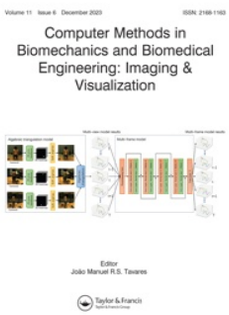

Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tciv20

Detection and prediction of diabetes using effective biomarkers

Mohammad Ehsan Farnoodian, Mohammad Karimi Moridani & Hanieh Mokhber

To cite this article: Mohammad Ehsan Farnoodian, Mohammad Karimi Moridani & Hanieh Mokhber (2024) Detection and prediction of diabetes using effective biomarkers, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 12:1, 2264937, DOI: 10.1080/21681163.2023.2264937

To link to this article: https://doi.org/10.1080/21681163.2023.2264937

Published online: 05 Oct 2023.

COMPUTER METHODS IN BIOMECHANICS AND BIOMEDICAL ENGINEERING: IMAGING & VISUALIZATION
2024, VOL. 12, NO. 1, 2264937
https://doi.org/10.1080/21681163.2023.2264937

Taylor & Francis
Taylor & Francis Group

Detection and prediction of diabetes using effective biomarkers

Mohammad Ehsan Farnoodian[a], Mohammad Karimi Moridani[b] and Hanieh Mokhber[b]

[a]Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran; [b]Department of Biomedical Engineering, Faculty of Health, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran

ABSTRACT

Diabetes is a prevalent and costly condition, with early diagnosis pivotal in mitigating its progression and complications. The diagnostic process often contends with data ambiguity and decision uncertainty, adding complexity to achieving definitive outcomes. This study addresses the diabetes diagnostic challenge through data mining and machine learning techniques. It involves training various machine learning algorithms and conducting statistical analysis on a dataset comprising 520 patients, encompassing both normal and diabetic cases, to discern influential features. Incorporating 17 features as classifier inputs, this research evaluates the diagnostic performance using four reputable techniques: support vector machine (SVM), random forest (RF), multi-layer perceptron (MLP), and k-nearest neighbor (kNN). The outcomes underscore the SVM model's superior performance, boasting accuracy, specificity, and sensitivity values of 98.78±1.96%, 99.28±1.63%, and 97.32±2.45%, respectively, across 50 iterations. The findings establish SVM as the preferred method for diabetes diagnosis. This study highlights the efficacy of data mining and machine learning models in diabetes diagnosis. While these methods exhibit respectable predictive accuracy, their integration with a physician's assessment promises even better patient outcomes.

ARTICLE HISTORY
Received 25 May 2023
Accepted 25 September 2023

KEYWORDS
Data mining; diabetes; SVM; detection; prediction

1. Introduction

The prevalence of diabetes among the world's population has increased over the decades to become one of the world's most dangerous diseases and among the fourth deadliest (Lin et al. 2020). Several complications are associated with diabetes, including damage to organs caused by defects in the metabolism of carbohydrates, fats, and proteins. As a result, it can lead to death in the worst-case scenario (Furman et al. 2019). Diabetic patients lose their lives every eight seconds, and diabetes affects one out of four people in the US. Each year about 1.6 million people are diagnosed with diabetes, and on average, 80 people per second die from diabetes (Kalra et al. 2018). In 2035, there will be around 500 million diabetics worldwide (Saeedi et al. 2019). There is no definite cure for this disease yet, and doctors and specialists are trying to control this disease with insulin and medicine. The best way, they believe, is prevention (Adu et al. 2019). Various prediction models have been implemented by researchers based on our healthcare data analysis. These models use data mining techniques, machine learning algorithms, computer simulation models, and a combination of these techniques (Wu et al. 2021). Data mining technology was used to build an innovative model for predicting type 2 diabetes that is stored in the knowledge discovery in databases, as well as a computational process to find patterns in large data sets presented in 2018 by Han Wu Sheng Qi et al (Han et al. 2018). The presented model uses artificial intelligence, machine learning, statistics, and database systems. The main goals of this method are pattern recognition, prediction, and clustering. There were many limitations to the first type 2 diabetes progression model (Sun et al. 2021), divided into units for cardiovascular patients, retinopathy, and neuropathy (Kuwata et al. 2017). In contrast, later simulation models were able to overcome some of these limitations.

In 2004, another model was presented based on the UK Prospective Diabetes Study (UKPDS), which predicts several complications related to diabetes using a system of equations. The important aspect of this model is the relationship between different types of difficulties at the patient's level (Clarke et al. 2004). Several researchers have used machine learning (ML) to predict diabetes using the Pima Indians diabetes dataset (PIDD). The PIDD dataset includes 9 features and 768 records describing female patients' conditions. The results of this method using the artificial neural network (ANN) algorithm on PIDD showed 88.6% accuracy (Jobeda and Foo 2021). Researchers found that AdaBoost can predict diabetes, coronary heart disease, and high blood pressure by analysing the Canadian primary care sentinel surveillance network (CPSSN) and using the decision tree model (Sharma and Shah 2021; Absar et al. 2022). Tigga et al. used logistic regression to predict diabetes using PIDD. This model showed a prediction accuracy is 75.32% (Tigga and Garg 2020). Sivashankari et al. classified diabetes based on patient information and treatment dimensions. They proposed an ensemble model and compared it with three algorithms: Naive Bayesian (NB), linear discrimination analysis (LDA), and logistic regression (LR). A stacked ensemble model has achieved 93.1% accuracy in predicting blood sugar disease (Sivashankari et al. 2022).

CONTACT Mohammad Karimi Moridani mkarimi.bme@gmail.com Department of Biomedical Engineering, Faculty of Health, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran

Additionally, Dritsas et al. have used medical data to predict diabetes. Also, k nearest neighbour (kNN), and random forest (RF) methods were used after data preprocessing for analysis (Dritsas and Trigka 2022). In another study, the patients and physicians knew about prevention in primary care, there was a clear relationship between predicted risk and treatment information, and patients and physicians communicated effectively with organisations such as the centres for disease control (CDC) and prevention (Committee on Improving the Quality of Cancer Care: Addressing the Challenges of an Aging Population; Board on Health Care Services; Institute of Medicine et al. 2013.). The national diabetes prevention program's lifestyle change program (DPP LCP) as a proposed treatment plan facilitates the referral process (Diabetes Prevention Program DPP Research Group 2002). Based on the findings of the studies and the importance of prevention in controlling diabetes and reducing the number of diabetics, this paper aims to improve prediction accuracy and prevent people from becoming infected at a young age by utilising statistical analyses and machine learning techniques such as MLP, kNN, RF, and SVM.

In this study, we address the critical issue of diabetes prediction and prevention using advanced statistical analyses and machine learning techniques, namely MLP, kNN, RF, and SVM. Our main contribution lies in the improvement of prediction accuracy, which has the potential to significantly impact healthcare interventions and enhance early detection efforts. By leveraging these approaches, our research aims to prevent individuals from contracting diabetes at a young age, consequently leading to improved healthcare outcomes and reduced burden on healthcare systems. This research presents a novel approach for detecting and predicting diabetes using data mining and machine learning techniques, incorporating four valid classifiers to achieve high accuracy, specificity, and sensitivity values.

This careful curation and subsequent utilisation of machine learning algorithms resulted in remarkably high accuracy, specificity, and sensitivity values, showcasing a promising approach for early diabetes detection and prediction.

The structure of the paper is as follows:

Section 2 explains how the data was collected and how the proposed method was developed. Also, this section uses some machine learning methods to express the simulation results. The results obtained in this paper are compared with those obtained by other research methods in section 3. Finally, section 4 describes future work that can be done in this field and concludes the paper.

2. Material and method

2.1. Data description

A UCI database containing 520 data is used in this study, and 17 features are retrieved from these participants. Data was collected using a direct questionnaire from the Sylhet Diabetes Hospital in Sylhet, Bangladesh (Islam et al. 2020). These data include the general variables of age and gender, as well as the variables of frequent urination and sudden weight loss, obesity, etc. Figure 1 shows the data distribution according to gender and frequent symptoms. The percentage of data distribution based on the average of different ages of the youngest (15–30 years), young (30–45 years), adults (45–60 years), elderly (60–75 years), and the oldest (75–90 years) is shown in Figure 2. These data are divided into two categories, with 320 data in the diabetes group and 200 in the normal group. In Figure 3, according to the data classification, there are 181 men and 19 women in normal people and 143 men and 173 women in people with diabetes.

