other diseases, one of the best ways to effectively manage diabetes lies in diagnosing the disease early before the effects become more serious.

Diabetes also has a significant economic impact, with estimated annual expenditures for direct costs relating to diabetes increasing from $966 billion in 2021 to a projected value in excess of over $1 trillion in 2045. In the US, spending on diabetes care and management makes up roughly 25% of all healthcare spending [2]. This also does not consider the indirect costs of diabetes, such as productivity loss. The cost of diabetes is expected to increase further due to an estimated increase in the prevalence of diabetes from 10.5% in 2021 to over 12% in 2045. In addition, the cost of diabetes per individual in North America was around $8650 in 2021. From this, it can be posited that, as a whole, people's health is worsening as a general observation, resulting in more people being affected by diabetes. Diabetes was diagnosed in 537 million individuals globally in 2021; by 2045, that figure is projected to rise to 783 million. [3]. This forecast can also be partly attributed to the rate at which the global population is increasing. Additionally, it is also estimated that half of all people affected by diabetes are undiagnosed.

Diabetes is usually diagnosed in one of two ways – in the more traditional way involving manual diagnosis by health practitioners – or by technology. Each of these methods has distinct advantages and disadvantages. While it is true that manual diagnosis by health practitioners allows for human expert insight, advances in technology have made this approach much more effective as time goes on and is becoming the preferred approach. Another advantage of technological approaches is that they require less time and resources to employ. Additionally, in the initial stages of the disease, indicators of diabetes can be easier to identify through technology than by manual examinations [4] while eliminating human errors and complications in the initial analysis

☆ This work is partly supported by VC Research (VCR 0000191) for Prof Chang.
* Corresponding author.
E-mail addresses: v.chang1@aston.ac.uk, victorchang.research@gmail.com (V. Chang).

https://doi.org/10.1016/j.health.2022.100118
Received 7 September 2022; Received in revised form 8 October 2022; Accepted 15 October 2022

stages. As the availability of electronic health record data continues to increase, it becomes more attractive to utilize automated diabetes diagnosis systems. In particular, artificial intelligence (AI) and machine learning (ML) approaches are the two main ways automated diabetes diagnosis systems can be built.

1.1. Aims and objectives

Broadly speaking, the main goal of this study is to develop an automated diabetes diagnosis system by utilizing a suite of ML models. The key objectives are outlined as follows:

- Identify the most significant risk factors for Type 2 diabetes and the correlation between them to apply ML techniques more effectively for diabetes diagnosis.
- Identify and analyze the diabetes data to improve its suitability by implementing sampling techniques and analysis.
- Compare the effectiveness of a range of different ML models by looking at several evaluation metrics to determine the best ML solution for diabetes diagnosis.
- Identify limitations with automated diabetes diagnosis systems and discuss how advances in this area can be explored to advance research in this area further

1.2. Research contributions

- The relationships between all dependent variables are explored to discover the most important contributing factors to a person developing diabetes. It was discovered that the main health indicators are Body Mass Index (BMI), age, blood pressure levels, cholesterol levels, general health, physical health, walking difficulties, and income are the main contributing risk factors for diabetes.
- Five machine learning classifiers were utilized to predict type 2 diabetes: Decision Tree, Logistic Regression, Random Forest, K-Nearest Neighbor, and Gaussian Naive Bayes classifiers. Univariable and multivariable attribute analysis was used to investigate the associations of potential risk factors with type 2 diabetes. Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset. Synthetic Minority Oversampling Techniques (SMOTE) were used to stabilize the imbalance in the output variable.
- People with low social incomes are more likely to be burdened by illness-related treatment because income and diabetes have a strong link. This is especially true in nations like the US, where treatment expenses are expensive. A more thorough study should be done in the future to address the rising cost diabetes has on the economy to reduce the financial burden of diagnostic and nondiagnostic costs.
- A thorough analysis of how to prevent death due to this condition, further information such as electrocardiogram results for diagnosed individuals, carnitine levels, and glucose level parameters should be investigated.

2. Related literature

The literature reviewed in this section focuses on studies employing multiple ML algorithms to achieve the diagnosis of diabetes in the initial stages of the disease.

Kaur and Kumari [5] explore predictive modeling and analytics for diabetes using ML. They utilized five different models to detect the condition: Support Vector Machine (SVM) with the linear kernel, SVM with the RBF kernel, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN), and Multifactor Dimensionality Reduction algorithms. Each model was evaluated based on parameters such as accuracy, recall, precision, F1 score, and ROC-AUC. The experiments highlighted positive results, with all models with SVM-linear providing the best accuracy (89%) and precision (88%). In addition, KNN provided the best recall (90%) and F1 score (88%). Overall, the work suggests that SVM-linear and KNN are the optimal algorithms for the diabetes dataset and are the best for a diabetes diagnosis.

Lu et al. [6] highlight the use of ML with a network approach for disease prediction of Type 2 diabetes. They used patient records from private health insurance formulated into a graph (bipartite) projected to the patient network. They used specific features and characteristics to train eight ML algorithms for the ability to predict. The experiments showed positive effects with AUC (ranging from 0.79–0.91) when the experiments were performed. The research results convey that a combination between ML techniques and network analysis can be positively used for successful disease risk prediction in the diabetes domain.

Nadeem et al. [7] convey a fusion-based ML approach to predicting onset diabetes. The architecture performs data fusion to prepare a coherent dataset from various locations to be better aligned to the ML algorithms — which are then developed through the fusion of two well-known ML algorithms that are used, which are SVM and ANN. The results demonstrate a classification accuracy of 94.67%, which exceeds the performance of other ML models for diabetes prediction by ~1.8%. However, the diagnosis and classification of severe medical conditions are challenging because of low volume and low-quality contextual data for the training and validation of algorithms which can compromise the results.

Sarwar et al. [8] demonstrate six well-known ML algorithms in the healthcare domain for diabetes prediction, namely SVM, KNN, Logistic Regression (LR), Decision Trees (DT), Random Forest (RF) and Naïve Bayes (NB). The predictions were made on the PIMA Indian dataset containing 768 records. In the experiments, it has been observed that SVM and KNN give the highest accuracy in predicting the conditions showing 77% accuracy. The research can be expanded to a larger dataset to aim for an accuracy of 99.99%.

Sneha and Gangil [9] display an analysis of diabetes for early condition prediction. The research proposes the design of a prediction algorithm using ML algorithms. The method proposed aims to focus on selecting attributes that aid in the early detection of diabetes using predictive analysis. The research results show that the DT and RF algorithms have the highest specificity of 98.20% and 98%, respectively, and are the best for analyzing diabetic data.

Hasan et al. [10] investigate diabetes prediction using an array of ML classifiers. This technology is particularly challenging when there is a limited number of labeled data and missing values in the diabetes datasets. The algorithms used are KNN, DT, RF, AdaBoost, NB, XGBoost and Multilayer Perceptron. The experiments in the research were all conducted using the Pima Indian diabetes dataset. The proposed classifier is the best performing with sensitivity, specificity, false omission rate, diagnostic odds ratio, and ROC-AUC. Furthermore, the classifier was shown at the time to outperform the state-of-the-art by 2%.

Sisodia, D. and Sisodia, DS [11] use classification algorithms to predict diabetes early on. Three ML algorithms are evaluated on various methods performing experiments on the Pima Indians diabetes dataset. The results showed an accuracy reading of 76.30% using an NB classification algorithm. The research can be expanded in the future by improving the automation of diabetes analysis using other ML algorithms.

Rajesh and Sangeetha [12] use a range of ML classifiers for predicting diabetes diagnosis, including NB, KNN, SVM, and a range of DT algorithms such as ID3 and C4.5. They conducted their research using the Pima Indians dataset. They also used an RF algorithm to obtain a result of 100% accuracy, but they determined that since their RF model suffered from data overfitting, the results should not be acknowledged. Instead, they achieved an accuracy of 90.62% by utilizing the C4.5 DT algorithm. Since this algorithm is frequently used and widely successful in medical applications, they concluded this model to be the best for diabetes diagnosis systems.

Kelarev et al. [13] focused on the application of DT and Ensemble classifiers to predict cardiovascular autonomic neuropathy in diabetes patients. Ensemble classifiers refer to hybrid ML models which take from more than one model in order to improve their performance. In their research, they used ensemble models based on DT, namely the ADTree, J48, NBTree, RandomTree, REPTree and SimpleCart DTs. Additionally, they investigated how various ensemble techniques could be optimized in such a way. Their best results were achieved by applying AdaBoost to ensemble Bagging and DT models, achieving an accuracy of 94.84%.

Ganie and Malik [14] also employed Ensemble classifiers for early-stage diabetes detection. They collected the dataset from different departments of hospitals of Jammu and Kashmir-UT, India, such as inpatient, outpatient, and emergency. By evaluating the effectiveness of three ensemble learning techniques, i.e., Voting, Boosting, and Bagging to several classical machine learning algorithms, this study found that Bagged Decision Tree outperformed with an accuracy rate of 99.14%.

Han et al. [15] studied the application of a multitude of different ML models for diabetes diagnosis prediction, including traditional models such as SVM, RF and DT, alongside ensemble models focused on using rule extractions from SVM. They found that by combining this approach with a RF classifier, the precision scores improved when compared to the base RF and SVM models (89.6% compared to 81.2% and 88.4%, respectively). Conversely, they found the recall scores of the ensemble classifier (44.3%) were lower than the base RF model (49.0%) but still higher than the base SVM model (40.0%). This was attributed to the SVM + RF ensemble using simpler rules than the base RF model. Nevertheless, they concluded that using the SVM + RF ensemble was still preferred over either of the base models.

In the work of Hassan et al. [16], they employed unsupervised methods for building a diabetes detection model. The model employed the K-means clustering technique to group the features and the Silhouette and Elbow methods were used to determine the optimal number of clusters. Then, they applied several algorithms (Multilayer Perceptron, RF, DT, SVM, and KNN) to the created cluster-based dataset and complete dataset, the evaluation results show that RF achieved the best accuracy of 99.57% on the cluster-based dataset.

