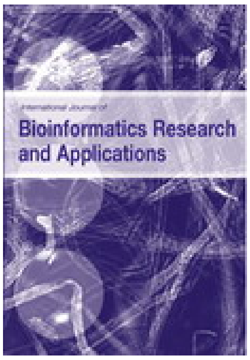


International Journal of Bioinformatics Research and Applications

ISSN online: 1744-5493 - ISSN print: 1744-5485
https://www.inderscience.com/ijbra

Predictive analysis for diabetes mellitus prediction using supervised techniques

Salliah Shafi Bhat, Madhina Banu, Gufran Ahmad Ansari

DOI: 10.1504/IJBRA.2024.10062801

Article History:

Received:	16 June 2023
Last revised:	20 June 2023
Accepted:	19 July 2023
Published online:	14 March 2024

Predictive analysis for diabetes mellitus prediction using supervised techniques

Salliah Shafi Bhat* and Madhina Banu

B.S. Abdur Rahman Crescent Institute of Science and Technology,
Chennai – 48, India
Email: salliahshafi678@gmail.com
Email: madhinamca05@gmail.com
*Corresponding author

Gufran Ahmad Ansari

Faculty of Science,
Dr. Vishwanath Karth MIT World Peace University (MIT-WPU),
Pune – 411038, India
Email: gufran.ansari@mitwpu.edu.in

Abstract: Diabetes mellitus is a silent disease. The worldwide prevalence of diabetes is rising quickly. Implementing lifestyle modifications and adopting necessary preventative measures for early detection of diabetes can help to avoid the development of diabetes. In this scenario, there is a need for simple, rapid, and accurate diagnostic methods. Various research is being carried out to increase the efficiency, effectiveness, dependability, and precision of these methods for identifying certain illnesses. Using diagnostic data from a dataset from the University of California Irvine (UCI) this research aims to examine whether a patient has diabetes or not. To predict diabetes four distinct machine learning algorithms (MLAs) have been used. With an accuracy rate of 98% the random forest was the most effective method. Other algorithms' accuracy rates range from 96% to 89% as well. In this paper machine learning prediction method for identifying diabetes patients is described. The author proposed a framework for early prediction of diabetes mellitus.

Keywords: diabetes mellitus; framework; machine learning; patient; accuracy.

Reference to this paper should be made as follows: Bhat, S.S., Banu, M. and Ansari, G.A. (2024) 'Predictive analysis for diabetes mellitus prediction using supervised techniques', *Int. J. Bioinformatics Research and Applications*, Vol. 20, No. 1, pp.78–96.

Biographical notes: Salliah Shafi Bhat have received his Master's in Information Technology from the Baba Ghulam Shah Badshah University Rajouri J&K and is currently pursuing PhD in Computer Applications from the BSA Crescent Institute of Science and Technology, India with specialisation in machine learning analytics. Her current research interests include the machine learning, artificial intelligence, and healthcare analytics.

Madhina Banu is currently working as an Assistant Professor in the Department of Computer Science and Engineering of B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamil Nadu, India. She completed her ME in Computer Science and Engineering from the Anna University, Chennai,

Tamil Nadu, India, in 2008 and Master of Computer Application (MCA) from the Barathidasan University Trichy, Tamil Nadu, India in 2006. She has published three articles in journal, international conferences, seminars, and workshops. She attended many national conferences, workshops, and faculty development programs. Her areas of interest are cloud computing, internet of things, and data mining, computer forensics, theory of computation, mobile and pervasive computing, and software testing.

Gufran Ahmad Ansari is a Professor in Department of Computer Science at Dr. Vishwanath Karth, MIT World Peace University Maharashtra, India. He also received his MCA from Dr. B.R. Ambedkar University Agra. He received his PhD from the Babasaheb Bhimrao Ambedkar. He has more than 20 years of experience in teaching. He has contributed to more than 60 research articles and conferences in reputed journals. He has published five book chapters and ten patents. He has also completed seven funded projects and has membership in more than 20 professional bodies. His research area is software engineering, machine learning, artificial intelligence, etc.

1 Introduction

Diabetes is a chronic condition that develops whenever the pancreatic secretion organ either fails to make enough insulin or refuses to use the insulin that is generated properly (Kabir et al., 2022). The patient's blood sugar level rises because they are unable to utilise the sugar that has been obtained from their food in an effective way. Diabetes commonly known as diabetes mellitus is a disorder in which the body does not produce enough insulin to enable the pancreas to break down blood sugar. It is the non-communicable chronic illness with the fastest rate of growth in the 21st century. It is based on heredity. Diabetes causes a variety of health issues including heart disease, stroke, nerve diseases, gallstones, eye loss, and kidney disease which significantly affect the person. Although obesity and inactivity are major contributors to diabetes their causes are never fully understood. In addition to height, weight, and inherited factors, diabetes is also influenced by all other factors that affect the blood sugar level such as nutrition (Bhat and Ansari, 2022). By 2045, 693 million people are predicted (Cho et al., 2018). According to various statistical studies (Saeedi et al., 2019) nearly half a billion individuals worldwide today suffer from diabetes which is a serious condition. According to predictions, the percentage could reach 25% of the total population in 2030 and 51% in 2045. Why does diabetes develop? The majority of the food we consume gets converted to glucose and absorbed into the circulation. The high insulin hormone aids in controlling the blood sugar level. This is accomplished by allowing blood glucose to enter your immune tissues wherever it will be used as nourishment. In the body, insulin is released by the pancreas. Over time this leads to serious consequences because too much blood sugar remains in the blood. Diabetes is incurable. However, diabetes can be controlled with regular exercise, a healthy diet, and medication use as recommended. Table 1 lists the basics of three forms of diabetes: type 1 diabetes (TID), type 2 diabetes (T2D), and gestational diabetes. Another type of diabetes known as pre-diabetes is thought to occur before T2D. In opposite to T2D the patient has a higher-than-normal blood glucose level when suffering from this disease (Ansari and Bhat, 2022). Pre-diabetes people are more likely to develop T2D when certain conditions and precautions are performed. Dangerous complications from diabetes could include heart disease, stroke, visual loss, disability, etc. The current methods for detecting diabetes involve lab tests to examine oral glycemic control and plasma glucose. Diabetes must be identified and treated as soon as possible in order to reduce its effects and continue treatment life quality. Early detection of diabetes can reduce its effects in adults, so the healthcare sector should benefit from this research. Machine learning are using data gathering methods to help health providers diagnose diseases in their early stages since they have access to the specific data in patient data. Using algorithms and statistics, data collection is the study of complex data to identify hidden and undetected patterns, correlations, and information that is challenging to identify (Muhammad et al., 2020). Diabetic patients may profit from this early detection and treatment due to the application of machine learning algorithms (MLAs) and the accessibility of active data. As machine learning enables a system or device to continuously learn from its past experiences and improve without explicit programming from other machines. The basic goal of machine learning is to create computers that can examine data and then use it as a component of their learning environment. As a result, machine learning allows machines to independently learn new things and modify their behaviour without the need for human input (Ghillani, 2022).