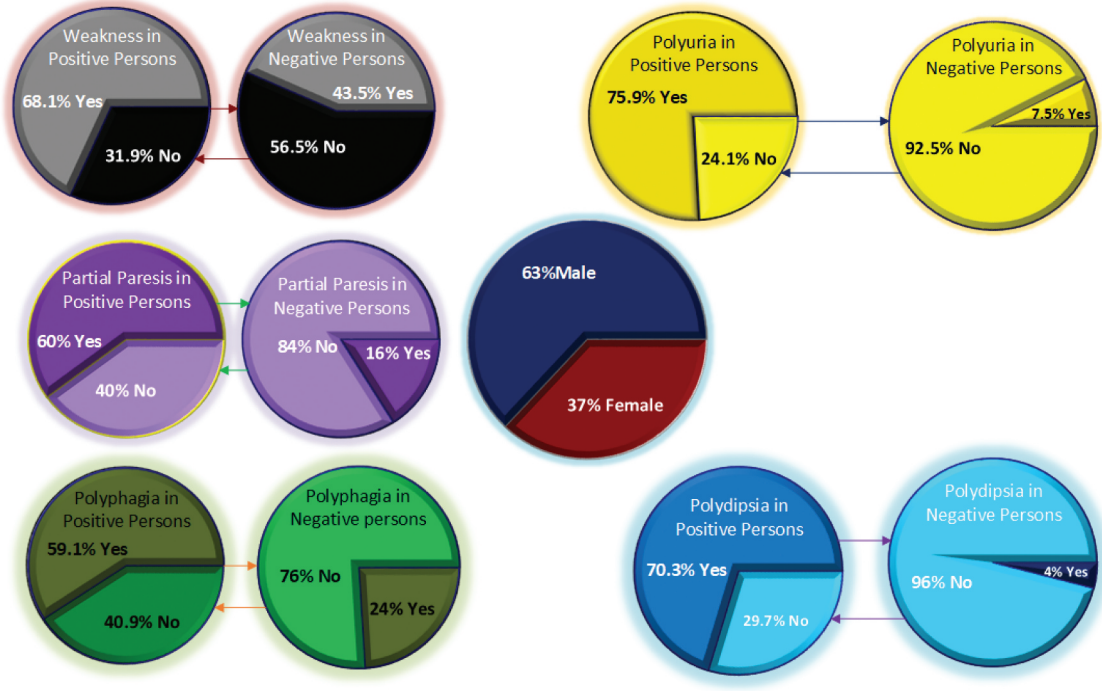

Figure 1. Distribution of data by gender and symptoms with a high dispersion index.

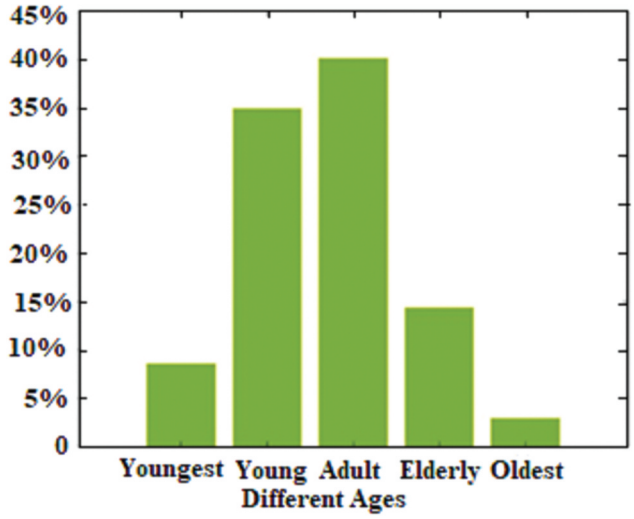

Figure 2. The percentage of data distribution according to the average age.

2.2. Proposed method

A machine learning algorithm can predict diabetes more accurately and efficiently with the most common data in normal and diabetic groups (Sanakal and Jayakumari 2014; Kumar Dewangan and Agrawal 2015; Qin et al. 2022; Kodama et al. 2022). This paper employs machine learning algorithms to predict the disease. The classification algorithm is one of the most common model learning methods for data prediction and analysis. In prediction methods, the values of some features are used to predict the value of a specific feature. A primary data set is divided into a training set and a test set in classification methods. Our model is tested and validated using the training and test data sets, and its accuracy is calculated.

Of the total data used in this research, 75% of them are used to train the models, and 25% of the rest is used to evaluate the obtained models. The confusion matrix was used to evaluate the model, a standard criterion for evaluating data mining models. 520 data into four groups 1) Selection of normal and diabetic people by influencing 5 factors with the highest dispersion index 2) Selection of the oldest normal and diabetic people 3) Selection of the youngest people Normal and diabetic 4) A random selection of normal and diabetic data were clustered.

To select diabetic people with the influence of 5 factors with the highest frequency index, we tried to use data with more than 5 influencing factors with the highest dispersion index, so there is less standard deviation. Suppose a diabetic person had 8 symptoms out of 14 symptoms. In that case, our condition for selecting this data had at least 5 influential factors with a high dispersion index if the same person had 3 factors out of 5 factors with a high dispersion index and having 5. Another factor was that data is not selected because it increases the standard deviation (if a person had 8 symptoms and 4 out of 5 symptoms, he would not be selected because the standard deviation is equal to the index data) 115 data from diabetic people were used for training, and 115 data from normal people were used for machine testing. The age range of the oldest diabetic people was between 61 and 90 years old, and normal people were between 54 and 72 years old. To train the machine, a selection of 56 data from diabetic people was used, and for testing it, a selection of 58 data from normal people was used. The youngest diabetic people were selected in the age range of 16 to 35 in men, 25 to 35 in women, and normal people in the age range of 26 to 37 in men and 28 to 36 in women. Fifty-one persons with diabetes and 51 normal data were used for training and testing the machine. As randomly

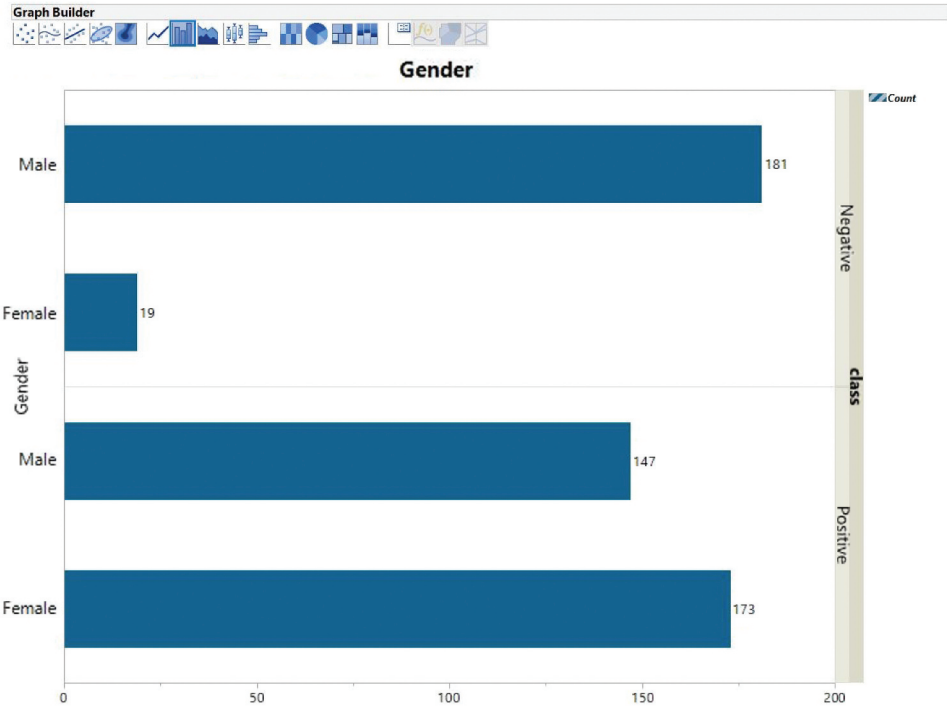

Figure 3. Data graph according to gender in normal and diabetic groups.

selected 520 data because there were only 200 data from normal people, and 200 data were randomly selected from 320 data of diabetic people so that the data for training and testing the machine is the same. MATLAB version 2022b neural network toolbox library and Feed-Forward Back Propagation algorithm were used to classify patients and normal people. The classifiers used in this paper to distinguish between two groups were SVM, RF, MLP, and kNN.