Ramesh et al. [17] created a healthcare monitoring framework to manage diabetes remotely. The framework uses ML algorithms embedded in real-time diabetes prediction and the interconnection of various devices (physician portals, smart patient devices) to enable healthcare experts to make informed decisions remotely, while enabling cost reduction and closed-loop communication.

Khanam and Foo [18] compared various ML techniques for diabetes prediction using the Pima Indian Diabetes Dataset. Throughout the study, they used seven ML algorithms (DT, KNN, RF, NB, AB, LR, SVM), with all algorithms achieving greater than 70% accuracy. Their study showed KNN and AB algorithms to achieve the best results for diabetes prediction, with both achieving accuracy scores of 79.42%. KNN had slightly higher precision and F1-measure scores than AB, so it can be considered slightly better in this regard.

Krishnamoorthi et al. [19] created a framework for disease prediction in healthcare by utilizing ML techniques. The study enforced ML techniques on the Pima Indian Diabetes Dataset. The results demonstrated that the LR performed better than the other ML algorithms. For diabetes disease, it has been highlighted that there is a correlation between glucose and BMI. One limitation of the study is that they used a structured dataset and aimed to test with an unstructured dataset. It is suggested that the methods used in the paper can be easily applied to other healthcare domains to predict other diseases.

Ahmed et al. [20] used decision-level Fusion ML as a decision support system for predicting diabetes. It has been noticed that many ML techniques have been used for the prediction, but the authors aimed to focus on the accuracy element with the various proposed models. For this study, two particular ML techniques have been used, namely SVM and ANN. As a result, the fuzzy detection system presented accuracy readings of 94.87%, demonstrating higher than other systems and saving many lives.

Goyal and Jain [21] proposed a novel approach by integrating a number of classification algorithms, such as NB, LR, J48 DT, SVM and RF using 10-fold cross-validation techniques and ensemble Bagging algorithms. Of these, the highest accuracy was demonstrated by SVM as a single classifier at 77.34%, whereas the accuracy of the ensemble LR-based model was even higher at 77.60%. Results of the experiment show a good form of increment in the accuracy of the classifier with low error rate and enhanced ROC-AUC when applied with 10-fold cross-validation. This type of approach is very helpful in medical diagnosis and can enable the medical practitioners to take appropriate decisions for chronic diseases, including diabetes.

Abdulhadi and Al-Mousa [22] explored the use of supervised learning models that could help assist doctors in the early detection of diabetes to improve the quality of patient's lives. Their paper presented multiple techniques that were used to train multiple models. The models they utilized were LR, Linear Discriminant Analysis, Linear SVM, Polynomial SVM, RF and Voting Classifier. Of these, RF achieved the highest accuracy of 82%.

Gupta et al. [23] proposed two prediction models through the use of quantum ML and deep learning techniques. The aim of their study was to assess and compare the predictive capabilities of these models when compared to other state-of-the-art models using the Pima Indian Diabetes Dataset. Of the two models proposed, the deep learning multilayer perceptron model proved to be the most effective predictor with 95% accuracy when using four hidden layers, and outperformed all other DL models by at least 7.36%. The quantum ML model returned an accuracy of 86% at best.

Sivaranjani et al. [24] applied ML algorithms after applying dimensionality reduction and feature selection to predict diabetes using the Pima Indian Diabetes Dataset. More specifically, their study used step forward feature selection and step backward feature elimination and principal component analysis and their application to SVM and RF models. The results of each classifier is evaluated using 5-fold cross validation. RF produced similar predictive performance regardless of whether step forward or step backward feature selection was employed, at 82.9%. SVM performed slightly worse, but the use of step backward feature selection proved more fruitful than step forward, increasing its performance by 1.5% up to 81.41%.

Finally, Lama et al. [25] study the use of ML methodologies with the Stockholm Diabetes Preventive Program to investigate the most important risk factors for developing Type 2 diabetes using SHAP dependence values. They discovered that the most significant contributors towards developing Type 2 diabetes were BMI, waist-hip ratio, age, blood pressure and genetics. Furthermore, they showed that having a combination of these risk factors increased the risk of developing diabetes multiplicatively.

To find the most accurate model to predict diabetes, Refat et al. [26] applied six machine learning models, namely XGBoost, RF, DT, SVM, LR, and KNN. They also used deep learning base classification techniques, including LSTM, ANN, and MLP. They discovered that has a 100% train and test accuracy of zero train and test loss, and one for precision, recall, F1-score, and ROC-AUC accuracy, with an execution time of 46.074 s, making it the most accurate model for this classification. Although the RF test accuracy is 92.3%, it should be emphasized that it is still effective. However, compared to ML models in this study, deep learning models performed worse. An ANN provided good performance for three different deep learning models with accuracy ranging from 88.5% to 92.3%.

3. Data and methodology

The methodology of the study has six key steps (Fig. 1), consisting of a literature survey, data acquisition, exploratory analysis, data preprocessing, model selection and model evaluation. Firstly, existing

Table 1
Comparison of accuracy results from a literature survey.

Study	Model(s)	Accuracy (%)
Kaur and Kumani [5]	SVM-Linear	89
Nadeem et al. [7]	SVM-ANN ensemble	94.67
Sarwar et al. [8]	SVM and KNN	77
Sissodia and Sissodia [11]	NB	76.30
Rajesh and Sangeetha [12]	C4.5 DT	90.62
Kelarev et al. [13]	Adaboost DT with Bagging	94.84
Ganie and Malik [14]	Bagged DT	99.14
Hassan et al. [16]	**K-means clustering and RF**	**99.57**
Khanam and Foo [18]	KNN and AB	79.42
Ahmed et al. [20]	Fusion ML Decision	94.87
Goyal and Jain [21]	LR ensemble	77.60
Abdulhadi and Al-Mousa [22]	RF	82
Gupta et al. [23]	Multilayer Perceptron	95
Sivaranjani et al. [24]	SVM with feed backward feature elimination	82.9

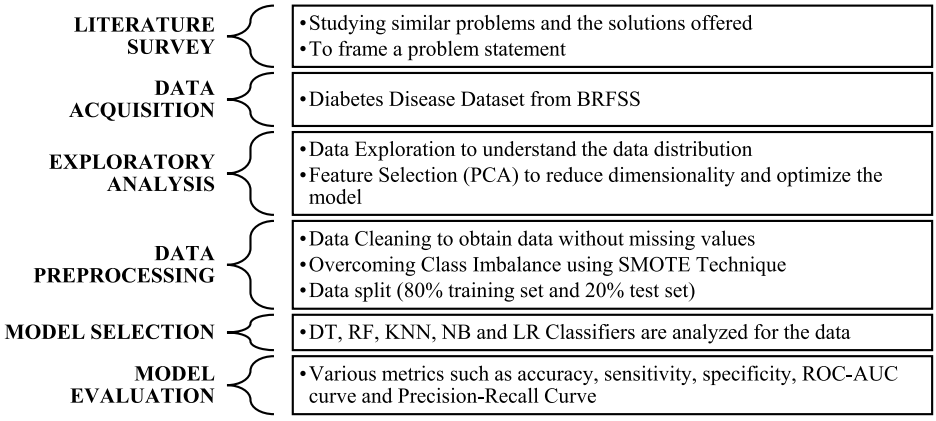

Fig. 1. Methodology of the study.

diabetes-related research was surveyed to develop the research questions and select an appropriate solution. Identifying a suitable dataset is important for constructing an effective classifier. Therefore, to accurately predict diabetes, the dataset was obtained from BRFSS, which provided sufficient data samples and attributes for training purposes. In the exploratory analysis, the dataset was cleaned to ensure that there were no missing values, and then exploratory data analysis was conducted to get a greater understanding of the data for pre-processing. DT, RF, KNN, NB and LR were selected as the algorithms to build the classifiers and the performance of the classifiers was analyzed and compared for accuracy, sensitivity, specificity, ROC-AUC curve, and Precision–Recall Curve.

3.1. Data description

The dataset used in this study is a subset of Behavioral Risk Factor Surveillance System (BRFSS) data. It is a clean dataset of 70,692 responses to the CDC's BRFSS2015 survey. Participants without diabetes, those with prediabetes, and those with diabetes are split equally. This dataset has 253,680 records and twenty-one unbalanced feature variables. A more detailed description of these variables is shown in Appendix. Diabetes_binary is the two-class binary target variable. A value of 0 indicates that the patient does not have diabetes, while a score of 1 indicates that they have prediabetes or diabetes.[1]

3.2. Exploratory data analysis

Exploratory Data Analysis is a method used to investigate and analyze data using various visual techniques. These techniques allow data scientists to spot anomalies, discover trends, check assumptions, and derive better insights for their tasks [27].

In our study, a correlation matrix is first computed to understand the relationship between each pair of variables. If there is a strong relationship between attributes, removal of the attributes should be considered. Through several types of data visualization techniques, the distribution of the attributes and the data balance are analyzed to improve understanding of the dataset and its features.

Feature engineering helps identify the features that hold the most relevant information to the predicted target [28]. PCA is used as the feature selection tool to reduce the dimensionality of data and optimize the model. Moreover, it can preserve as much information present in the complete data as possible [29].

3.3. Data pre-processing

Before feeding the data to the model, it must be partitioned into train and test sets. The training partition is used to train the model, and the test partition is used to assess its performance. The provided dataset is split in an 80:20 ratio for training and testing, respectively. The choice of this ratio follows the Pareto principle, the main idea of which is that 80% of the effect in most cases comes from 20% of the causes [30]. When the dataset was cleaned, 5% of null values were added by using sample functions to all of the variables except for the target variable.