Table 1 Diabetes description

Types of diabetes	*Description*	*Characteristics*
T1D	This type of diabetes happens when the pancreas in the body stops producing insulin thus patients must take synthetic glucose every day.	This type is often known as juvenile diabetes because it is typically diagnosed in youngsters.
T2D	This type of diabetes develops when the body begins to react more slowly to the body's own produced insulin. This typically occurs as a result of weight gain and a lack of physical activity.	90% of those who have diabetes of T2D.
Gestational diabetes	Typically this type of diabetes occurs during the pregnancy of women.	There is a possibility that both the mother and child could develop diabetes T2D as an outcome.

1.1 Machine learning in healthcare

The technique of teaching machines to extract and recognise patterns using data and an algorithm is known as a subfield of artificial intelligence (AI). By using historical and current data with a variety of examples and variables that are then normalised, and converted into algorithms to get the necessary results, MLA plays a significant role in predicting various diseases. In the healthcare sector, a variety of chronic diseases can be identified, diagnosed, and forecasted using machine learning and deep learning. Communicable diseases can also be predicted and treated and medical conditions can be mentioned using recommender systems.

1.2 Machine learning technique

In the healthcare sector, MLAs are frequently used for diagnosis detection, prediction, and treatment. Many healthcare systems for improving medical care use ML-developed algorithms. Supervised learning, unsupervised learning, and semi-supervised learning are the three main categories of machine learning approaches. In order to identify a doctor's behavioural responses and health difficulties, these algorithms first evaluate the data using statistical conclusions. These algorithms for instance observe changes in the patient's flexibility, changes in habits, and health difficulties that may affect everyday activities. They also identify and detect variations in patterns of eating and resting. Healthcare systems and smart healthcare applications can use the behavioural patterns identified by these algorithms to suggest lifestyle modifications to patients, as well as better treatments and healthcare methods.

1.2.1 Supervised machine learning

It is a unique type of machine learning technique that is employed to predict outcomes and make decisions based on data. This method is used to find relationships among sample data and train from labelled data in order to predict events in the future. In supervised learning predicting is a unique type of machine learning technique that is employed to predict outcomes and make decisions based on data. For classification and regression issues, supervised learning models are frequently used in classification, the target variable is categorical. In regression the target variable is numeric. Although supervised learning is certainly effective, it has the problem of demanding many labelled data to create a large-scale dataset (Acharya et al., 2022). For example, Weather forecasting, disease prediction, fraud detection, score prediction, and risk assessment are a few examples of the areas where supervised machine learning is frequently employed.

1.2.2 Unsupervised machine learning

It combines supervised and unsupervised learning methodologies to create ML models. For the purpose of creating machine learning models, this method is used to analyse both labelled and unlabeled data. With this approach, a model is trained using some labelled data before being applied to the remaining unlabeled data to provide predictions. The data labelling process takes time and involves human efforts. Combining both methods to improve the classification performance and prediction capabilities of models is the primary goal of semi-supervised learning (Taghizadeh-Mehrjardi et al., 2022). Semi-supervised learning is widely used in fields like web mining which classifies web pages, text classification which locates names in text, video extraction, and which classifies persons in media articles.

1.2.3 Semi-supervised learning

The training samples of data are not labelled or categorised. The process helps to identify data and extracts hidden patterns. The many clusters are produced as output of the unsupervised learning process which uses unlabeled data as its input. These clusters can only form because of the relationships between the data samples. A cluster is formed from all of the samples and variables that strongly correlate with one another. The data samples that show little correlation with one another are grouped into various clusters. The two main categories of these algorithms are clustering and association algorithms namely

- *Clustering:* It is utilised when we wish to identify the built-in groupings in the data.
- *Association:* The algorithm is employed whenever you want to build rules that may represent a significant correlation between data samples, such as the chance that someone who purchases A will also purchase B.

The contribution of the paper is that author uses the framework for analysing the data for DM disease prediction.

Data has been collected through online Kaggle from UCI. The K nearest neighbour (KNN), support vector classification (SVC), random forest (RF), and Naïve Bayes (NB) machine learning classifiers have been used to develop models for binary classification problems. Before developing the machine learning models exploratory data analysis was carried out to ensure the data quality evaluation. These classifiers were run on the Windows operating system utilising Python 3.8 with the Jupyter Integrated Development Environment in the open-source framework Anaconda 2021. Regarding a number of statistical metrics, including precision, recall, F1-score, false positive (FP) rate, false negative (FN) rate, etc. The results obtained from the experimental investigation are acceptable. To minimise hospitalisation, useless treatments, and the cost of lab tests this research will examine the DM disease's parameters A will purchase B.

The major objective of paper includes

- In this study, the framework for analysing the data for DM disease prediction has been collected through online Kaggle from the UCI.
- KNN, SVC, NB, and RF machine learning classifiers have been used to develop models for binary classification problems.
- Before developing the machine learning models, exploratory data analysis was carried out to ensure the data quality evaluation. These classifiers were run on the Windows operating system utilising Python 3.8 with the Jupiter Integrated Development Environment in the open-source framework Anaconda 2021.

The below are the various sections that make up this research. The previous literature review in DM is represented in Section 2 of the literature review. Section 3 is about the methodology for the early prediction of diabetes. Section 4 result and discussion are discussed in this paper and finally, in Section 5 conclusion and future scope explain the future study of this work.