2.2.1. Support vector machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that separates data samples represented as points in space using a line or hyperplane. This separation is such that the data points on the same side of the line are similar and placed in the same group. New data samples will be placed in one of the existing categories after being added to the same space. When the data points cannot be separated by a straight line or a straight hyperplane, the problem is called non-linear. In such a situation, the SVM kernels come into play and increase the dimensions of the space so that the data points are linearly separable (Moridani et al. 2019).

2.2.2. Random forest

The random Forest Algorithm is a popular machine learning algorithm from the artificial intelligence subcategory that belongs to the supervised learning technique. It is used in machine learning for classification and regression problems (predicting and expressing one variable's changes based on another variable's information). Random Forest Algorithm is a classification that includes several decision trees (DT) in different subsets of the data set and takes the average to improve the prediction accuracy of that data set. Instead of relying on a decision tree, a random forest predicts the prediction from each tree based on the majority of votes and considers the final result as the output. More trees in the forest lead to higher accuracy and avoid the problem of overfitting (VijiyaKumar et al. 2019).

2.2.3. Multi-layer perceptron

The multi-layer perceptron neural network is one of the most widely used examples of neural networks. The efficiency of this neural network in solving a problem depends on the topology considered for the network. If the network's topology is more straightforward than the requirement, there is no possibility of learning in the neural network. If the topology is more complex than the requirement, the problem of overtraining will occur, and the generalisation power of the obtained network will be very low. A MLP neural network will be obtained by stacking several perceptron. That is, we will have several layers of neurons in such a network. Multi-layer perceptron is a class of feed-forward artificial neural networks. An MLP consists of at least three layers: an input layer, a hidden layer, and an output layer. Except for the input nodes, each node is a neuron that uses a non-linear activation function. MLP uses a supervised learning technique for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that are not linearly separable (Mohammad 2022).

2.2.4. K nearest neighbor

k-Nearest Neighbors (kNN) is a simple supervised machine learning algorithm with easy implementation. This algorithm can be used to solve classification and regression problems. A supervised machine learning algorithm is an algorithm that relies on labelled input data to learn from and label it. In this algorithm, a sample is classified by the majority vote of its neighbours, and this sample is determined in the most general class among k nearest neighbours. K is a positive integer value and is generally small. If $k = 1$, the sample is determined in the class of its nearest neighbours. An odd value of k is useful because it prevents equal votes. The k-nearest neighbour method is used for many methods because it is effective, non-parametric, and easy to implement. However, the classification time is long, and it isn't easy to find the optimal value of k. The best choice of k depends on the data. In general, a large value of k reduces the effect of noise on classification, but the boundary between classes becomes less distinct (Suyanto et al. 2022).

We outline the key parameters for each classifier along with their values.

MLP Classifier:

Number of Hidden Layers: 2
Number of Nodes in Hidden Layers: 10
Activation Function: ReLU
Learning Rate: 0.001
Number of Epochs: 100

SVM Classifier:

Kernel: Radial Basis Function (RBF)
Regularisation Parameter (C): 1.0
Kernel Coefficient (Gamma): scale

RF Classifier:

Number of Trees: 100
Maximum Depth: 5
Minimum Samples Split: 2
Minimum Samples Leaf: 1
Maximum Features: square root of the number of features
Decision Tree (DT) Classifier:
Maximum Depth: 5
DecisionTree (DT) Classifier
Minimum Samples Split: 2
Minimum Samples Leaf: 1
Maximum Features: means all features are considered

2.3. Evaluation

Indicators are used to determine the performance of the diagnostic system to identify normal and patients. The values of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values were calculated due to the equality of the clustering data in the two diabetic and normal classes. In machine learning and statistics, we often use the terms true positive and true negative, yet people still get confused, and so their matrix is known as a confusion matrix. The confusion matrix can also be used to calculate the model's accuracy, sensitivity, and specificity. By analysing these measures, we

can gain a better understanding of how the model performs. The formulas (1) to (3) can be used to calculate these measures, and confusion matrix can be used for the calculation (Mohammad et al. 2021).

$$\text{Sensitivity(Sen)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{Specificity(Spe)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2}$$

$$\text{Accuracy(Acc)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{3}$$

In Figure 4, the blocks illustrate the steps of data processing to detect normal and diabetic patients. After splitting the data into two groups (train and test), the preprocessed data is used for training the model, and then the test data is used for testing. The model is then evaluated by using parameters of sensitivity, specificity, and accuracy after the test data has been applied to the model.

In our study, the dataset may have contained missing values, which could have adversely affected the performance of the machine learning algorithms. To address this issue, we employed the K-Nearest Neighbors (KNN) algorithm for missing value imputation.KNN imputation is a data imputation technique that replaces missing values with the average value of the K-nearest data points based on their feature similarity. This method is widely used in handling missing data because it preserves the underlying data distribution and can improve the accuracy of the subsequent analyses.

Feature scaling is a critical preprocessing step to ensure that all features have a similar scale and prevent any single feature from dominating the analysis due to its larger magnitude. Z-Score Normalisation, also known as standardisation, is a common method used for feature scaling. In this step, we standardised each feature by subtracting the mean of the feature and dividing by its standard deviation. This process transforms the features into a normalised range between 0 and 1. Min-Max Normalisation is preferred as it maintains the relative relationships among the features while avoiding the impact of outliers and improving the convergence of machine learning algorithms. By scaling the features to a common range, Min-Max Normalisation ensures that all attributes contribute equally to the learning process, thus enhancing the performance and accuracy of the predictive model. By applying KNN for missing value imputation and Min-Max Normalisation for feature scaling, we aimed to enhance the quality and reliability of our dataset, thereby improving the performance of the machine learning algorithms used in diabetes detection and prediction.

3. Simulation results

The database study showed that some symptoms are more frequent than others, which can help us diagnose and predict diabetes. The symptoms of extreme weakness, blurred vision, and overeating were seen in both groups. The symptoms of frequent urination, excessive thirst, and severe weight loss, among the most common symptoms of diabetes (Kumar and Suresh 2021), were more common in the group of people with diabetes. People's age also affects the occurrence of symptoms. JMP software was used to check the database more closely. Figure 1 represents the above explanations. Examining the characteristics of people's gender and age showed that men in the age range (of 55–60 years old) and women in the age range (of 35–40 years old) showed symptoms of the disease and were diagnosed with diabetes. The interpretation of the obesity graph showed that this characteristic could not be a determining factor in predicting a person with diabetes. The reason is that this factor has little dispersion in diabetic and non-diabetic people.

In the Alopecia data, 24.4% of the people with diabetes and 50.5% in the group of normal people had this factor due to the high percentage of participation in non-diabetic people and the low percentage of occurrence in people who were positive for diabetes., is not a determining factor

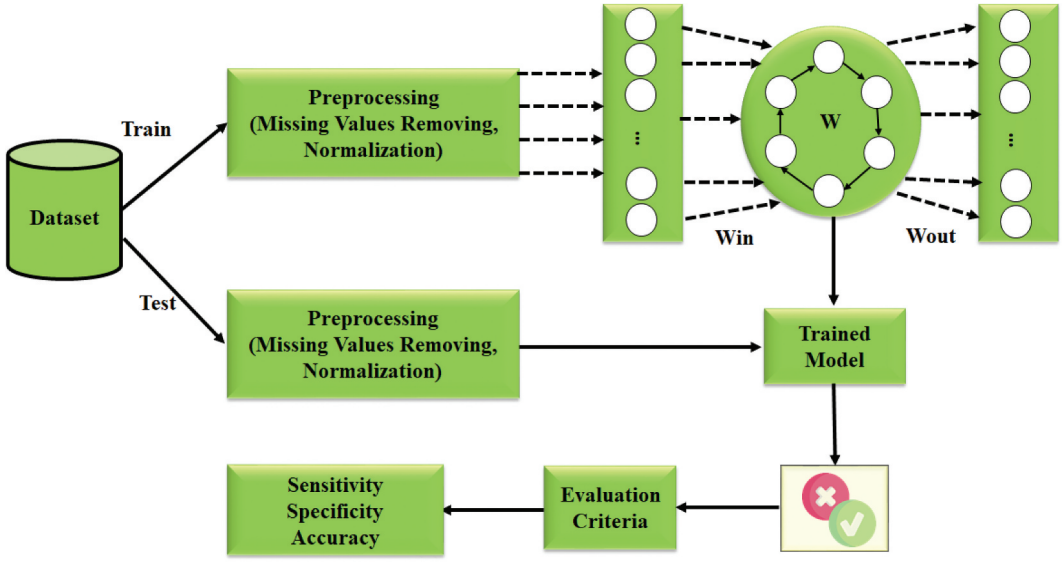

Figure 4. An overview of the steps in developing models using ML algorithms.