Training the model with this missing data might result in inaccurate performance. Therefore, data entries containing missing values were removed from the dataset for better results. The MinMax Scaler was applied to normalize the data values before fitting the dataset. The given dataset is extremely biased, with just 15% of positive samples and the rest being negatives. To address this problem, SMOTE techniques were applied to oversample entries from the minority class, resulting

[1] https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

in an equal distribution dataset with over 200,000 data entries. As the data is categorial, and the target variable is binary in nature, we apply classification models to determine the accuracy of prediction.

3.3.1. SMOTE

SMOTE [31] is a statistical sampling technique used to increase the number of examples in the dataset uniformly. The component creates new instances based on minority cases. It takes the entire dataset as an input, but it increases the percentage of only the minority cases. The proportion of the majority of cases is unaffected by the implementation of SMOTE.

The new occurrences are different from minority cases that already exist. Instead, each target class and its close neighbors are sampled from the feature space by the algorithm. The algorithm then creates fresh examples incorporating traits from both the target case and its neighbors. With this method, each class has access to additional characteristics, and the samples are more inclusive.

3.3.2. ADASYN

In order to produce more synthetic data for minority class examples that are more challenging to learn than minority class examples that are easier to learn, ADASYN can be used to utilize weighted distribution for separate minority class instances. The ADASYN technique thereby enhances learning by reducing the bias brought on by the class imbalance and adaptively shifting the classification decision boundary toward the challenging samples.

The dataset used to train the ML classifiers was imbalanced, with 80% of the cases having a target variable value of 0 and 20% of the cases having a target variable value of 1. Such a large data imbalance could potentially result in a high likelihood of inaccurate results, and we need only to increase the percentage of minority class to balance instances in the target variables evenly. Therefore, SMOTE techniques are shown to be more suitable when compared to ADASYN for this dataset.

3.3.3. Principal component analysis

Principal component analysis (PCA) [32] is an unsupervised statistical method that reduces the dataset dimensions. The use of PCA facilitates linking variables by identifying relationships between them. It thereby simplifies the data dimensionality but also retains the important existing trends and patterns within the data.

3.3.4. T-distributed stochastic neighbor embedding

T-distributed stochastic neighbor embedding [33] (t-SNE) converts a high dimensional dataset into a low dimensional graph while retaining a substantial portion of the original data. It does this by giving each data point a specific place on a two- or three-dimensional map. By locating clusters in the data, this approach makes sure that an embedding preserves the meaning in the data. While reducing dimensionality, t-SNE seeks to keep comparable examples near together and dissimilar instances away. T-SNE is non-linear, so it can capture the structure of trickier manifolds and involves hyperparameters, unlike PCA.

Based on the above features of both PCA and t-SNE, it is clear that PCA is a more suitable approach as only the variables that do not contribute much to the target variable should be removed. There is no such requirement for keeping comparable samples used in t-SNE together in the dataset.

Datasets with high dimensionality can lead to overfitting when utilizing ML classifiers. The dataset used for this study consists of twenty-one variables, meaning that it can be considered to have high dimensionality. After applying PCA, eight variables, namely BMI, age, high blood pressure, high cholesterol, general health, physical health, walking difficulty and income [34], were identified, contributing the most to the prediction of a diabetes diagnosis.

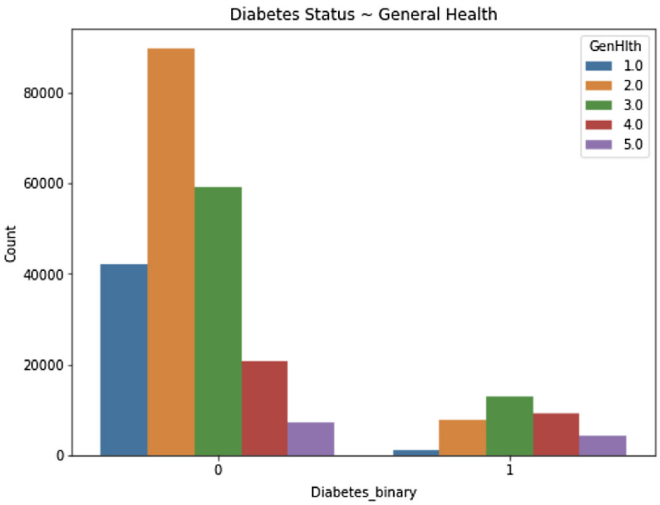

Fig. 2. Bar plot showing diabetes status and general health.

3.4. Machine learning classifiers and evaluation metrics

Five different ML algorithms were applied to build the classifiers using the training set. A summary of the algorithms used is summarized in Table 2:

A number of algorithms have been used for the prediction of diabetes, and we have identified the most popular of these and selected these five algorithms to predict our data for the following reasons:

- When using XGBoost, optimal performance is achieved on datasets with at least eight variables. Therefore, it was determined not to be appropriate for this dataset.
- SVM is not ideal for this dataset because it has slow calculation times when processing large datasets.
- Linear Regression and KNN regressor were not selected as they are regression models and not suitable for this study.
- ANN and other related deep learning techniques were not chosen as they would potentially be overfitting the model.
- After analyzing all these scenarios, the following five classifiers were selected: DT, RF, KNN, LR and NB.

After training the classifier, four results are obtained by comparing the output with the predetermined labels: true positive, true negative, false positive and false negative, which constitute the confusion matrix. Evaluation metrics based on the confusion matrix are used to verify the correctness of the ML model, including Accuracy, Sensitivity, Specificity, ROC-AUC curve, and Precision–Recall curve.

The pseudocode for the ML classifiers is outlined in Table 3.

4. Implementation and results

4.1. Exploratory data analysis

General health is measured on a numerical scale of 1 to 5. From Fig. 2, we can determine that those with a general health score of 3–5 are more prone to develop diabetes. In this chart, class 0 indicates no diabetes and 1 indicates prediabetes or diabetes. There are over 80,000 patients with a general health of 3, 20,000 with a general health of 4, and approximately 10,000 with a general health of 5 for class 0 (no diabetes), indicating that many people who are not diabetic should improve their overall health to avoid contracting the disease since they are at high risk by having a general health score in this range. For the more than 20,000 patients with class 1 diabetes or prediabetes, they must improve their physical health with medication under the guidance of a doctor to improve their life quality and overall physical health. The

Table 2
ML classifiers.

ML classifier	Summary
DT	DT is a tree-structured classifier wherein the split at each node is chosen while the internal and leaf node samples are kept to a minimum. A maximum depth of 50 was chosen and criterion entropy was used to evaluate the leaf cleanliness.
RF	RF builds decision trees out of several samples and classifies them based on the majority vote from each tree. As the data was clean and binary, we applied ten trees as depth. In order to acquire the performance, criteria entropy and a random state of 42 were used.
KNN	The nearest neighbors are found using a mix of Ball Tree (BT), KD Tree (KDT), or brute-force search. The Manhattan or Euclidean distance functions are used for computing the classifications. A k-value of 3 is used as it was determined to be the best k-value for the dataset.
LR	LR is most commonly used for the classification of binary classes. It internally uses the sigmoid function to learn the linear relationships between variables. The solvers and penalty can be changed to see visible differences in the classification. In this study, a random state of 0 was used.
NB	Naive Bayes is a probabilistic AI calculation based on the Bayes Theorem that is used in a wide range of classification problems.

Table 3
Pseudocode for early prediction of diabetes using ML classifiers.

```
1   Import dataset and data
2   Dimensionality Reduction: PCA
3   Class Imbalance: SMOTE
4   Identify the ML classifiers that will be used for the classification
5   MLC=[ DecisionTreeClassifier ( ), RandomForestClassifier ( ), KNeighborsClassifier ( ),
      LogisticRegression ( ), NB ( )]
6   for (i=0;i<5;i++) do
      model = MLC[i]
      model.fit ( );
      model.predict ( );
      print(classification_report ( ), accuracy_score ( ), precision_score ( ), recall_score ( ),
        f1_score ( ), hamming_loss ( ), roc_curve ( ), roc_auc_score( ))
    end
```

government creates several health programs for them to enroll in and improve their lifestyles by changing their diet and exercising apart from taking their medication.

The distribution of patients' health states compared to their diabetes status is measured on a scale of 1 to 5, with 1 being the best and 5 being the worst. Furthermore, we can see that the most significant difference in the form of the distribution is that the density of class 0 is more on a scale of 1 to 3. Still, the density of class 1 is more widely dispersed from 1 to 5, indicating that the health state of diabetic patients is generally well managed.

The overall form and distribution of the recommendations for education compared to the diabetes status of patients (quartiles relatively close to each other) are similar in the next example. However, there are more outliers in the case of the positive class.

In a comparison of patient income when compared to diabetes status, we can see that the density of the plot increases with income level, indicating that the higher the income, the higher the chances of getting the disease. In contrast, people with class 1 diabetes have a plot evenly distributed on a scale of 1 to 8, indicating that all income level groups have the disease, while it is denser among the income levels of 5 to 8.

In Fig. 4, it can be seen that 96.5% of patients had high cholesterol levels, and 94.7% are not heavy drinkers. These are both considered severe risk factors for developing diabetes.

Diabetes has a moderate risk of stroke and heart disease, according to this study. The majority of patients did not have a stroke or heart disease, as seen in Fig. 5. Furthermore, diabetic patients had a fractional value of an extra number of strokes or heart attacks compared to healthy people. Even healthy people can have a stroke or heart attack, but their risk is lower than that of individuals who have already been diagnosed with diabetes.

From Fig. 6, it is shown that patients with no diabetes have lesser mental health issues when compared to patients with diabetes. While diabetes patients have slightly high mental health issues as their data points are more predominant and continuous, it is also clear that patients who have higher physical activity are less likely to have diabetes than patients with low activity.

In Fig. 7, the BMI distribution is the most frequent between 22 and 30. BMI is one of the most crucial risk factors for developing diabetes. Higher BMI correlates strongly with the probability of occurrence of the disease.

The correlation plot shown in Fig. 8 shows the correlation between all the features and there is no linear correlation between feature variables.