2 Literature review

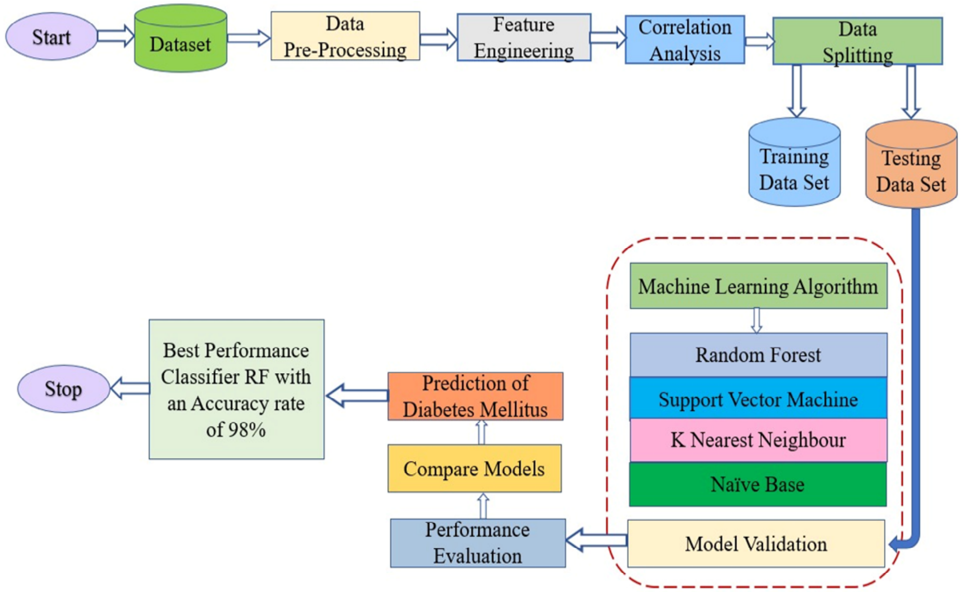
Several models that have been employed by various researchers to examine the classification of diabetes data are given below. Radial basis function, polynomial and linear kernels were compared to the conventional QDA, LDA, and NB in the work by Maniruzzaman et al. (2017). To determine the optimal cross-validation methodology, the authors also carried out extensive testing. The research shows that the K = 10 cross-validation technique with GP – the most effective classifier for predicting diabetes is one that is based. In terms of accuracy (81.97%), sensitivity (91.79%), specificity (63.33%), and positive and negative predictive values, GP-based models often perform worse than other models (84.91% and 62.50%, respectively). Different parameters from the dataset were utilised by Komi et al. (2017) to describe various categorisation strategies. For the diabetes prediction in this study, a modest data sample was used. The parameter of maternity, however, was not taken into account by the researchers when predicting diabetes. The five algorithms utilised were SVM, GMM, EM, ANN, and logistic regression. And the researchers came to the conclusion that artificial neural networks, or ANNs, had the best accuracy for diabetes prediction. Multilayer perception, Hoeffding tree (HT), RF, Jrip, BayeNet, and decision tree algorithms were utilised by Mercaldo et al. (2017) to make predictions. The best progressive feature selection approach for feature selection is used by the researchers in this experiment. The research comes to the conclusion that HTs provide good accuracy. The HT method model employed in this study makes use of real-world data and achieves precision values of 0.770 and recall values of 0.775. To accurately predict the possibility of diabetes, Vyas et al. (2018) applied three distinct machine learning classifiers, SVM, NB and DT. With an AUC of 0.819, they were able to prove that NB is the model that performs the best. The median value in the study by Hasan et al. (2020) was chosen instead of the mean value due to the fact that the median value has a less pronounced central tendency toward the attribute distribution mean. They have utilised a variety of ML classifiers, including MLP and KNN, RF, decision trees, NB, and AdaBoost. In order to improve diabetes prediction they have additionally combined ML models and used an ensemble classifier. Since AUC is unaffected by the class distribution it is decided that this measure will be used to weight the ML model in place of accuracy. The classifier suggested in this study performs well with a sensitivity of 0.789, specificity of 0.934, false omission rate of 0.092, diagnostic odds ratio of 66.234 and AUC of 0.950, outperforming prior results in this field by 2% of AUC. Two distinct datasets, PIDD and 130 US hospital diabetes datasets were used by Alehegn et al. (2019). KNN, RF, J48, and NB are the ML methods utilised in this paper. The suggested method offers superior accuracy of 93.62% when using combining Meta classifiers in the instance of PIDD. Predictive analysis is used in the strategy put forth by Sneha and Gangil (2019) to identify the features that assist in the early detection of diabetes. The comparison shows that the decision tree model and RF both have the highest specificity scores, with 89.2% and 90% respectively, therefore these models are the best for analysing diabetes. With 82.30% Naïve Bayesian provides the highest accuracy value. A PIMA Indian dataset was analysed by Hina et al. (2017) using a variety of classification methods including NB, RF, J48, logistic regression, and zero R. Diabetes is predicted by compared. For the DM disease, Tigga and Garg (2020) created a prediction model. For the investigation, a dataset with 952 instances and 18 attributes was obtained. It also made use of the PIMA dataset. The classifiers for machine learning that are utilised are RF, LR, KNN, SVM, NB and DT. With a percentage of 87.10% for collected data and 75% for the PIMA dataset, the accuracy obtained for RF was the highest. Two datasets, including PIDD and a breast cancer dataset, were used by Kumari et al. (2021) and were taken from the UC Irvine (UCI) Repository. In order to make predictions, three ML classifiers are used. They are the RF, LR, and NB models. With a percentage of 79.08% for PIMA data and 97.27% for breast cancer data using soft voting classifiers the accuracy for both datasets is the greatest. For the early-stage detection, classification, and prediction of diabetic disease, Butt et al. (2021) have suggested a machine learning-based strategy. The Indian PIMA dataset was utilised. RF, multilayer perception (MLP), and LR classifiers are employed. With a score of 87.26%, MLP has the best accuracy. In this research the implementation of four algorithm, viz., RF, support vector machine, KNN and NB and the results of the analysis were contrasted using different statistical metrics. After implementation and evaluation, RF had the greatest accuracy of all classifiers at 98%.

3 Methodology

3.1 Proposed model for early prediction of diabetes mellitus

The researcher techniques are highlighted in the methodology section. The research is divided into a number of sections that explain machine learning, data analysis, the datasets utilised, the suggested framework, and the methodologies used to evaluate the framework. In this research, machine learning and data analysis approaches are used to create a machine learning technique that can determine from a variety of diagnostic parameters whether a patient has diabetes or not. The proposed model is shown in Figure 1 for creating machine learning models for The UCI Machine Learning Repository gathered lifestyle dataset was used to create the ML models used in this framework. To improve the relevance assessment of the data, the pre-processing method has been used to identify outliers, eliminate inconsistent results, find missing values, and classify duplicate values. Independent and dependent characteristics have been has been identified by correlation analysis. To overcome under and overfitting issues in the data, the dataset was divided into a 70:30 ratio, where 70% of the data was used to train classifiers and 30% of the data was used to test/validate the models. Reliable results have been obtained by applying hyperparameter modification. More statistical computations have been made in order to verify the data and improve DM illness prediction.