in predicting diabetes. Considering the small dispersion of the muscle stiffness feature in the data, this factor, like the case of obesity, cannot be an influencing factor in the diagnosis of diabetes. According to the studies, 42.2% of diabetic people have had a positive test, including this feature, and 30% of normal people have received a positive result. Among these features, Partial Paresis can play a decisive role in predicting diabetes, with 60% positive results in diabetic people and 16% in normal people. It was investigated that temperament is one of the determining factors, although it is not one of the five influencing factors. Also, blurred vision, with a percentage of 54.6%, can be used as an essential factor in predicting and diagnosing diabetes. In general, 5 factors that showed more than 50% frequency in diabetic people include weakness at 68.1%, Polyphagia at 59.1%, Polydipsia at 70.3%, Polyuria at 75.9%, and sudden weight loss at 52% among the cases that can play a decisive role in predicting diabetes. We achieved significant results, among which we can say how the polyuria factor has changed in different ages. For example, in the age range of 35–40 years, the highest frequency is related to polyuria. The polyuria factor is the most common among women with diabetes aged 35 to 40 and the most common among men with diabetes between the ages of 55 and 60. This factor did not have dispersion in normal women, but it had the highest distribution in men aged 70 to 75. Figure 5 shows the changes related to the polyuria factor in the normal and diabetic groups according to age and gender.

The polydipsia factor in women with diabetes between the ages of 35 and 40 has been widely distributed, and in normal women, this factor has not been distributed; but in the men aged 55 to 60, the highest dispersion was in people with diabetes, and the age 40 to 60 in normal people included the highest distribution. Figure 6 shows the changes in the polydipsia factor in different age groups and according to the gender and condition of people in terms of being normal or diabetic.

The partial paresis factor, which was the third most influential factor in women with diabetes, this index was most prevalent in the age group of 40 to 50 years and normal women in the age group of 20 to 30 and 50 to 60 years; It has the highest prevalence in men with diabetes aged 50 to 60 and the highest prevalence in normal men aged 40 to 50. The partial paresis changes in women and men according to the type of disease and age group are shown in Figure 7.

Severe weakness was one of the influential factors, with the highest dispersion index observed in women with diabetes aged 35 to 40 and normal women aged 50 to 60. The highest dispersion index is seen in men with diabetes between 55 and 60 and normal men between 50 and 55. The fifth influential factor was polyphagia, most common in women with diabetes aged 40 to 50 and normal women aged 20 to 30. The highest prevalence of polyphagia is seen in men with diabetes and normal between the ages of 50 and 60. A comparison of the degree of severe weakness and polyphagia in women and men is shown in Figures 8 and 9.

Tables 1–4 shows the results obtained using different classifiers and the effective features of diabetes in both young and older adults and in general. The results show that the SVM classifier has performed better than other classifiers in diagnosing diabetes. The sensitivity, specificity, and accuracy of the proposed model using the SVM classifier when the Top 5 features are selected to feed the classifier are 98.78 ± 1.96%, 99.04 ± 0.94%, and 99.74 ± 0.48%respectively. Using other classifiers showed that RF, MLP, and kNN are in the following ranks after SVM. Figure 10 shows the receiver operating characteristic (ROC) diagram for different classifiers. The area under curve (AUC) value for the SVM classifier is 0.97 and has more value than other classifiers.

Table 5 presents the confusion matrix for the SVM classifier, showcasing a comparative analysis between using all features

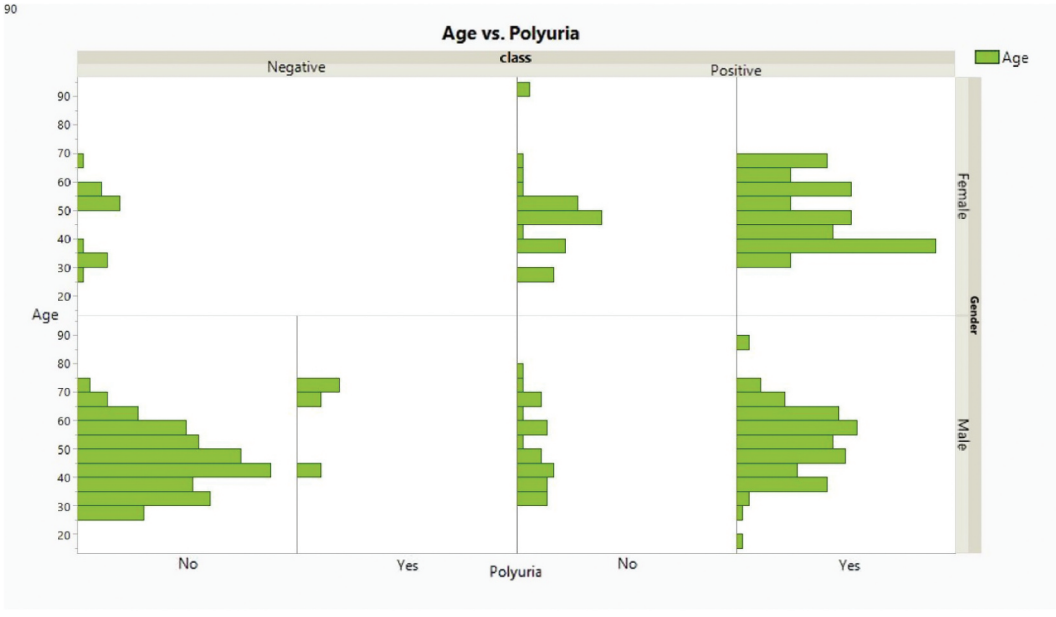

Figure 5. Comparison of polyuria factor changes in normal and diabetic groups based on age and gender.

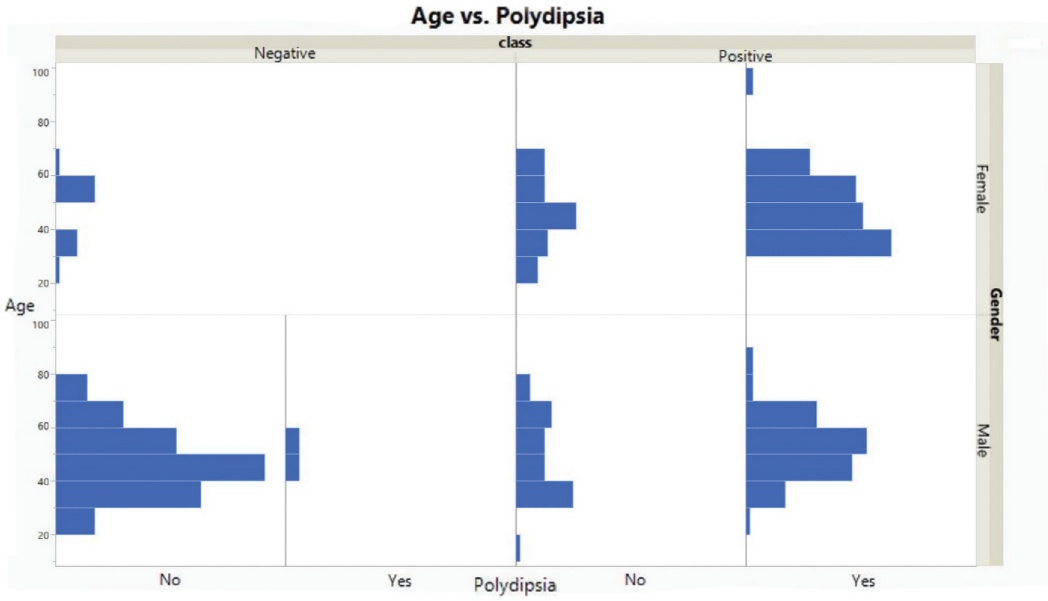

Figure 6. Comparison of polydipsia factor changes in normal and diabetic groups based on age and gender.