The data is equally sampled using SMOTE for class 0 and class 1 in the output target variable to get more efficient predictions when ML algorithms are applied with less biased results (see Fig. 9).

As shown in Fig. 10, the components below 0.4 are not contributing much towards the target variable and will make little to no difference in the output, so that they can be dropped.

The SMOTE technique is applied to overcome the imbalance in the output variable. PCA is applied to reduce the number of features and select only those that contribute the most to the target variable, leaving eight components (see Fig. 11).

4.2. Machine learning models

4.2.1. Logistic regression

Regression analysis is a predictive modeling analysis since it always deals with predictions. Regression is divided into three categories. The first one is linear regression. One of the important tools used in the natural sciences is logistic regression, which is strongly associated with neural networks in sociology. It simply classifies and determines if a specific event is occurring or not, which addresses many categorization problems. The output of LR predicts the outcome in binary form for a discrete dependent variable.

It uses discrete dependent variables and is a statistical classification model. The logistic function or sigmoid function, which accepts a value between 0 and 1, is how LR operates. The phrase "logistic" derives from

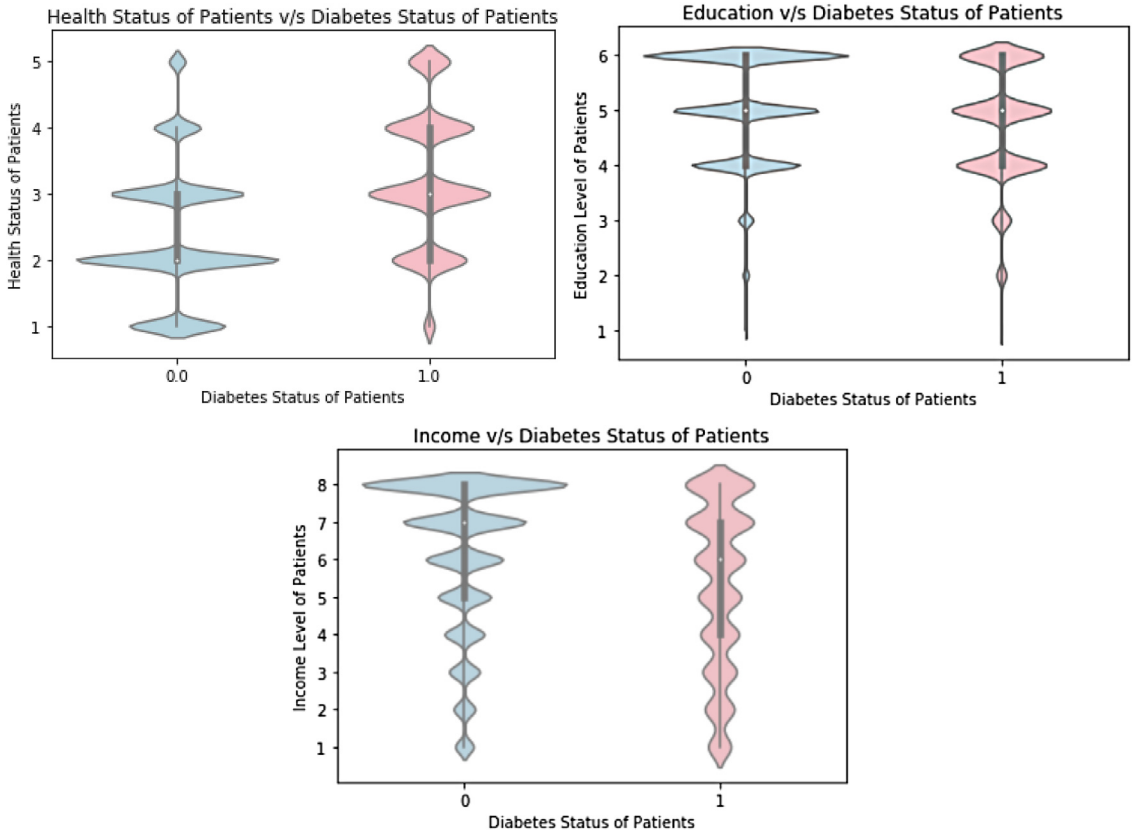

Fig. 3. Violin plots showing the diabetes status of patients and income.

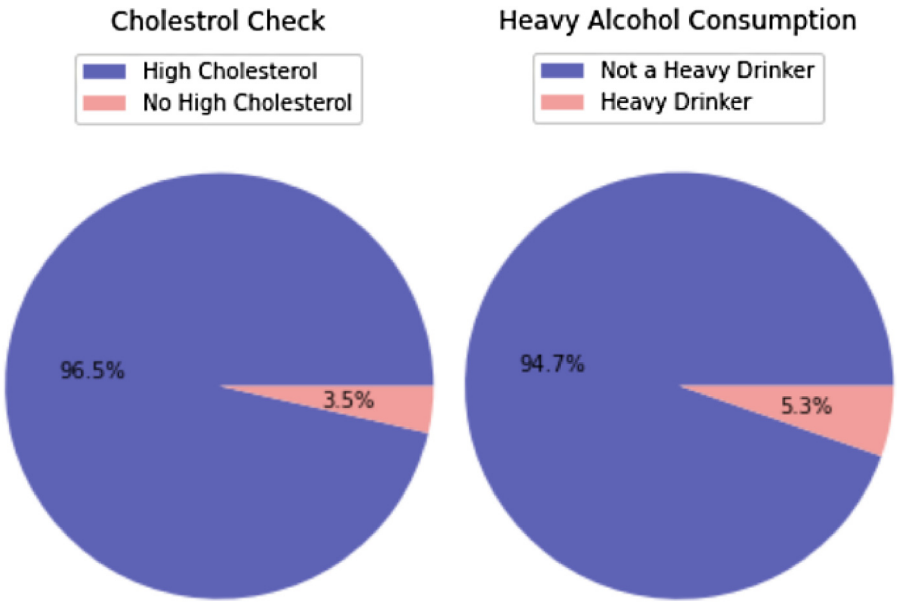

Fig. 4. Pie charts for cholesterol levels and heavy alcohol consumption.

the algorithm's primary function, the Logit function. It is a controlled learning algorithm that deals with likelihood and only allows for two alternative outcomes: yes or no, true or false, 0 or 1, high or low. The probability is rounded to 0 if it is less than 0.5 and to 1 if it is more than 0.5. The S-shaped sigmoid function presented below takes real values and plots them in the [0,1] range, as shown in (1).

$$\cos(y) = \frac{1}{(1 + \exp(-y))} \tag{1}$$

4.2.2. Support Vector Machine (SVM)

SVM is known to perform well in many areas, like bioinformatics, content, picture acknowledgment, speaker ID, pattern recognition in photographs, and target recognition. This categorization technique works with both linear and non-linear data. It performs classification by building a hyperplane in high- or infinite-dimensional space that, under ideal circumstances, isolates the data into two classes that can be used for classification or regression, as shown in Fig. 12. For a given set of objects, several different separation hyperplanes are feasible.

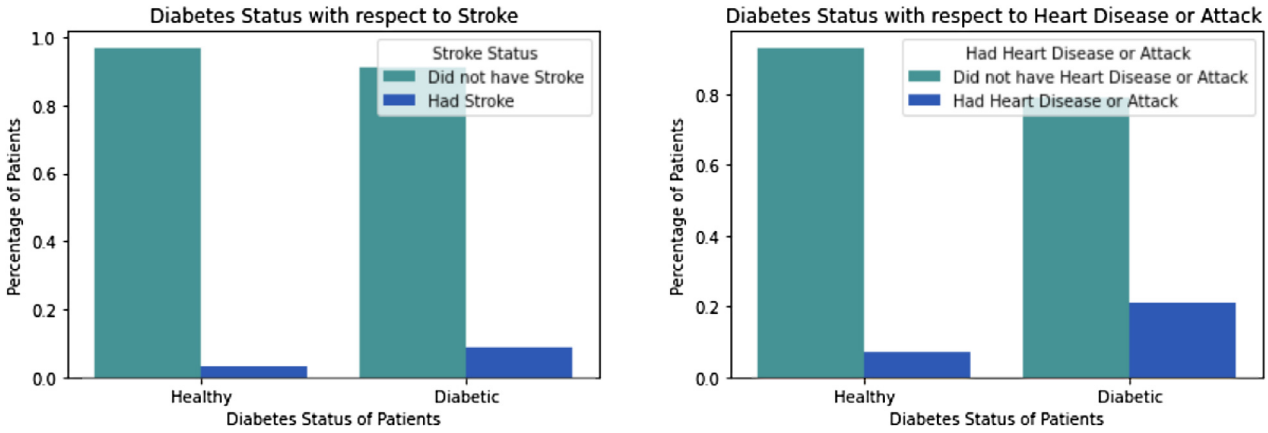

Fig. 5. Bar charts for diabetes status with respect to stroke and heart disease.

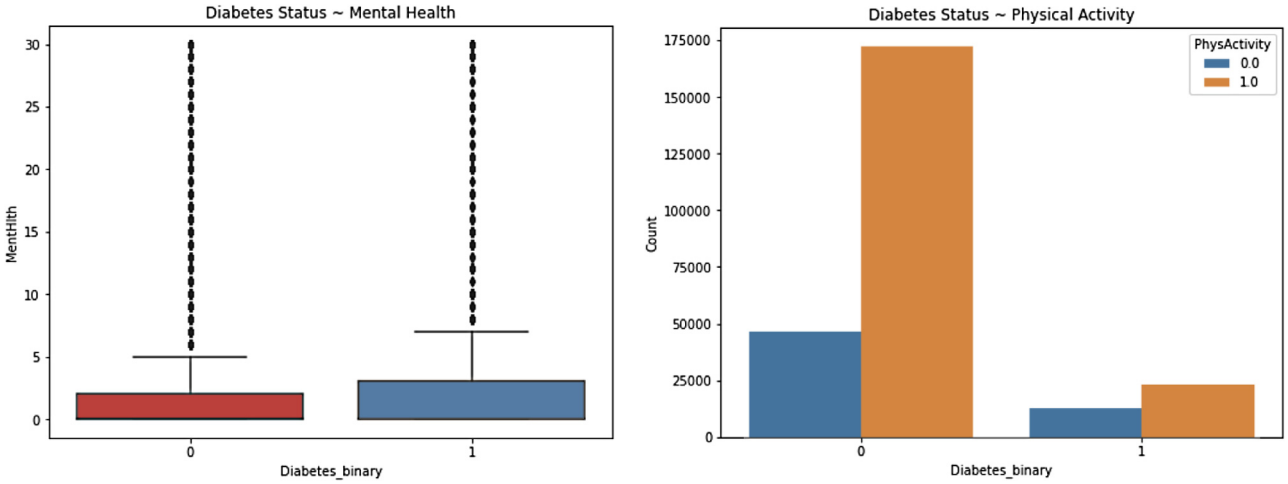

Fig. 6. Box plots for diabetes status with respect to mental health and physical activity.