Figure 1 Framework for proposed model (see online version for colours)



3.2 Dataset

The UCI Machine Learning Repository primarily published the dataset is used in this research. The dataset includes information on recently diagnosed or at-risk patients symptoms and signs of diabetes. The dataset includes 520 records of patients with 12 attributes. The dataset description and the relevant attributes are shown in Table 2. The twelve attributes are age, gender, thirst, insulin, skin thickness, family history of diabetes, smoker, glucose, fruits, BMI, blood pressure, and outcome. The outcome attribute target or dependent variable consists of two binary values: 1 and 0. A person's diabetes is indicated by a number between 0 and 1, where 1 means they have diabetes. The remaining characteristics take independent characteristics or variables into consideration.

Table 2 Attributes of UCI Machine Learning Repository

S. no.	*Attributes*	*Description*	*Min*	*Max*	*Mean*	*Std. dev.*
1	Age	Age of a person in years	16	90	48.02	12.15
2	Gender	A person is male or female	0	1	0.36	0.48
3	Thirst	How frequently a person drinks water throughout the day	0	1	0.49	0.5
4	Insulin	Insulin level of a person	0	1	0.44	0.49
5	Skin thickness	Skinfold thickness of triceps (mm)	0	1	0.41	0.49
6	Family history of diabetes	Whether any family person is diabetic or not	0	1	0.58	0.49
7	Smoker	Whether a person is a smoker or not	0	1	0.45	0.49
8	Glucose	Two-hour oral glucose tolerance test for evaluating plasma glucose levels	0	1	0.22	0.41
9	Fruits	Whether the person eats fruits or not	0	1	0.44	0.49
10	BMI	Body mass index (weight in kg)	0	1	0.48	0.5
11	Blood pressure	Blood pressure in mm (Hg)	0	1	0.24	0.42
12	Outcome	Diabetic or non-diabetic	…	…	…	…

3.3 Data pre-processing

Pre-processing is the transformation of unprocessed data into a compact and ordered dataset before it is analysed. Databases may have a variety of quality control problems. To enable accurate analysis, pre-processing aims to assess and improve the quality of the data (Ansari and Bhat, 2022). The transformations carried out on the data before analysis are referred to as pre-processing. This method converts raw data into a dataset that can be analysed. Various pre-processing procedures were used, including min-max, variance, deviation, a dataset that has undergone normalisation, mean scaling, and the removal of missing values. Additionally, variations were taken out of the dataset. Pre-processing revealed that certain records for the attributes in the data collection had the value '0', which could not be performed; isolated cases could be located in the data and some records had values. The mean values were used to replace any missing, unusual, or unavailable values of '0' in the data. As a result, an impossible-high dataset devoid of other manipulation was obtained.

3.3.1 Data visualisation

The histogram is one of the greatest methods for data visualisation that is more frequently employed in machine learning. It is related to splitting a continuous variable over a predetermined period of time or interval. Normal distribution, outliers, skewness, etc. are all represented by each distribution. Skewness can be evaluated to show which ranges deviate from the average.

Figure 2 displays the histogram distribution for several diabetes dataset attributes, including age, gender, thirst, insulin, skin thickness, family history of diabetes, smoker, glucose, fruits, BMI, blood pressure, and outcome.

- Despite select outlier values, plots of some characteristics, such as blood pressure, glucose, body mass index, and skin thickness are regularly distributed.
- Age, thirst, and family history of diabetes plots are skewed, among other features.
- Blood pressure and glucose levels increase a person's risk of getting diabetes.

3.4 Feature engineering

Selecting the most pertinent input factors that are most correlated with the target attribute is a technique for reducing the feature set. The predictive model performs better and costs less to compute and store as a result of having low resolution. The size of the feature set can be decreased with a variety of benefits. For instance, as the removal of useless features implies the reduction of noise, overfitting of the data can be reduced. Since the removal of misleading input variables, the model accuracy has also improved. Furthermore, the smaller amount of data implies that the model's construction and training required less time. The three kinds of traditional feature selection techniques used in supervised learning are filter, wrapper, and intrinsic techniques. The two feature selection techniques we employ in the current study are the filter-based method and the wrapper method. The feature engineering processes are built using various combinations of these strategies with the optimal number of features for feature extraction.