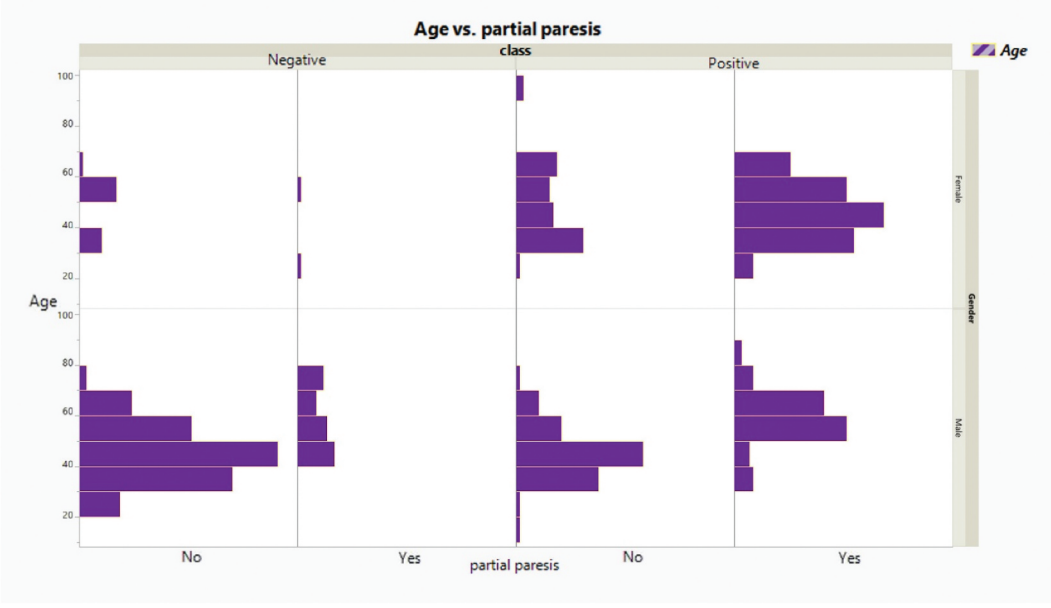

Figure 7. Comparison of partial paresis factor changes in normal and diabetic groups based on age and gender.

and a subset of top features for all patients in our study. The confusion matrix is a critical tool for evaluating the performance of a classifier model, enabling us to assess the accuracy and effectiveness of the SVM algorithm in predicting patient outcomes.

In this study, we utilised a dataset comprising both normal and patient cases, with an equal distribution of 200 samples. The SVM classifier was trained and tested using two different feature sets: the complete set of all features and a curated subset of top features identified through rigorous analysis.

The confusion matrix reveals the four key metrics essential for classifier evaluation:

- True Positive (TP): This represents the percentage of actual patient cases that were correctly classified as positive (patient) by the SVM classifier.
- False Negative (FN): This shows the percentage of actual patient cases that were misclassified as negative (normal) by the SVM classifier.
- False Positive (FP): This indicates the percentage of actual normal cases that were incorrectly classified as positive (patient) by the SVM classifier.

True Negative (TN): This depicts the percentage of actual normal cases that were correctly classified as negative (normal) by the SVM classifier.

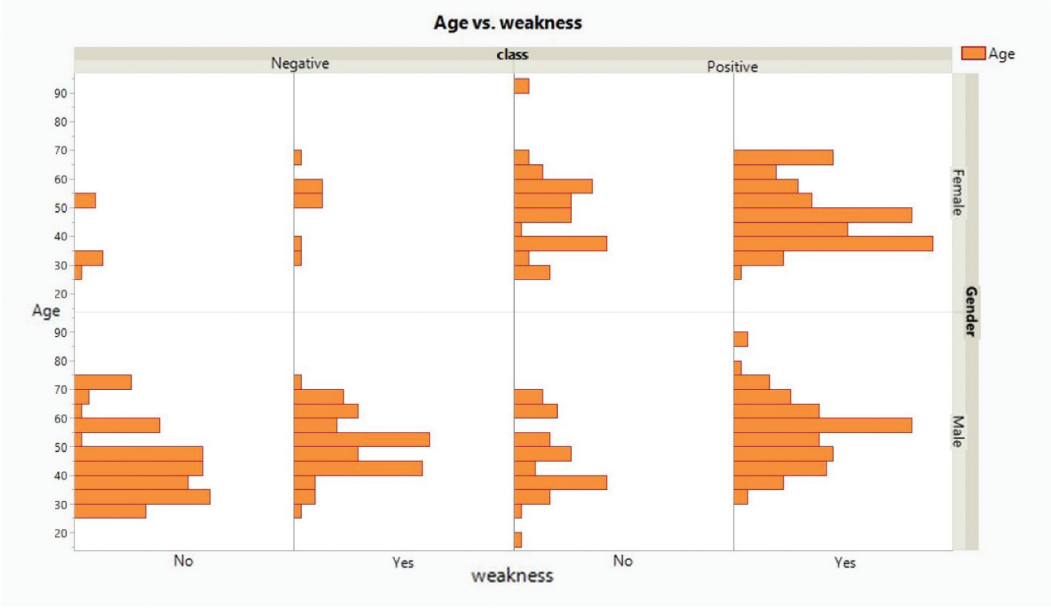

Figure 8. Comparison of weakness factor changes in normal and diabetic groups based on age and gender.

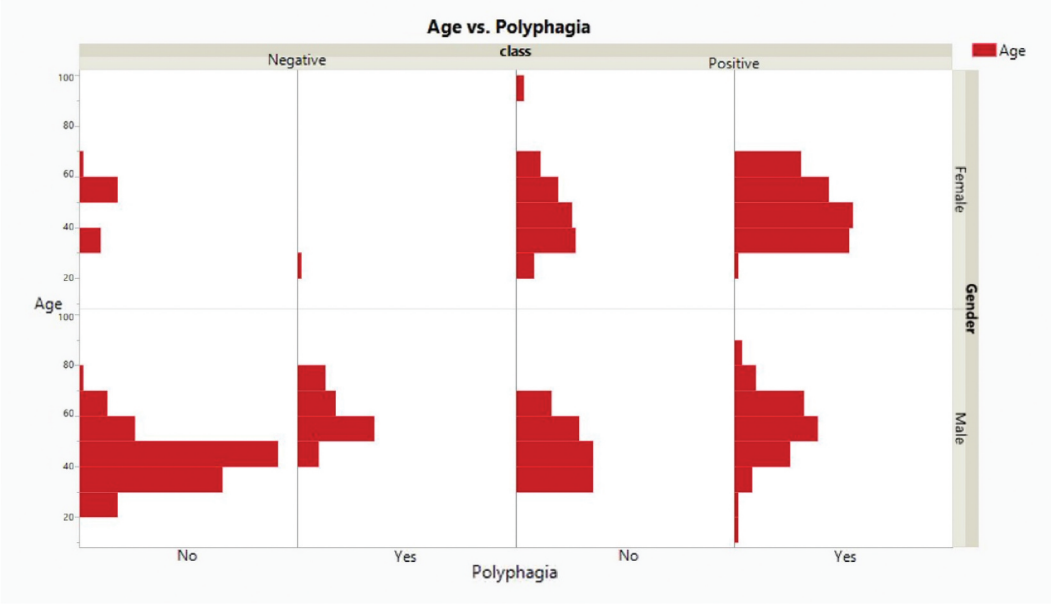

Figure 9. Comparison of polyphagia factor changes in normal and diabetic groups based on age and gender.

Table 1. The evaluation parameters of SVM classifier.

	Top 5 Features			All Features		
Metric	All Patients	Youngest Patients	Oldest Patients	All Patients	Youngest Patients	Oldest Patients
Sen (%)	97.32 ± 2.45	98.21 ± 2.85	98.71 ± 2.67	92.43 ± 4.23	93.87 ± 3.13	93.65 ± 3.46
Spe (%)	99.28 ± 1.63	99.96 ± 0.72	99.98 ± 0.54	93.76 ± 3.57	94.16 ± 2.65	94.35 ± 2.32
Acc (%)	98.78 ± 1.96	99.04 ± 0.94	99.74 ± 0.48	93.90 ± 3.34	94.74 ± 2.17	94.87 ± 1.95

Table 2. A description of the parameters used in RF classifier evaluation.