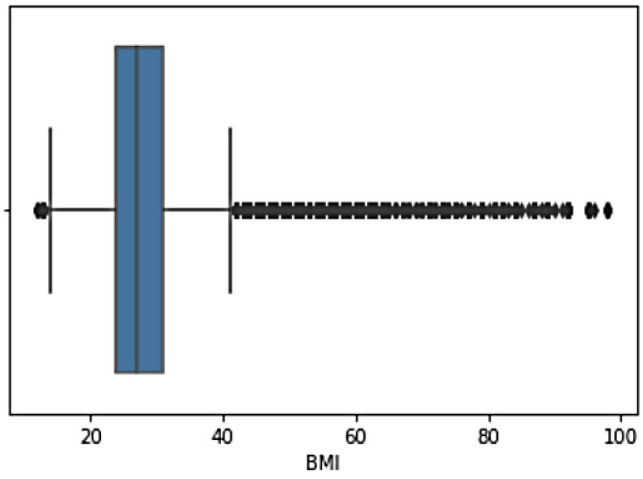

Fig. 7. BMI distribution among patients.

The hyperplanes should not be located close to the datapoints but should be selected, so they are far from the information focal points for each categorization. Support vectors are used to describe datapoints that are more closely spaced from the hyperplane [35].

4.2.3. K-Nearest Neighbor (KNN)

KNN [36] is a classification algorithm that presents a method to manage data requests that assess how likely it is that data is thought to be a person from a social event based on several data restrictions currently in place. This tactic is known as a sluggish methodology since it would not build a structure utilizing the train set unless an instructional assortment request was made. It also goes by the name "non-parametric approach". This indicates that since the technique is based on data, it makes no assumptions about hidden data scattering. Red triangles, blue squares, and green circles are the datapoints that represent the various classes, as shown in Fig. 13.

4.2.4. Random Forest

RF is defined as a process that creates numerous decision trees by referring to each tree to make decisions. Typically, n number of datapoints are picked from the dataset, and by combining them, a stable decision is produced. If there are more guesses, the average of all predictions is used. The classification and regression problems are resolved using the RF [37] technique. As depicted in Fig. 14, a considerable number of trees are produced in the forest throughout this process. More trees mean a healthier forest, which produces findings with great accuracy.

The RF architecture is made up of several trees. Every tree offers a particular selection. By averaging all of the options, the most recent prediction is evaluated.

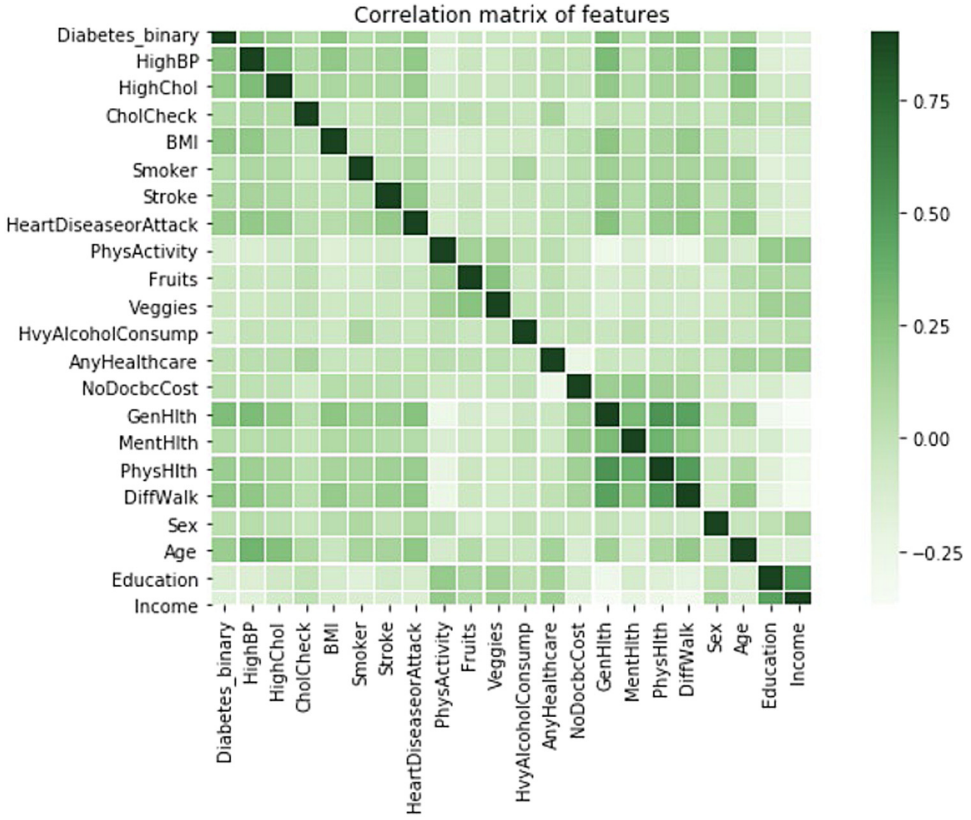

Fig. 8. Correlation plot of data features.

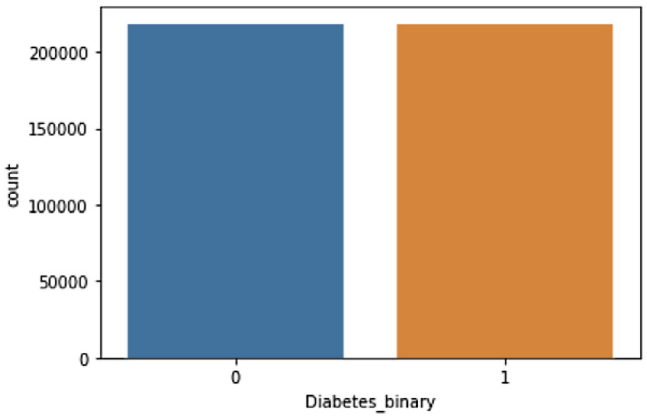

Fig. 9. Count plot for the target variable.

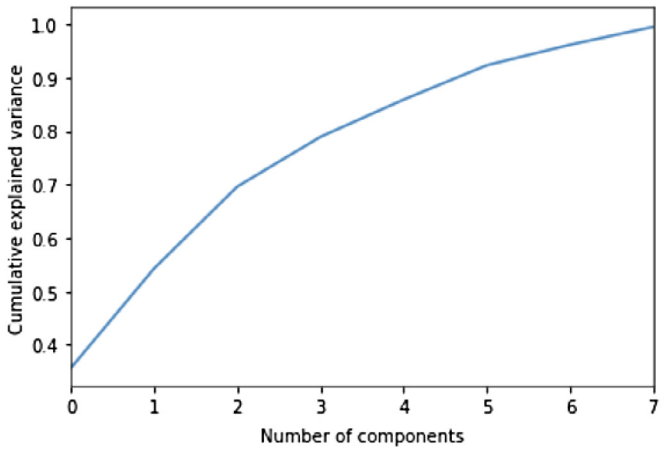

Fig. 10. Cumulative explained variance compared to the number of components.

4.2.5. Naive Bayesian

Naive Bayesian is a term used to describe a grouping algorithm, a probabilistic classifier that relies on the autonomous supposition between several indicators. The method that uses the dataset as its source of data conducts research and predicts the class grade using Bayes' Theorem. It determines whether there is a chance that the input data will be classified and helps predict the class of the test for obscure information. This style of organizing works well with large datasets. Using the Bayes Theorem equation shown below, it is possible to calculate the back likelihood for each class.

$$P\left(\frac{a}{z}\right) = \frac{\left(P\left(\frac{z}{a}\right) * P(a)\right)}{P(z)} \tag{2}$$

$$P\left(\frac{a}{z}\right) = \left(P\left(\frac{z1}{a}\right) * P\left(\frac{z2}{a}\right)\right) * \ldots * P\left(\frac{zn}{a}\right) * P(c) \tag{3}$$

- P(a/z) represents the target/object class's back likelihood for the provided predictor.
  - The target/object class's earlier likelihood is P(a).
  - The probability, or P(a/c), measures how likely a predictor is given a class.
- The earlier predictor likelihood is denoted by P(z).

4.2.6. Decision Trees

Decision Trees are used to extract data from a vast number of available datasets using decision rules. Data that can be quickly stored and further classified using a decision tree is simply categorized. In this

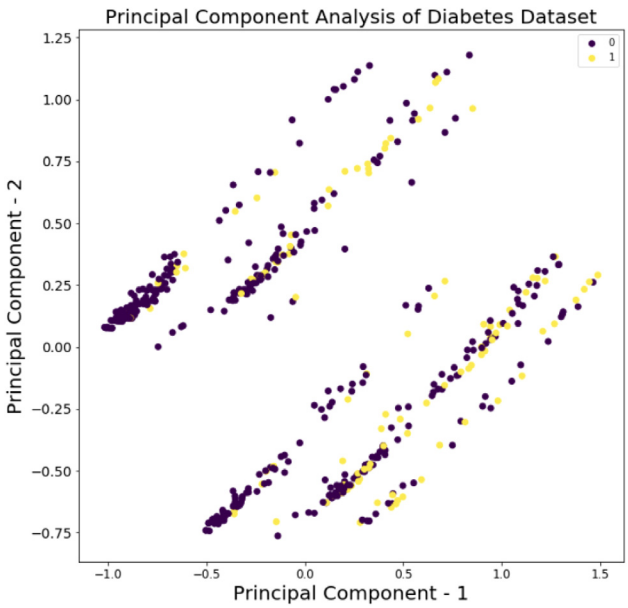

Fig. 11. Dimensionality reduction of diabetes dataset.