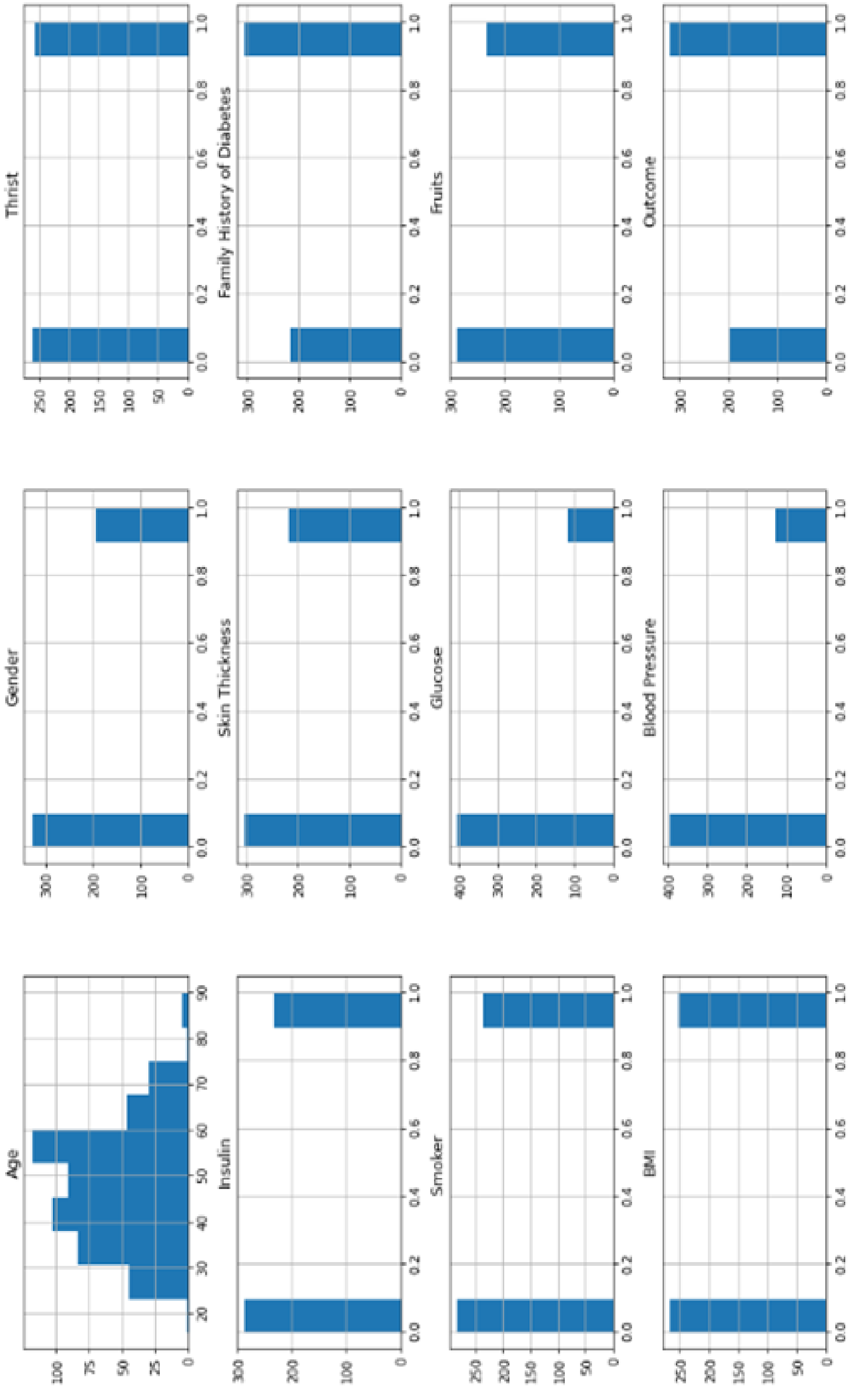
3.4.1 Filter-based method

The analysis of the statistical correlations between the input features and the target variable forms the foundation of filter-based techniques. The input features that have a positive people that have a high correlation with the target variable are chosen, while the others are excluded. Each characteristic is given a score based on its association with the target variable which is used to filter the data and choose the most relevant ones (Otchere et al., 2022).

3.4.2 Wrapper method

Wrapper approaches employ various feature set subsets and assess the effectiveness of an ML model for each subset. The subset with the greatest performance evaluation score is chosen from each subset. The best feature subset can be discovered using both systematic and unstructured methods. The first approach might make use of a best-first search strategy while the second approach could make use of a randomised algorithm.

Figure 2 Histogram of various dataset variables (see online version for colours)

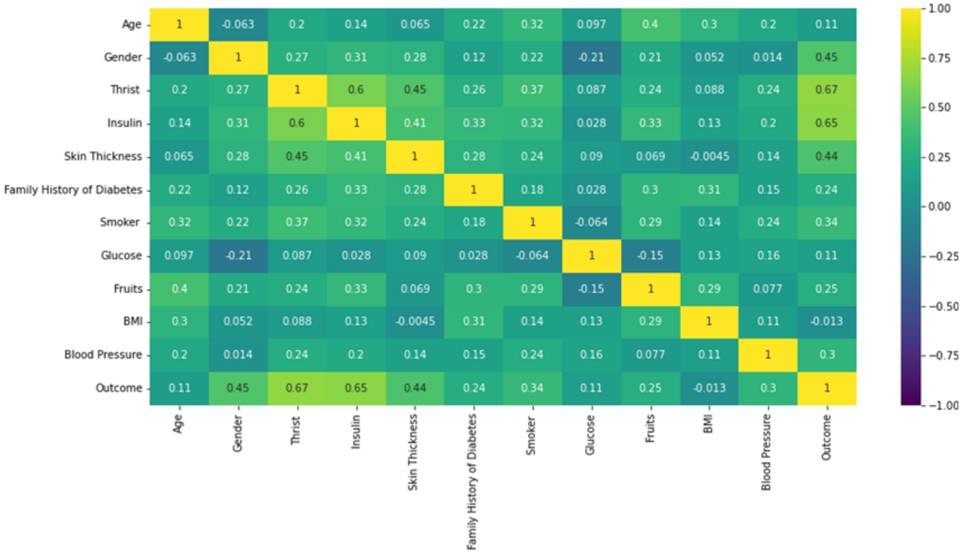


3.5 Correlation analysis

When generating a useful dataset overview, it is frequently simpler to take the relationship between features into care. A statistic known as correlation determines how strongly two variables move in relation to one another. The correlation map based on the diabetes dataset shows that when two variables move in opposite directions, there is a positive correlation and a negative correlation is shown in Figure 3.

The dataset is analysed and a heat map is used to show how the values are correlated. This leads to the conclusion that the most closely connected characteristics to the target variable are glucose, BMI, age, and blood pressure. In Figure 3 the correlation between ‘glucose’ and ‘outcome’ is 0.063 higher than that of any other feature. This suggests that if ‘glucose’ rises, diabetes may be developed.

Figure 3 Heat map showing the relationship between the variables (see online version for colours)



3.5.1 Training and testing dataset

The training dataset’s purposes are distinct from those of the test dataset. The primary improves in the model’s real classification while the latter aids in verifying the model’s results. Additional training data is usually beneficial because it gives the model more real-world examples on which to base its predictions. Additionally, more test data is beneficial for understanding the generalisability of the model being applied. Generally, 70% of the dataset is utilised for training and the remaining 30% is used for testing (Ullah et al., 2022). An imbalanced sample of positive and negative cases makes up the PIDD dataset. Once the models are finished, the stability of the models is evaluated using K-fold. The dataset is divided into K distinct subsets or folds as the initial step. K times are iteratively run through the model. The dataset is trained on the K – 1st fold for each iteration. With the test dataset, the Kth fold is evaluated. K has been given a value of 10 in this research. After the model is completed the dataset is divided into a train set and a test set. Final results were derived by averaging test sample values over ten iterations. Through hyperparameter tuning the top models that were chosen were improved and optimised. The confusion matrix that was reported based on model prediction is detailed in the result and discussion.