	Top 5 Features			All Features		
Metric	All Patients	Youngest Patients	Oldest Patients	All Patients	Youngest Patients	Oldest Patients
Sen (%)	95.54 ± 4.65	96.98 ± 3.74	96.42 ± 3.67	90.08 ± 5.04	91.96 ± 3.95	91.12 ± 3.90
Spe (%)	97.87 ± 3.09	97.64 ± 3.23	97.09 ± 2.99	91.19 ± 3.90	92.45 ± 3.47	92.17 ± 4.87
Acc (%)	96.12 ± 3.39	97.52 ± 3.11	97.57 ± 2.04	91.32 ± 3.87	92.36 ± 3.71	92.32 ± 4.20

Table 3. Analysing the parameters of MLP classifiers.

	Top 5 Features			All Features		
Metric	All Patients	Youngest Patients	Oldest Patients	All Patients	Youngest Patients	Oldest Patients
Sen (%)	90.12 ± 5.23	91.34 ± 5.15	92.62 ± 4.73	86.39 ± 6.90	87.42 ± 5.65	87.89 ± 5.09
Spe (%)	92.18 ± 4.78	92.30 ± 4.10	93.16 ± 4.49	87.27 ± 6.29	88.18 ± 5.28	88.90 ± 4.90
Acc (%)	93.25 ± 4.29	94.67 ± 4.29	94.49 ± 3.79	87.71 ± 6.10	89.19 ± 5.12	89.39 ± 4.38

Table 4. Evaluation parameters for kNN classifier.

	Top 5 Features			All Features		
Metric	All Patients	Youngest Patients	Oldest Patients	All Patients	Youngest Patients	Oldest Patients
Sen (%)	87.22 ± 6.26	89.45 ± 5.36	89.18 ± 5.19	83.37 ± 7.11	85.11 ± 6.54	85.54 ± 5.79
Spe (%)	88.18 ± 6.18	90.36 ± 5.87	90.47 ± 5.37	84.14 ± 6.98	86.37 ± 6.11	86.39 ± 6.44
Acc (%)	89.10 ± 6.04	91.83 ± 5.14	91.28 ± 5.23	85.74 ± 6.35	87.72 ± 5.90	87.08 ± 5.72

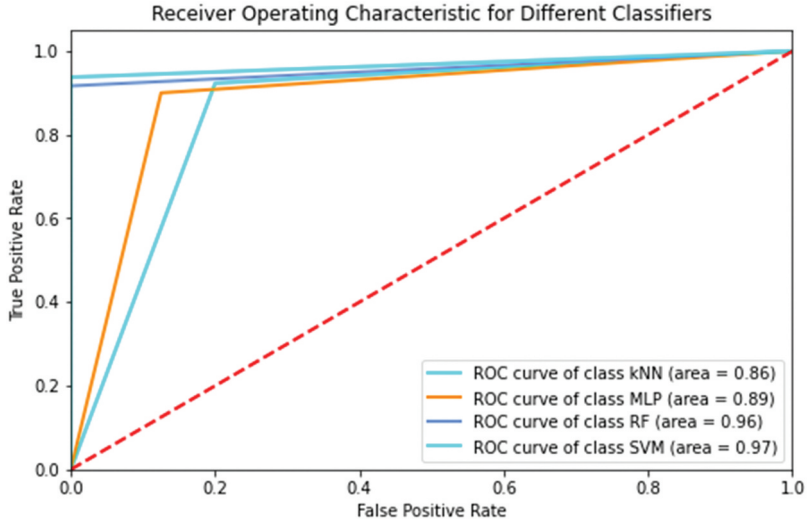

Figure 10. Receiver operating characteristics for different classifiers.

Table 5. Confusion matrix for SVM classifier: a Comparison between all features and Top features for all Patients.

		Predicted	
		Positive	Negative
All Features			
Actual	Positive	TP=92	FN=8
	Negative	FP=6	TN=94
Top Features			
Actual	Positive	TP=98	FN=2
	Negative	FP=0	TN=100

Table 6. Summary of diabetes identification studies .

Author (s)	Number of Samples	Number of Features	Method/Classifier Name (s)	Accuracy
Kaur et al (Kaur and Chhabra 2014).	1	1	Random blood glucose check	Above 66%
Saxena et al (Saxena et al. 2014).	100	11	kNN and SVM	66% (average)
Hwang et al (Huang et al. 2015).	768	8	DT, RF, SVM, BN	Above 60%
Ram D. et al (Joshi and Dhakal 2021).	768	8	Logistic Regression	78%
Sidong W et al (Wei et al. 2018).	768	8	Deep Neural Network	78%
Roshan B (Roshan et al. 2019).	768	8	Gradient Boosting, Logistic Regression, Naive Bayes	86%

By comparing the results of using all features and top features, we can gain insights into the impact of feature selection on the SVM classifier's performance. The accuracy, sensitivity, and specificity values for each feature set are presented, allowing us to evaluate the overall predictive power of the SVM model in both scenarios.

4. Discussion

Polyuria and excessive thirst were more than 70% of the distribution in diabetic people. It can be seen that these two factors have a more significant impact on the disease and can be considered as the symptoms of type 1 diabetes, which is the body's reaction to high blood sugar due to Consider the destruction of the islets of Langerhans and the absence of insulin in the body, which in very severe cases, excess blood sugar is excreted through urine. Of course, these factors can also be seen in people with type 2 diabetes, which happens due to improper control of diabetes. (Type 2 diabetes, in case of incorrect control, progresses to type 1 diabetes and insulin injection). Obesity had a low frequency in the data. The evaluation and clustering found that this factor is not a good choice for predicting an early diabetes diagnosis. Data analysis showed that other factors, such as itching and blurred vision, can effectively diagnose and predict diabetes with high accuracy.

Now, by using the data that are most frequent in normal and diabetic groups, machine learning can be used to predict diabetes better and more accurately. This article uses machine learning algorithms to predict this disease's early stages. The classification algorithm is one of the most common model learning methods for prediction in data analysis. In prediction methods, the values of some features are used to predict the value of a specific feature. In classification methods, the primary data set is divided into training and test data sets. The model is built using the training data set and the test data set for testing and validation and calculating the model's accuracy. Of the total data used in this research, 75% of them are used to train the models, and 25% of the rest is used to evaluate the obtained models. The confusion matrix and ROC curve were used to evaluate the model, a common criterion for evaluating data mining models.

Kaur et al. predicted type 2 diabetes using the Pima data set; it was shown that the insulin serum level is the most important feature in diabetic people; if the insulin level is above 800, it indicates that the person has diabetes (Kaur and Chhabra 2014). Saxena et al. presented an integrated approach between diagnosis for kNN and SVM methods of type 2 diabetes on the Pima dataset. They concluded that kNN also in this data set to diagnose type 2 diabetes. It is the most effective artificial intelligence algorithm and has reached an average accuracy of 66% (Saxena et al. 2014). Hwang et al. have used DT, RF, SVM, and Bayesian network algorithms to predict type 2 diabetes. By using this method, they reached an accuracy of over 60% (Huang et al. 2015). Ram D. et al. used the logistic regression algorithm to achieve a 78% accuracy rate (Joshi and Dhakal 2021). In a study by Sidong W et al., the deep neural network model achieves 78% accuracy compared to SVM, DT, and NB (Wei et al. 2018). Roshan B. utilised Gradient Boosting, Logistic Regression, and Naive Bayes algorithms to build the models. With Gradient Boosting, the highest accuracy was 86% (Roshan et al. 2019).

Diabetes symptoms can be different in different blood sugar ranges. Therefore, future articles predicting the occurrence of symptoms in different blood sugar ranges can be strategic. Also, with the advancement of artificial intelligence in medicine, the use of a comprehensive model along with all factors affecting the incidence of diabetes, and the use of deep learning models, we can hope to improve the results.

Research shows that the number of diabetic patients worldwide is constantly increasing. As a result, to improve the quality of health, it is necessary to create a model that can be used to predict and diagnose diabetes earlier. To predict diabetes disease, it is necessary to consider a system that can collect data from a significant source of diabetic patients. Also, this system should be highly reliable. Today, the stages of pre-diabetes and prevention are much more important than controlling and treating diabetes. By general analysis of the data, it was found that the age of onset of diabetes symptoms in men and women is different; the highest age dispersion index in women in the field of onset of symptoms is in the ages (35–40 years old) and men in the ages (55–60 years old).