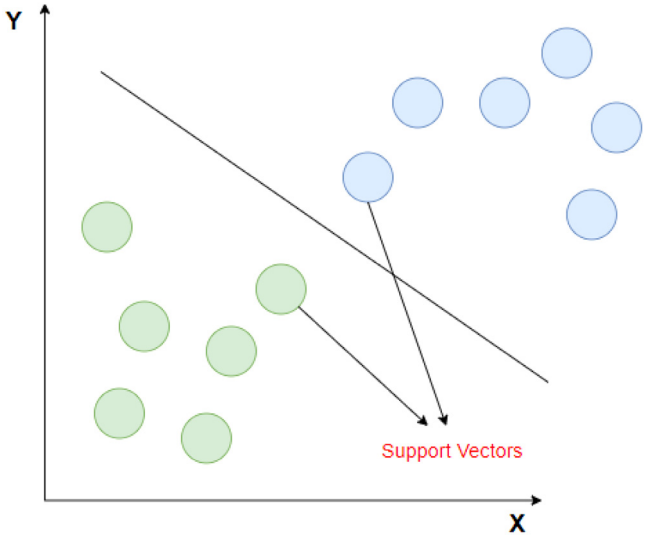

Fig. 12. Representation of support vectors in the SVM model.

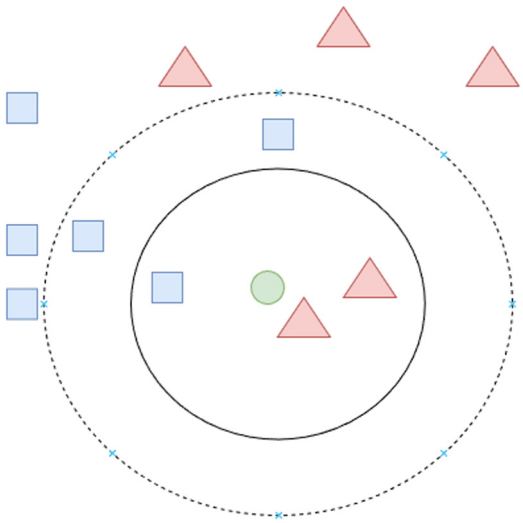

Fig. 13. KNN representation for k = 3.

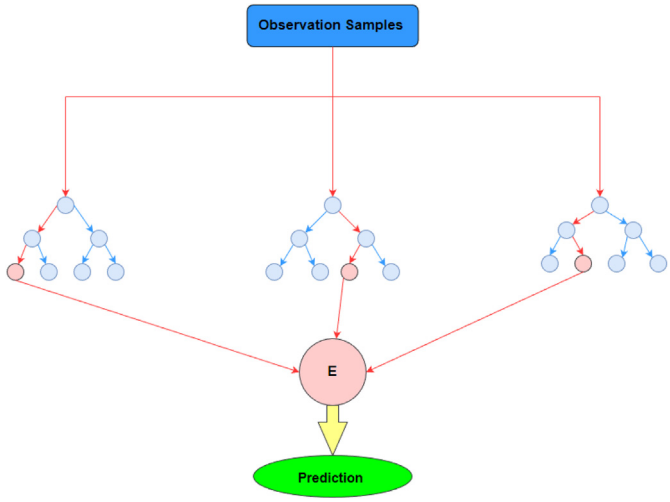

Fig. 14. Random Forest.

study, we discuss various decision-tree-based techniques for classifying data (see Fig. 15).

For example, according to the DT above, smokers tend to pass away early. If a person does not smoke, whether or not they drink is the next consideration. A person becomes old and dies if they do not drink or smoke.

A person dies elderly if they drink alcohol, do not smoke, and weigh less than 90 kg. If a person drinks but does not smoke, their weight is considered. Finally, if a person does not smoke, does drink, and weighs more than 90 kg, they are more likely to die early.

4.3. Evaluation of results

Before feeding the data to the model, the dataset is cleaned, and the data values are normalized before fitting the dataset. SMOTE techniques are used to oversample entries from the minority class, resulting in an equal distribution dataset with over 200,000 rows. The dataset is split in an 80:20 ratio in accordance with the Pareto principle.

After training, the classifiers are evaluated using several metrics based on the confusion matrix (Fig. 16). In this matrix, TP (true positive) denotes the number of samples for which the model correctly predicted the positive category. TN (true negative) is the number of samples for which the model correctly predicted the negative category. FP (false positive) refers to the number of samples for which the model incorrectly predicted the positive category. FN (false negative) is the number of samples in the negative category that the model incorrectly predicts.

The metrics include accuracy, sensitivity, specificity, ROC-AUC curve, and Precision–Recall curve. Confusion matrices and the formulae used to calculate these metrics are defined below.

Accuracy computes the number of correct predictions a model makes, and hamming loss is the number of incorrect predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

In comparison to all other models for Diabetes (Table 4), RF has the highest accuracy (82.26%), which means that RF was 82.26% correct in distinguishing cases with and without diabetes in the test sample. NB performed the worst, with an accuracy rate of 70.56%. However, accuracy is not an adequate performance measure for assessing the power of the whole model on its own and is not truly predictive.

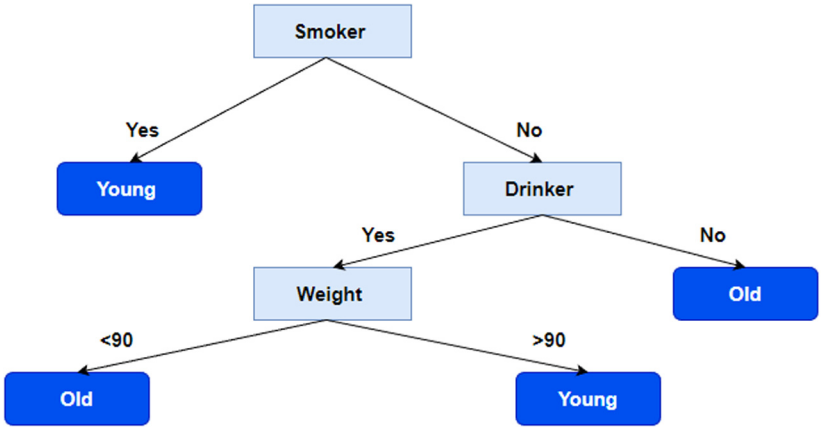

Fig. 15. Decision Tree flowchart.

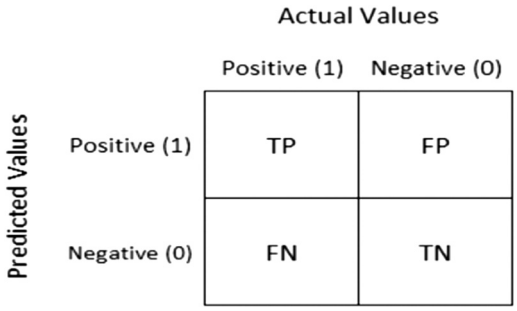

Fig. 16. Confusion matrix.

	precision	recall	f1-score	support
0	0.81	0.84	0.83	43671
1	0.83	0.80	0.82	43663
accuracy			0.82	87334
macro avg	0.82	0.82	0.82	87334
weighted avg	0.82	0.82	0.82	87334

Fig. 17. Random Forest confusion matrix.

Precision, Recall, and Specificity are used alongside accuracy to provide a more balanced evaluation approach.

$$\text{Hamming loss} = 1 - \text{Accuracy} \tag{5}$$

The Hamming loss is the proportion of incorrect labels to total labels. Hammering loss is determined in multi-class classification as the hamming distance between 'actual' and 'predictions.' Hamming loss penalizes only the individual labels in multi-label categorization. The lower the loss, the better the model performance. RF has the lowest Hamming loss of 17.74%.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

In disease detection, predicted false positives can lead to misdiagnosis and cause wastage of healthcare resources, and improving the precision of diagnostic models can help to improve this problem. Precision quantifies the number of correctly forecast positive observations: This is accomplished by counting the samples that were correctly predicted as positive (TP) and dividing them by the total number of positive predictions, correct or incorrect (TP, FP). According to Table 4, of all the classifiers evaluated, the RF model had the highest precision of 83.47%, followed by DT at 83.02%, indicating that they correctly predicted more than 83% of all cases predicted to be diabetic. LR performed the worst in this perspective, with a Precision rate of 70.56%.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

Recall or sensitivity, like precision, aims to figure out the proportion of true positives that were accurately detected. It accomplishes this by dividing the correctly predicted positive samples (TP) by the total number of positives, either correctly or incorrectly predicted as positive (TP, FN). Recall measures the number of correct positive predictions out of all possible positive predictions made. The highest sensitivity, i.e., recall, is 80.45% for RF, meaning 80.45% of the cases with diabetes in the test set were correctly selected by the RF classifier. NB is the worst performer in this perspective, with a recall rate of 67.07%. False negatives should be avoided, as missing the presence of the disease can lead to treatment being delayed and can cause real damage [38].