3.6 Machine learning technique

Algorithms estimated the dataset of RF, SVC, NB, and KNN. These methods were selected because they are extensively cited in the literature and when employed with the current dataset yield findings that are notably better.

- *RF:* The RF model is an ensemble model made up of uncut decision trees. The output of various DTs is combined by the RF to produce a single result. The DT is used as a base for row and column sampling. As the base learner numbers increase, the variety decreases, or the opposite arises. K is a viable option for cross-validation. It is regarded as a crucial bagging technique. The following formula for RF = DT (base learner) + bagging (row sampling with replacement) + feature bagging (column sampling) + aggregation (mean/median, majority vote).
- *SVC:* A popular statistically-based supervised machine learning technique used for regression and classification applications is the SVC. It is known to function with both linear and nonlinear data and is highly effective at overcoming the problems caused by dimension (Ghosh et al., 2019). Particularly with small datasets and high dimensional feature spaces, it performs well. By identifying a hyperplane with the largest margin. SVC splits training samples into discrete groups when working with linear data. It also determines the closest points to the margin edge, the support vectors, and the distance separating them from the n-dimensional hyperplane with the largest gap. Equation (1) the mathematical formula for maximising the margin is given below, where w is the weight vector, x is the input vector, and b is the bias.

  $$\text{Minimise} = \frac{1}{2}\|\text{w}\|1^2 \tag{1}$$

  Subject to $y_i((w \cdot x) + b) > 0$.

  SVM employs a kernel-based strategy when working with nonlinear data, utilising some kernel functions and the kernel trick to select the ideal hyperplane to linearly separate the data. The list of kernel functions that were looked at in this research to determine the optimum is provided below. Equation (2), where the linear kernel function is represented by the constant integer c.

  $$K\left(x_i, x_j\right) = x^{t}_{ix_j} \tag{2}$$

  The polynomial kernel function is shown in equation (3) where d is the degree of polynomial in the slope r is constant.

  $$K\left(x_i, x_j\right) = \left(yx^{t}_{ix_j} + r\right)^{d,y} > 0 \tag{3}$$

  The RBF kernel function is represented by equation (4), where is the gamma and exp (x_i, x_j^2) is the Euclidean distance between two points x_i and x_j.

$$K\left(x_i, x_j\right) = \exp\left(-y\left\|x_i - x_j^2\right\|\right), y > o \quad (4)$$

The Sigmoid kernel function is shown in equation (5), where r is a constant term and is the slop

$$K\left(x_i, x_j\right) = \tanh\left(yx_{ix_j}^{t} + r\right) \quad (5)$$

- *NB:* Supervised machine learning method uses the Bayes theorem to solve designated problem statements. The NB classifier is an easy-to-use classification method that makes use of the notion of conditional probability to speed up the creation of quick AI and ML models with improved predicting abilities (Bhat and Ansari, 2022). According to the statement, The Bayes theorem asserts that the link between the probability of the premise P(H) before gathering evidence and the probability of the premise P(H) after gathering evidence P(H|E) given premises H and evidence E. NB is depicted as in equation (6) as

$$P(E) = \frac{P(H) \cdot P(E)}{P(E)} \quad (6)$$

- *KNN:* One of the fundamental algorithms utilised in supervised machine learning methods is KNN. Its capability for counting allows it to categories similarities between newly available examples and data instances, all of which are based on assessments of class similarity. To determine a data point's nearest neighbour based on the value of K, the KNN algorithm's basic operating concept is to locate or access the data point with the shortest distance between them. Euclidean distance, Manhattan distance, and Murkowski distance can all be used to measure or calculate the distances in KNN classifiers. KNN is a simple and lazy classifier because it lacks intelligence mechanisms at various stages of the process (Abubakr, 2020).

3.7 Performance evaluation

The effectiveness of classification was measured using the following metrics.

Accuracy, precision, recall, F1-score, and confusion matrix:

- *Accuracy:* The ratio of successfully diagnosed diabetic patients to the total number anticipated is the measure of accuracy. The mathematical description of precision is seen in equation (7).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

- *Precision:* Equation (8) calculates precision as the proportion of correctly recognised diabetic individuals to all diabetic patients.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

- *Recall:* Equation (9) is used to compute recall which is the proportion of correctly identified diabetic patients to the total population of that class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

- *F1-score:* The symmetrical mean of the sensitivity and precision is the F1-score shown in equation (10)

$$\text{F1-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

- *Confusion matrix:* An overview of classification model performance is provided by a commonly referred to as ‘model performance’. The performance measure evaluates classification problems involving two or more classes in machine learning. The true positive (TP), FP, true negative (TN), and FN numbers make up a confusion matrix.

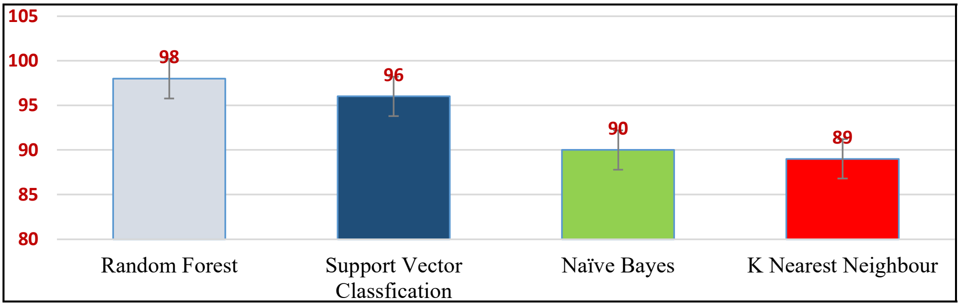
4 Result and discussion

The results of an analysis of exploratory data are presented and shown in this section. The health sector is still in the educational stage of healthcare parameter analysis predicting future health conditions for numerous chronic diseases based on these criteria. Every area of research and engineering especially in the field of healthcare is using MLA. Prior to developing machine learning, an inferential statistical correlational analysis of a dataset for diverse tasks is necessary. Exploratory data analysis is the term for this procedure (EDA). The relationship between the dependent and independent parameters utilised in the dataset has been determined through the examination of correlation coefficients. There is a connection between the dependent and independent variables of 1 and +1. Determine the relevance of the features in a dataset using correlation coefficient analysis (CCA). If the dependent and independent variables are well-correlated the feature set is identified as suitable for use in machine-learning models. The features used demonstrate a high and positive correlation between several independent or dependent attributes, such as BMI, gender, and other factors are closely associated. The dataset’s statistical metrics such as count, mean, standard deviation, minimum and maximum values, etc. have been calculated. For each attribute in the dataset various statistical measures have been listed using describe () function from the pandas library package. The accuracy of four supervised ML systems for DM prediction using a dataset is shown in Figure 4. Table 3 compares the datasets used in this study and the PIMA diabetes dataset to show how machine learning techniques can be utilised to more accurately diagnose diabetes mellitus.