Diabetes is known as one of the hidden diseases, and this sentence means that the symptoms appear earlier than the disease itself; Now, by examining the data and the effectiveness of diabetes symptoms, it was found that symptoms such as frequent urination, excessive thirst, partial paresis, severe weakness, and overeating due to their high prevalence are among the important symptoms in early diagnosis and prediction of diabetes. Other symptoms can also be effective, but considering that the increase in the number of parameters can cause the complexity of the system and the slowness of the response speed, in the future, with the help of deep learning models, more parameters can be used to investigate the relationship between the increase in parameters affecting diabetes and the effectiveness of the prediction model. Mapping-based methods that have brilliant results in disease diagnosis can be used in the future to improve the results (Karimi Moridani et al. 2013; Moridani and Bardineh 2018; Moghadam et al. 2021;

Mohammad and Khameneh‌zahra 2022; Mahmoudi et al. 2022). In recent years, the integration of artificial intelligence (AI) in the field of medical imaging and diagnostics has shown tremendous potential for revolutionising healthcare practices (Ahmet 2011; Akben et al. 2016; Kubilay and Kaba 2022; Sunnetci and Alkan 2023). Our plan for future research involves leveraging deep learning models and architectures to further enhance the predictive capabilities of our diabetes prediction system. We aim to compare the performance of deep learning models with the traditional machine learning techniques used in this study (Anand et al. 2022). This extension will enable us to comprehensively evaluate the potential benefits of deep learning in diabetes prediction and contribute to the ongoing advancement of predictive healthcare models.

Table 6 provides a comprehensive summary of various research studies focused on the identification of diabetes using different machine learning techniques. Each row represents a specific study and includes essential information, such as the dataset used, the number of features, the type of classifier employed, and the achieved accuracy.

Conclusion

In conclusion, our study has highlighted the importance of early diabetes detection and prediction using machine learning techniques. The SVM classifier, when trained on a curated set of features, has shown promising results in distinguishing between normal and diabetic cases. Key symptoms such as frequent urination, excessive thirst, and others have proven to be critical in the early diagnosis of diabetes.

Looking ahead, we recognise the potential of deep learning models and advanced artificial intelligence techniques to further improve the accuracy and reliability of diabetes prediction. Mapping-based methods and comprehensive models incorporating multiple parameters hold promise for refining prediction models.

The advancement of medical AI and the establishment of a reliable data collection system from a significant pool of diabetic patients are crucial for future research in the field. Emphasising pre-diabetes stages and preventive measures is essential for promoting public health and enhancing the quality of diabetes management.

In summary, this study contributes to the growing body of knowledge on diabetes prediction, and the findings provide valuable insights for researchers and healthcare practitioners alike. With continued advancements in technology and data analysis, we envision substantial progress in predicting and preventing diabetes in the future.

While our study offers valuable insights into diabetes prediction using the classifiers, we recognise its limitations and the potential dark side. The small dataset size, feature selection, and class imbalance are factors that may impact the model's performance. External validation and a more extensive comparison with other classifiers would provide a more robust evaluation.

Nevertheless, our research contributes to the field of diabetes prediction and emphasises the significance of early diagnosis. We believe that by addressing the study's limitations and exploring advanced techniques, such as deep learning models, we can enhance the accuracy and reliability of diabetes prediction in the future. It is essential to continue advancing medical AI and data collection efforts to improve healthcare outcomes for individuals at risk of diabetes.

Abbreviations

ANN	Artificial Neural Network
AUC	Area under Curve
CDC	Centers for Disease Control
CPCSSN	Canadian Primary Care Sentinel Surveillance Network
DT	Decision Tree
FN	False Negative
FP	False Positive
kNN	k Nearest Neighbor
LDA	Linear Discrimination Analysis
LR	Logistic Regression
ML	Machine Learning
MLP	Multi-Layer Perceptron
NB	Naive Bayesian
PIDD	Pima Indians Diabetes Dataset
RF	Random Forest
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UKPDS	UK Prospective Diabetes Study

Disclosure statement

No potential conflict of interest was reported by the author(s)

Funding

No source of funding for this work.

Notes on contributors

Mohammad Ehsan Farnoodian received a B.S. degree in biomedical engineering-bioelectric from Tehran Medical Science, Islamic Azad University, Tehran, Iran, and earned his M.S. degree in biomedical engineering-bioelectric from Science and Research branch, Islamic Azad University, Tehran, Iran, in 2023. He is passionately dedicated to the examination and interpretation of biomedical data, particularly in the context of disease prediction and detection. His academic pursuits involve in-depth exploration of biomedical data analysis intricacies, with a specific focus on employing data-driven approaches for disease anticipation and identification.

Mohammad Karimi Moridani received a BS degree in electrical engineering-Electronic from 2006, and he obtained MS and Ph.D. degrees in biomedical engineering-bioelectric in 2008 and 2015, respectively. Currently, he serves as an assistant professor in the biomedical engineering department at Tehran Medical Science, Islamic Azad University in Tehran, Iran. His research focuses on biomedical signal and image processing, nonlinear time series analysis, and cognitive science, with specific applications ranging from ECG, HRV, and EEG signal processing for disease detection and prediction to epileptic seizure prediction, pattern recognition, image processing for facial and beauty recognition, watermarking, and more. He is driven by a passion to contribute meaningfully to the scientific community and employs data-driven methodologies to address critical challenges in healthcare and related fields.

Hanieh Mokhber received a B.S. degree in biomedical engineering-bioelectric from Islamic Azad University of Tehran Medical science. Her scholarly endeavors involve a meticulous exploration of the complexities of

biomedical data analysis, with a specific and unwavering emphasis on harnessing data-driven methodologies to anticipate and identify various diseases.

ORCID

Mohammad Karimi Moridani http://orcid.org/0000-0003-0793-3797

Authors' contributions

All authors evenly contributed to the whole work. All authors read and approved the final manuscript.

Availability of data and materials

The data used in this paper is cited throughout the paper.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

References

Absar N, Das EK, Shoma SN, Khandaker MU, Miraz MH, Faruque MRI, Tamam N, Sulieman A, Pathan RK. 2022. The efficacy of machine-LearningSupported smart System for heart disease prediction. Healthcare. 10(6):1137. doi: 10.3390/healthcare10061137.

Adu MD, Malabu UH, Malau-Aduli A, Malau-Aduli BS, Rodda S. 2019. Enablers and barriers to effective diabetes self-management: a multi-national investigation. PloS One. 14(6):e0217771. doi: 10.1371/journal.pone.0217771.

Ahmet A. 2011. Analysis of knee osteoarthritis by using fuzzy c-means clustering and SVM classification. Scientific Research And Essays. 6 (20):4213–4219. doi: 10.5897/SRE11.068.

Akben SB, Alkan A, Gao Z-K. 2016. Visual Interpretation of Biomedical time series using Parzen Window-based Density-Amplitude Domain Transformation. PloS One. 11(9):e0163569. doi: 10.1371/journal.pone.0163569.

Anand D, Arulselvi G, Balaji G, Chandra GR. 2022. A deep convolutional extreme machine learning classification method to detect bone Cancer from histopathological images". Int J Intell Sys Appl Eng. 10(4):39–47.

Clarke PM, Gray AM, Briggs A, Farmer AJ, Fenn P, Stevens RJ, Matthews DR, Stratton IM, Holman RR. 2004. A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the United Kingdom Prospective diabetes study (UKPDS) outcomes model (UKPDS no. 68). Diabetologia. 47(10):1747–1759. doi: 10.1007/s00125-004-1527-z.

Committee on Improving the Quality of Cancer Care: Addressing the Challenges of an Aging Population; Board on Health Care Services; Institute of Medicine. 2013. Levit A, Balogh EP, Nass SJ, Ganz PA, editors. *Delivering high-quality cancer care: charting a new course for a system in crisis*. Washington (DC): National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK202146/

Diabetes Prevention Program (DPP) Research Group. 2002. The diabetes prevention program (DPP): description of lifestyle intervention. Diabetes Care. 25(12):2165–2171. doi: 10.2337/diacare.25.12.2165.

Dritsas E, Trigka M. 2022. Data-driven machine-learning methods for diabetes risk prediction. Sensors. 22(14):5304. doi: 10.3390/s22145304.

Furman D, Campisi J, Verdin E, Carrera-Bastos P, Targ S, Franceschi C, Ferrucci L, Gilroy DW, Fasano A, Miller GW, et al. 2019. Chronic inflammation in the etiology of disease across the life span. Nat Med. 25 (12):1822–1832. doi:10.1038/s41591-019-0675-0.