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{8}$$

Specificity measures the percentage of actual negatives that were accurately identified. This is accomplished by dividing the number of correctly predicted negative samples by the total number of samples that were either correctly or mistakenly forecasted as negative (TN, FP). The specificity of DT and RF reached over 84% with 84.05% and 84.07%, respectively, indicating that they can correctly identify over 84% of the cases without diabetes in the test set. A detection system with a high specificity contributes to the issue of over-medicalization, as diagnosing a patient without diabetes as a person with diabetes can cause anxiety and unnecessary follow-up procedures [39].

$$\text{F} - \text{measure} = \frac{2\,(Recall \times Precision)}{Recall + Precision} \tag{9}$$

An effective diabetes detection system should avoid both missed and misdiagnosis, but accuracy and specificity are conflicting performance metrics. F-measure accounts for both precision and recall. It can have a maximum value of 1, signifying flawless precision and recall, and a minimum value of 0 if either precision or recall is zero. RF has the highest F1 score of 82.26%, while it achieved the best performance in all other evaluation metrics, which validates it as the best diabetes classifier in this study.

Table 4 and Fig. 18 shows the comparison of the performance metrics of accuracy, precision, sensitivity, specificity, F1 score and hamming loss. These metrics evaluate the classifiers' performance [40,41]. According to Table 4, RF was the best classifier for diabetes detection on the test set in terms of every metric. Moreover, as shown in Fig. 17, RF was more accurate in predicting cases with diabetes (Precision: 83%) than in predicting cases without diabetes (Precision: 81%), whereas it was less able to identify cases with diabetes from all

Table 4
Performance metrics for the ML models.

Classifier	Accuracy	Precision	Sensitivity	Specificity	F1-Score	Hamming loss
DT	81.02%	83.02%	77.98%	84.05%	81.10%	18.98%
RF	**82.26%**	**83.47%**	**80.45%**	**84.07%**	**82.26%**	**17.74%**
KNN	80.55%	81.20%	79.50%	81.59%	80.54%	19.45%
LR	72.64%	72.06%	73.95%	71.34%	72.64%	27.36%
NB	70.56%	72.09%	67.07%	74.04%	70.52%	29.44%

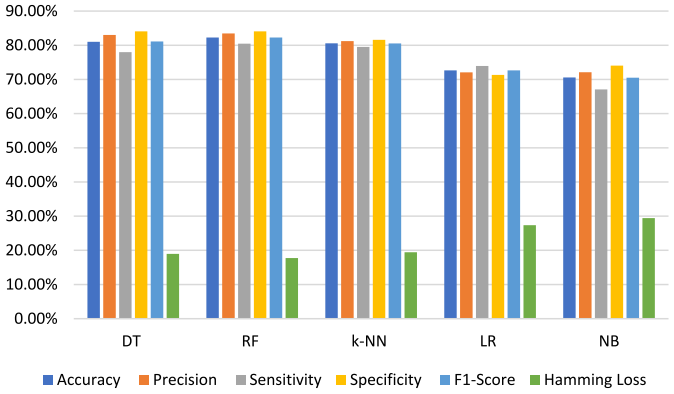

Fig. 18. Visual comparison of ML model performance.

positive samples (Recall: 80%) than it was to identify cases without diabetes (Recall: 84%). For the DT and KNN (Table 4), although DT (Accuracy: 81.02%) was only 0.47% more accurate than KNN (Accuracy: 80.55%) when classifying all test samples, DT (Precision: 83.02%) was 1.82% more accurate than KNN (Precision: 81.20%) in predicting cases with diabetes, and DT (Specificity: 84.05%) was 2.46% less likely to misdiagnose diabetes than KNN (Specificity: 81.59%). In addition, the scores of the hamming loss have slight variation, with KNN having a higher hamming loss. A lesser value of hamming loss indicates a better classifier. Therefore, DT is a better classifier than KNN. KNN (Sensitivity: 79.50%) had the advantage that it was 1.52% less likely to miss a diabetes case than DT (Sensitivity: 77.98%). NB performed worst among the classifiers, and it may have the issue of missed diagnosis as its precision score is less than 70%.

The Precision–Recall curve is perfect when it is right-angle or perpendicular. The ROC-AUC curve and Precision–Recall curve for RF in Fig. 19 are highly perpendicular, showing the model to be a good fit. The AUC score of the RF classifier is 0.8226, and its hamming loss of 0.1774 reaffirms that it is comparatively better than other classifiers to predict the disease.

5. Conclusions

5.1. Research contributions

The key contribution of our research was the development of predictive models that used ML to detect people who were developing diabetes. This work presented a study of five classifiers (DT, RF, KNN, LR and NB) for predicting the likelihood of diabetes. The RF classifier achieved the highest accuracy of 82.26%.

This research assessed the prediction of diabetes based on the key features. With the enhanced capability of the ML algorithms in classification, the model can significantly aid medical practitioners in the diagnosis.

5.2. Limitations of the study and future work

One future direction of this study is to identify the gene and clinical variables that affect diabetes. A more suitable dataset can give more in-depth information on the variables that can be helpful in predicting the disease in a better way. In order to achieve this, cross-validation techniques should be used to improve the performance of the model and include SHAP values to understand which features have the highest impact on the outcome variable. In order to compete with state-of-the-art models achieving above 90% accuracy, the implementation of ensemble models should be utilized.

5.2.1. Federated machine learning

Originally proposed by Google, the idea behind federated machine learning (FML) is to construct ML models from data that is distributed across several devices or servers to prevent data leakage [42]. This allows for a more collaborative and decentralized approach while also reducing distribution costs, data distribution imbalances and reliability. While employing similar strategies to distributed machine learning, FML places a heavier emphasis on data privacy and protection, particularly during the stage of training the model. Additionally, federated database systems share a lot of similarities in methodology but put more of an emphasis on database operations as opposed to protecting data privacy. While FML is not associated with improving the performance of ML models, particular attention needs to be paid to approaches that strive to improve data privacy. In the age of Big Data and rapidly improving information systems, this is true more than ever, and its importance will only continue to increase. Ultimately, utilizing FML massively increases data privacy. In projects similar to ours, whereby sensitive patient data is analyzed, FML should be employed when appropriate to ease people's concerns in this regard.

5.2.2. ADASYN

Adaptive synthetic sampling (ADASYN) is a sampling approach for learning from imbalanced datasets. Similar to SMOTE, the purpose of ADASYN is to address class imbalance within datasets to reduce bias associated with imbalanced data and shift the classification decision boundaries towards the more difficult data points. In order to achieve this, data from the minority class is adaptively generated, which is harder to learn. In the initial ADASYN proposal [43], performance was compared against SMOTE across a standardized list of datasets. ADASYN was found to outperform SMOTE when the same ML models were used after addressing the data imbalance with these two techniques. Furthermore, it was noted that to improve performance further, ADASYN should be integrated into ensemble ML classifiers by using bootstrap sampling techniques and then embedding ADASYN into the sampled data.

5.2.3. Ensemble classifiers

Ensemble ML classifiers are a more recent development and are well-known for increased predictive performance compared to traditional approaches. Ensemble classifiers combine algorithms from two or more ML models to outperform each of the individual models. The most common technique used in ensemble classifiers is boosting. The aim of boosting is to improve the performance of weak learning algorithms, such as classification rules or decision trees [44]. AdaBoost and its variants are a commonly used set of iterative boosting algorithms which forces weak learning models to focus on the more complex data points by performing more iterations and producing additional classifiers [45]. Bootstrap aggregating, or bagging, is another common component of such classifiers, whereby composite classifiers are created by combining the outputs of the various models to produce

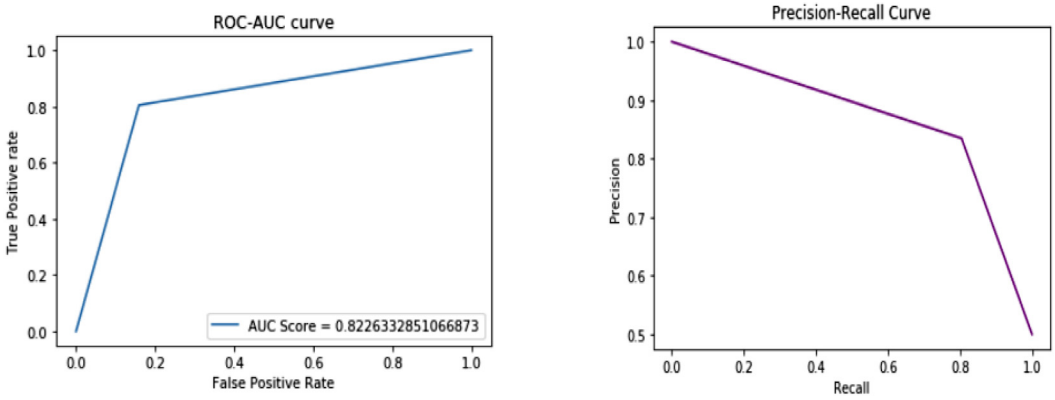

Fig. 19. ROC-AUC & Precision–Recall curve for RF classifier.

Table A.1
Description of dataset.

S. No	Features	Description
1	Diabetes binary	0 = no diabetes, 1 = diabetes
2	HighBP	0 = no high BP 1 = high BP
3	HighChol	0 = no high cholesterol 1 = high cholesterol
4	CholCheck	0 = no 1 = yes
5	BMI	Body Mass Index
6	Smoker	Have you smoked at least 100 cigarettes in your entire life? 0 = no 1 = yes
7	Stroke	You had a stroke. 0 = no 1 = yes
8	HeartDiseaseorAttack	Coronary Heart Disease (CHD) or myocardial infraction (ML) 0 = no 1 =yes
9	PhysActivity	physical activity in past 30 days - not including job 0 = no 1 = yes
10	Fruits	Consume Fruits 1 or more times per day 0 = no 1 = yes
11	Veggies	Consume Vegetables 1 or more times per day 0 = no 1 = yes
12	HvyAlcoholConsump	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes
13	AnyHealthcare	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes
14	NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost ? 0 = no 1= yes
15	GenHlth	Would you say that in general your health is ? Scale 1–5 1= excellent 2 = very good 3 = good 4 = fair 5 = poor
16	MentHlth	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good ? scale 1–30 days
17	PhysHlth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good ? scale 1–30 days
18	DiffWalk	Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
19	Sex	0 = female 1 = male
20	Age	13-level age category (_AGEG5YR see codebook) 1 = 18–29 9 = 60–64 13=80 or older
21	Education	Education level (EDUCA see codebook) scale 1–6 1 = Never attended school or only kindergarten 2 = Grades through 8 (Elementary) 3 = Grades 9 through 11 (some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)
22	Income	Income scale (INCOME2 see codebook) scale 1–8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more

a more robust prediction. It can be hypothesized that implementing ensemble classifiers and associated techniques would improve predictive performance, as other research extensively suggests this to be the case. In summary, to improve upon the results conducted in this study, ADASYN-based ensemble classifiers can be developed, fine-tuned, and optimized to build a more effective automated diabetes diagnosis system.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgment

This work is partly supported by VC Research, UK (VCR 0000191) for Prof Chang.