Table 3 Comparison of machine learning model

Algorithms	*Accuracy with PIMA dataset*	*Accuracy with dataset used in this study*
RF	91.26% (Mujumdar and Vaidehi, 2019)	98%
SVM	89.16% (Refat et al., 2021)	96%
KNN	86% (Bhat et al., 2022)	89%
NB	87.16% (Rani, 2020)	90%

Figure 4 Accuracy of algorithms (see online version for colours)

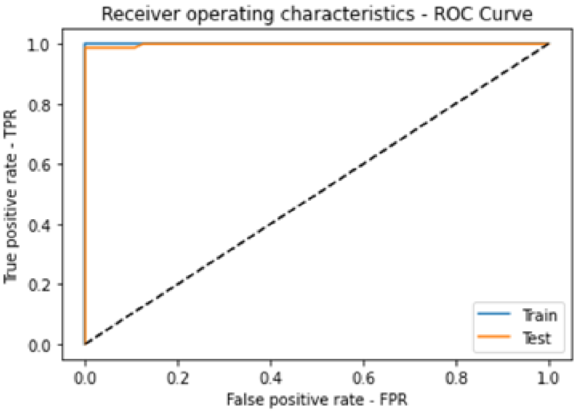


RF beat other classifiers in terms of performance achieving the highest accuracy of 98% followed by SVM, NB, and DT at 96%, 90, and 89%, respectively. To find mislabelled or erroneous DM illness predictions, the confusion matrix in Table 4 was employed. It contrasts actual values with anticipated values using four variables: TP, TN, FP, and FN stand for TP and TN, respectively (FN). FP (type 1) and FN (type 2) errors in this binary classification problem have been recognised and their measurements have been computed.

Table 4 Parameter evaluations using analytics

Classifier	*True positive*	*True negative*	*False positive*	*False negative*
RF	55	1	1	73
SVC	53	3	2	72
NB	50	6	6	68
KNN	53	3	9	65

Figure 5 ROC curve of RF (see online version for colours)



Binary classifier metrics are evaluated using receiver operating characteristic (ROC) at different criteria. The classifiers FP and TP rates are designed using the x- and y-axes. AUC is a metric that classifiers use to distinguish between the FPT and TPR. Figures 5–8 display the ROC curves for each classifier used to predict DM illness, illustrating the model’s ability to distinguish between the various categories.

Figure 6 ROC curve of SVC (see online version for colours)

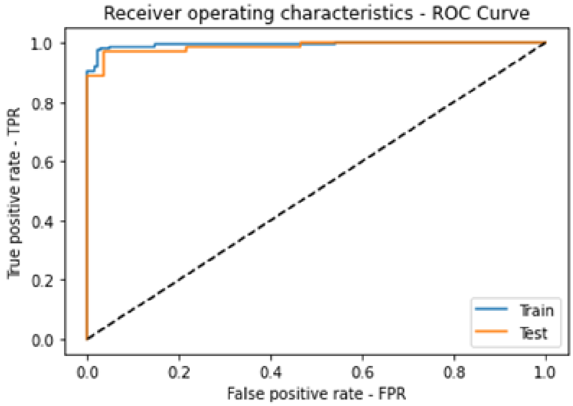


Figure 7 ROC curve of NB (see online version for colours)

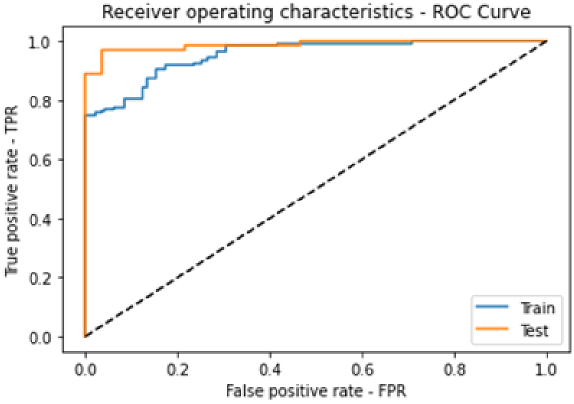


Figure 8 ROC curve of KNN (see online version for colours)

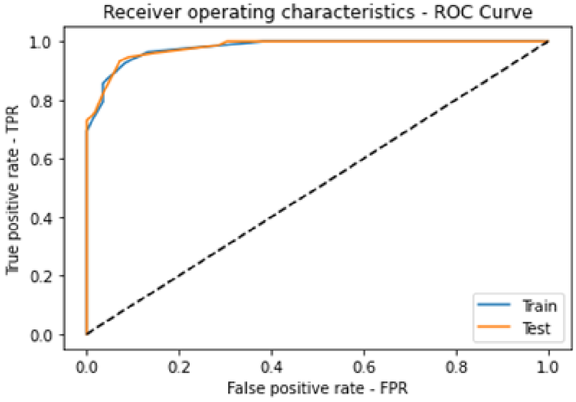
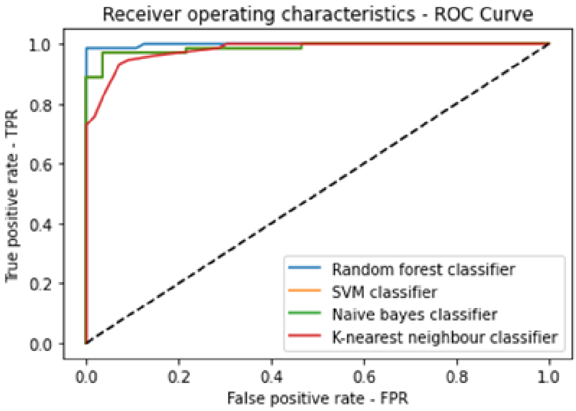


Figure 9 ROC curve for all algorithms (see online version for colours)



The ROC curve shown in Figure 9 proves that the final model is more effective.