Han W, Yang S, Huang Z, Jian H, Wang X. 2018. Type 2 diabetes mellitus prediction model based on data mining. Inform Med Unlocked. 10:100–107. doi: 10.1016/j.imu.2017.12.006.

Huang GM, Huang KY, Lee TY, Weng J. 2015. An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients. BMC Bioinform. 16(Suppl 1):S5. doi: 10.1186/1471-2105-16-S1-S5.

Islam MF, Ferdousi R, Rahman S, Bushra HY. 2020. UCI machine learning repository: early-stage diabetes risk prediction dataset. Available online: [accessed on 5 July 2021] https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.

Jobeda JK, Foo SY. 2021. A comparison of machine learning algorithms for diabetes prediction. ICT Express. 7(4):432–439. doi: 10.1016/j.icte.2021.02.004.

Joshi RD, Dhakal CK. 2021. Predicting type 2 diabetes using Logistic regression and machine learning approaches. Int J Env Res Pub He. 18 (14):7346. doi: 10.3390/ijerph18147346.

Kalra S, Jena BN, Yeravdekar R. 2018. Emotional and psychological needs of people with diabetes. Indian J Endocrinol Metab. 22(5): 696–704. doi: 10.4103/ijem.IJEM_579_17.

Karimi Moridani M, Setarehdan SK, Nasrabadi AM, Hajinasrollah E. 2013. Mortality risk assessment of ICU cardiovascular Patients using physiological variables universal. J Biomedical Engineering. 1(1):6–9. doi: 10.13189/ujbe.2013.010102.

Kaur G, Chhabra A. 2014. Improved J48 classification algorithm for the prediction of diabetes. Int J Comput Appl. 98(22):13–17. doi: 10.5120/17314-7433.

Kodama S, Fujihara K, Horikawa C, Kitazawa M, Iwanaga M, Kato K, Watanabe K, Nakagawa Y, Matsuzaka T, Shimano H, et al. 2022. Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: a meta-analysis. J Diabetes Investig. 13:900–908. doi: 10.1111/jdi.13736.

Kubilay MS, Kaba E. 2022. Fatma Beyazal Çeliker, Ahmet Alkan, comparative parotid gland segmentation by using ResNet-18 and MobileNetV2 based DeepLab v3+ architectures from magnetic resonance images. Concurrency And Computation: Practice And Experience. 35(1):e7405. doi: 10.1002/cpe.7405.

Kumar Dewangan A, Agrawal P. 2015. Classification of diabetes mellitus using machine learning techniques. Int J Eng Appl Sci. 2:257905.

Kumar A, Suresh K. 2021. Diabetes mellitus: a stitch in time saves nine early diagnosis and management minimizes complications- a case study. Glob J Obes Diabetes Metab Syndr. 8(2):014–017. doi: 10.17352/2455-8583.000052.

Kuwata H, Okamura S, Hayashino Y, Tsujii S, Ishii H, Herder C, for the Diabetes Distress and Care Registry at Tenri Study Group. 2017. Higher levels of physical activity are independently associated with a lower incidence of diabetic retinopathy in Japanese patients with type 2 diabetes: a prospective cohort study, diabetes distress and Care registry at Tenri (DDCRT15). PloS One. 12(3):e0172890. doi: 10.1371/journal.pone.0172890.

Lin X, Xu Y, Pan X, Xu J, Ding Y, Sun X, Song X, Ren Y, Shan P-F. 2020. Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. Sci Rep. 10(1):14790. doi: 10.1038/s41598-020-71908-9.

Mahmoudi N, Moridani MK, Khosroshahi M, Moghadam ST. 2022. Epileptic seizure prediction using geometrical features extracted from HRV signal. In: Suma V, Fernando X Du KWH, editors. Evolutionary computing and Mobile sustainable networks. Lecture notes on Data Engineering and Communications Technologies. Singapore: Springer; p. 116. doi:10.1007/978-981-16-9605-3_33.

Moghadam FS, Moridani MK, Jalilehvand Y. 2021. Analysis of heart rate dynamics based on non-linear lagged returned map for sudden cardiac death prediction in cardiovascular patients. Multidim Syst Sign Process. 32(2):693–714. doi: 10.1007/s11045-020-00755-8.

Mohammad KM. 2022. An automated method for sleep apnoea detection using HRV. Journal Of Medical Engineering & Technolog. 46(2):158–173. doi: 10.1080/03091902.2022.2026504.

Mohammad KM, Khameneh ,zahra K. 2022. Mahsa Shahipour Shamsabad, designing an intelligent System to detect stress levels during driving. International Arab Journal Of Information Technology. 9(1):81–89. doi: 10.34028/iajit/19/1/10.

Mohammad KM, Sina Jabbari Behnam S, Heydar M. 2021. Automatic epileptic seizure detection based on EEG signals using deep learning. Artificial Intelligence Evolution. 2(2):96–106. doi: 10.37256/aie.2220211123.

Moridani MK, Bardineh HY. 2018. Presenting an efficient approach based on novel mapping for mortality prediction in intensive care unit cardiovascular patients. MethodsX. 5:1291–1298. doi: 10.1016/j.mex.2018.10.008.

Moridani M, Zadeh MA, Mazraeh ZS. 2019. An efficient automated algorithm for distinguishing normal and abnormal ECG signal. IRBM. 40 (6):332–340. doi: 10.1016/j.irbm.2019.09.002.

Qin Y, Wu J, Xiao W, Wang K, Huang A, Liu B, Yu J, Li C, Yu F, Ren Z. 2022. Machine learning models for data-driven prediction of diabetes by lifestyle type. Int J Environ Res Public Health. 19(22): 15027. doi: 10.3390/ijerph192215027.

Roshan B, Kumar Mourya A, Chauhan R, Kaur H. 2019. Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Appl Sci. 1(9):1–8. doi: 10.1007/s42452-019-1117-9.

Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, Colagiuri S, Guariguata L, Motala AA, Ogurtsova K, et al. 2019 Nov. IDF diabetes Atlas Committee. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International diabetes federation diabetes Atlas, 9th edition. Diabetes Res Clin Pract. 157:107843. DOI:10.1016/j.diabres.2019.107843.

Sanakal R, Jayakumari T. 2014. Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine. Int J Comput Trends Technol. 11(2):94–98. doi: 10.14445/22312803/IJCTT-V11P120.

Saxena K, Khan Z, Singh S. 2014. Diagnosis of diabetes mellitus using K nearest neighbor algorithm. IJCST. 2(4):36–43.

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

Sivashankari R, M S, Mohammad Kamrul H, Saeed Rashid A, Alsuhibany Suliman A, Sayed A-K. 2022. An empirical model to predict the diabetic positive using stacked ensemble approach. Front Public Health. 9. doi:10.3389/fpubh.2021.792124.

Sunnetci KM, Alkan A. 2023. Biphasic majority voting-based comparative COVID-19 diagnosis using chest X-ray images. Expert Syst Appl. 216:119430. doi: 10.1016/j.eswa.2022.119430.

Sun Q, Tang L, Zeng Q, Gu M. 2021. Assessment for the correlation between diabetic retinopathy and metabolic syndrome: a cross-sectional study. Diabetes Metab Syndr Obes. 14:1773–1781. doi: 10.2147/DMSO.S265214.

Suyanto S, Meliana S, Wahyuningrum T, Khomsah S. 2022. A new nearest neighbor-based framework for diabetes detection. Expert Syst Appl. 199:116857. doi: 10.1016/j.eswa.2022.116857.

Tigga NP, Garg S. 2020. Prediction of type 2 diabetes using machine learning classification methods. Procedia Comput Sci. 167:706–716. doi: 10.1016/j.procs.2020.03.336.

VijiyaKumar K, Lavanya B, Nirmala I and Caroline SS, "Random forest algorithm for the prediction of diabetes," 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)1–5 (2019). 10.1109/ICSCAN.2019.8878802

Wei S, Zhao X, and Miao C. A comprehensive exploration to the machine learning techniques for diabetes identification. In 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), pages 291–295. IEEE, 2018. 10.1109/WF-IoT.2018.8355130

Wu WT, Li YJ, Feng AZ, Li L, Huang T, Xu A-D, Lyu J. 2021. Data mining in clinical big data: the frequently used databases, steps, and methodological models. Mil Med Res. 8(1). doi: 10.1186/s40779-021-00338-z.