Appendix

See Table A.1.

References

[1] A. Stokes, S.H. Preston, Deaths attributable to diabetes in the United States: comparison of data sources and estimation approaches, PLoS One 12 (1) (2017) e0170219.

[2] M.C. Riddle, W.H. Herman, The cost of diabetes care—an elephant in the room, Diabetes Care 41 (5) (2018) 929–932.

[3] J. Elflein, Estimated number of diabetics worldwide in 2021, 2030, and 2045, 2022, [online] Available at: https://www.statista.com/statistics/271442/number-of-diabetics-worldwide/ (accessed 27 Sep, 2022).

[4] J. Chaki, S.T. Ganesh, S.K. Cidham, S.A. Theertan, Machine learning and artificial intelligence based diabetes mellitus detection and self-management: a systematic review, J. King Saud Univ.-Comput. Inf. Sci. (2020).

[5] H. Kaur, V. Kumari, Predictive modelling and analytics for diabetes using a machine learning approach, ACI 18 (1/2) (2020) 90–100, http://dx.doi.org/10.1016/j.aci.2018.12.004.

[6] H. Lu, S. Uddin, F. Hajati, M.A. Moni, M. Khushi, A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus, Appl. Intell. 52 (3) (2022) 2411–2422, http://dx.doi.org/10.1007/s10489-021-02533-w.

[7] M.W. Nadeem, H.G. Goh, V. Ponnusamy, I. Andonovic, M.A. Khan, M. Hussain, A fusion-based machine learning approach for the prediction of the onset of diabetes, Healthcare 9 (10) (2021) 1393, http://dx.doi.org/10.3390/healthcare9101393.

[8] M.A. Sarwar, N. Kamal, W. Hamid, M.A. Shah, Prediction of diabetes using machine learning algorithms in healthcare, in: 2018 24th International Conference on Automation and Computing (ICAC) Newcastle Upon Tyne, United Kingdom, 2018, pp. 1–6, http://dx.doi.org/10.23919/IConAC.2018.8748992.

[9] N. Sneha, T. Gangil, Analysis of diabetes mellitus for early prediction using optimal features selection, J. Big Data 6 (1) (2019) 13, http://dx.doi.org/10.1186/s40537-019-0175-6.

[10] M.K. Hasan, M.A. Alam, D. Das, E. Hossain, M. Hasan, Diabetes prediction using ensembling of different machine learning classifiers, IEEE Access 8 (2020) 76516–76531.

[11] D. Sisodia, D.S. Sisodia, Prediction of diabetes using classification algorithms, Procedia Comput. Sci. 132 (2018) 1578–1585.

[12] K. Rajesh, V. Sangeetha, Application of data mining methods and techniques for diabetes diagnosis, Int. J. Eng. Innov. Technol. (IJEIT) 2 (3) (2012).

[13] A.V. Kelarev, A. Stranieri, J.L. Yearwood, H.F. Jelinek, Empirical study of decision trees and ensemble classifiers for monitoring of diabetes patients in pervasive healthcare, in: 2012 15th International Conference on Network-Based Information Systems, IEEE, 2012, pp. 441–446.

[14] S.M. Ganie, M.B. Malik, An ensemble machine learning approach for predicting type-II diabetes mellitus based on lifestyle indicators, Healthcare Anal. 2 (2022) 100092.

[15] L. Han, S. Luo, J. Yu, L. Pan, S. Chen, Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes, IEEE J. Biomed. Health Inf. 19 (2) (2014) 728–734.

[16] M.M. Hassan, S. Mollick, F. Yasmin, An unsupervised cluster-based feature grouping model for early diabetes detection, Healthcare Anal. (2022) 100112.

[17] J. Ramesh, R. Aburukba, A. Sagahyroon, A remote healthcare monitoring framework for diabetes prediction using machine learning, Healthcare Technol. Lett. 8 (3) (2021) 45–57.

[18] J.J. Khanam, S.Y. Foo, A comparison of machine learning algorithms for diabetes prediction, ICT Express 7 (4) (2021) 432–439.

[19] R. Krishnamoorthi, S. Joshi, H.Z. Almarzouki, P.K. Shukla, A. Rizwan, C. Kalpana, B. Tiwari, A novel diabetes healthcare disease prediction framework using machine learning techniques, J. Healthcare Eng. (2022).

[20] U. Ahmed, G.F. Issa, M.A. Khan, S. Aftab, M.F. Khan, R.A. Said, T.M. Ghazal, M. Ahmad, Prediction of diabetes empowered with fused machine learning, IEEE Access 10 (2022) 8529–8538.

[21] P. Goyal, S. Jain, Prediction of type-2 diabetes using classification and ensemble method approach, in: 2022 International Mobile and Embedded Technology Conference (MECON), IEEE, 2022, pp. 658–665.

[22] N. Abdulhadi, A. Al-Mousa, Diabetes detection using machine learning classification methods, in: 2021 International Conference on Information Technology, ICIT, IEEE, 2021, pp. 350–354.

[23] H. Gupta, H. Varshney, T.K. Sharma, et al., Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction, Complex Intell. Syst. 8 (2022) 3073–3087, http://dx.doi.org/10.1007/s40747-021-00398-7.

[24] S. Sivaranjani, S. Ananya, J. Aravinth, R. Karthika, Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction, in: 2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS, Vol. 1, IEEE, 2021, pp. 141–146.

[25] L. Lama, O. Wilhelmsson, E. Norlander, L. Gustafsson, A. Lager, P. Tynelius, et al., Machine learning for prediction of diabetes risk in middle-aged Swedish people, Heliyon 7 (7) (2021) e07419.

[26] M.A.R. Refat, M. Al Amin, C. Kaushal, M.N. Yeasmin, M.K. Islam, A comparative analysis of early stage diabetes prediction using machine learning and deep learning approach, in: 2021 6th International Conference on Signal Processing, Computing and Control, ISPCC, IEEE, 2021, pp. 654–659.

[27] A.M. Malik, A.K. Sagar, S. Sahana, Prediction of cardiopathy using exploratory data analysis, in: 2021 IEEE 6th International Conference on Computing, Communication and Automation, ICCCA, IEEE, 2021, pp. 117–122.

[28] A. Thakkar, R. Lohiya, Attack classification using feature selection techniques: a comparative study, J. Ambient Intell. Humaniz. Comput. 12 (1) (2021) 1249–1266.

[29] C.L. Chowdhary, D.P. Acharjya, Segmentation and feature extraction in medical imaging: a systematic review, Procedia Comput. Sci. 167 (2020) 26–36.

[30] H.B. Harvey, S.T. Sotardi, The pareto principle, J. Am. College Radiol. 15 (6) (2018) 931.

[31] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357.

[32] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdiscip. Rev. Comput. Stat. 2 (4) (2010) 433–459.

[33] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).

[34] T. Sharma, M. Shah, A comprehensive review of machine learning techniques on diabetes detection, Vis. Comput. Ind., Biomed. Art 4 (1) (2021) 30, http://dx.doi.org/10.1186/s42492-021-00097-7.

[35] World Health Organization, Diagnostic Criteria and Classification of Hyperglycaemia First Detected in Pregnancy (No. WHO/NMH/MND/13.2), World Health Organization, 2013.

[36] G.O. Campos, A. Zimek, J. Sander, R.J.G.B. Campello, B. Micenková, E. Schubert, I. Assent, M.E. Houle, On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study, Data Min. Knowl. Discov. 30 (4) (2016) 891–927, http://dx.doi.org/10.1007/s10618-015-0444-8.

[37] J.W. Lee, J.B. Lee, M. Park, S.H. Song, An extensive comparison of recent classification tools applied to microarray data, Comput. Statist. Data Anal. 48 (4) (2005) 869–885, http://dx.doi.org/10.1016/j.csda.2004.03.017.

[38] L. Zwaan, H. Singh, The challenges in defining and measuring diagnostic error, Diagnosis 2 (2) (2015) 97–103.

[39] A. Swift, R. Heale, A. Twycross, What are sensitivity and specificity? Evidence-Based Nursing 23 (1) (2020) 2–4.

[40] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, X. Gao, Predictive models for diabetes mellitus using machine learning techniques, BMC Endocr. Disord. 19 (1) (2019) 101, http://dx.doi.org/10.1186/s12902-019-0436-6.

[41] L. Zhang, Y. Wang, M. Niu, C. Wang, Z. Wang, Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study, Sci. Rep. 10 (1) (2020) http://dx.doi.org/10.1038/s41598-020-61123-x, Art. (1).

[42] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Trans. Intell. Syst. Technol. 10 (2) (2019) 1–19.

[43] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IEEE, 2008, pp. 1322–1328.

[44] L. Rokach, Ensemble-based classifiers, Artif. Intell. Rev. 33 (1) (2010) 1–39.

[45] A. Shahraki, M. Abbasi, Ø. Haugen, Boosting algorithms for network intrusion detection: A comparative evaluation of real AdaBoost, gentle AdaBoost and modest AdaBoost, Eng. Appl. Artif. Intell. 94 (2020) 103770.