5 Conclusions and future scope

MLT is a benefit to the healthcare industry since they have the ability to forecast outcomes by analysing the massive amounts of health data to get insightful knowledge and help practitioners and providers improve care. MLAs have demonstrated improved results when applied to a variety of statistical metrics for the diagnosis, prognosis, prediction, and detection of many chronic diseases. The determination of these methods in healthcare analytics is the exclusive subject of this paper. The essential role the practical advantages of machine learning paradigms in the field of healthcare. Along with their capacity to classify and predict diseases, many machine learning techniques have been discussed. This paper discusses some well-known MLAs for healthcare analytics based on their learning mechanisms. To reduce the suffering of unnecessary laboratory testing, etc. a framework has been described and put into practice for the improved prediction of DM disease. In order to expand the scope of the current research, these machine learning classifications will be applied to other related diseases that share data with diabetic disease. In order to improve the accuracy and dependability of the framework employed in this research, machine learning ensemble and hybridisation techniques will be applied. In future work, the model can be trained using fresh patient data. It is believed that as the amount of data rises, a better estimate can be generated.

References

Abubakr, J.M.B. (2020) 'Customer churn prediction in telecom using machine learning in big data platform', Vol. 18, No. 19, pp.567–574.

Acharya, S. et al. (2022) 'Analyzing milk foam using machine learning for diverse applications', *Food Analytical Methods*, Vol. 15, No. 12, pp.3365–3378.

Alehegn, M., Joshi, R.R. and Mulay, P. (2019) 'Diabetes analysis and prediction using random forest, KNN, Naïve Bayes and J48: an ensemble approach', *Int. J. Sci. Technol. Res.*, Vol. 8, No. 9, pp.1346–1354.

Ansari, G.A. and Bhat, S.S. (2022) 'Exploring a link between fasting perspective and different patterns of diabetes using a machine learning approach', *Educational Research*, Vol. 12, No. 2, pp.500–517.

Bhat, S.S. and Ansari, G.A. (2021) 'Predictions of diabetes and diet recommendation system for diabetic patients using machine learning techniques', *2021 2nd International Conference for Emerging Technology (INCET)*, IEEE.

Bhat, S.S. and Ansari, G.A. (2022) 'Prediction of diabetes mellitus using machine learning', *Machine Learning and Deep Learning in Efficacy Improvement of Healthcare Systems*, pp.93–108, CRC Press.

Bhat, S.S., Selvam, V., Ansari, G.A., Ansari, M.D. and Rahman, M.H. (2022) 'Prevalence and early prediction of diabetes using machine learning in North Kashmir: a case study of district Bandipora', *Computational Intelligence and Neuroscience*, Vol. 2022, Article ID 2789760.

Butt, U.M. et al. (2021) 'Machine learning based diabetes classification and prediction for healthcare applications', *Journal of Healthcare Engineering*, Vol. 20, pp.20–29.

Cho, N.H. et al. (2018) 'IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045', *Diabetes Research and Clinical Practice*, Vol. 138, pp.271–281.

Ghillani, D. (2022) *Deep Learning and Artificial Intelligence Framework to Improve the Cyber Security*, Authorea Preprints.

Ghosh, S., Dasgupta, A. and Swetapadma, A. (2019) 'A study on support vector machine based linear and non-linear pattern classification', *2019 International Conference on Intelligent Sustainable Systems (ICISS)*, IEEE.

Hasan, M.K. et al. (2020) 'Diabetes prediction using ensembling of different machine learning classifiers', *IEEE Access*, Vol. 8, pp.76516–76531.

Hina, S., Shaikh, A. and Sattar, S.A. (2017) 'Analyzing diabetes datasets using data mining', *Journal of Basic and Applied Sciences*, Vol. 13, No. 16, pp.466–471.

Kabir, M. et al. (2022) 'Therapeutic potential of dopamine agonists in the treatment of type 2 diabetes mellitus', *Environmental Science and Pollution Research*, Vol. 16, pp.1–20.

Komi, M. et al. (2017) 'Application of data mining methods in diabetes prediction', *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, IEEE.

Kumari, S., Kumar, D. and Mittal, M. (2021) 'An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier', *International Journal of Cognitive Computing in Engineering*, Vol. 2, pp.40–46.

Maniruzzaman, M. et al. (2017) 'Comparative approaches for classification of diabetes mellitus data: machine learning paradigm', *Computer Methods and Programs in Biomedicine*, Vol. 152, pp.23–34.

Mercaldo, F., Nardone, V. and Santone, A. (2017) 'Diabetes mellitus affected patients classification and diagnosis through machine learning techniques', *Procedia Computer Science*, Vol. 112, pp.2519–2528.

Muhammad, L.J., Algehyne, E.A. and Usman, S.S. (2020) 'Predictive supervised machine learning models for diabetes mellitus', *SN Computer Science*, Vol. 1, No. 5, p.240.

Mujumdar, A. and Vaidehi, V. (2019) 'Diabetes prediction using machine learning algorithms', *Procedia Computer Science*, Vol. 165, pp.292–299.

Otchere, D.A. et al. (2022) 'Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions', *Journal of Petroleum Science and Engineering*, Vol. 208, p.109244.

Rani, K.M.J. (2020) 'Diabetes prediction using machine learning', *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, Vol. 6, pp.294–305.

Refat, M.A.R., Al Amin, M., Kaushal, C., Yeasmin, M.N. and Islam, M.K. (2021) '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, pp.654–659.

Saeedi, P. et al. (2019) 'Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas', *Diabetes Research and Clinical Practice*, Vol. 157, p.107843.

Sneha, N. and Gangil, T. (2019) 'Analysis of diabetes mellitus for early prediction using optimal features selection', *J. Big Data*, Vol. 6, No. 10, p.13.

Taghizadeh-Mehrjardi, R. et al. (2022) 'Semi-supervised learning for the spatial extrapolation of soil information', *Geoderma*, Vol. 426, p.116094.

Tigga, N.P. and Garg, S. (2020) 'Prediction of type 2 diabetes using machine learning classification methods', *Procedia Computer Science*, Vol. 167, pp.706–716.

Ullah, N. et al. (2022) 'An efficient approach for crops pests recognition and classification based on novel DeepPestNet deep learning model', *IEEE Access*, July, Vol. 10, pp.73019–73032.

Vyas, A., Kadakia, U. and Jat, P.M. (2018) 'Extraction of professional details from web-URLs using DeepDive', *Procedia Computer Science*, Vol. 132, pp.1602–1610.