Global Journal of Computer and Engineering Technology (GJCET)

Vol. 5, No.1, January, 2025

ISSN: 2767-1933

DIABETES PREDICTION USING MACHINE LEARNING: A SYSTEMATIC REVIEW

Dolly Verma, RSR Rungta College of Engineering & Technology, India (Dv1622766@gmail.com)
Ashish Tamrakar, RSR Rungta College of Engineering & Technology, India(Ashish.tamrakar1987@gmail.com)

ABSTRACT

Today, diabetes is a serious and very common disease in the world. It is not only detrimental to the blood but also increases the risk of complications, including nerve damage, blindness, kidney damage, heart diseases, etc., which causes many deaths per year. Diabetes is a common health problem that develops due to poor lifestyle and bad habits such as a poor diet and lack of exercise. To predict diabetes, various data mining and machine learning methods are used, such as k-nn, support vector machine, random forests, logistic regression, decision tree, gradient boosting, naïve bayes, ada boosting, gaussian naive bayes, linear discriminant analysis, extra trees, bagging, deep neural network. We examined new developments in machine learning to increase diabetes prediction and diagnosis. In this article, to detect patients with diabetes, machine learning methods are used. The classification model identified diabetic patients with a higher accuracy.

Keywords: Random forest, support vector machine, logistic regression, k-nearest neighbor, diabetes, machine learning, classification, prediction, and accuracy.

1. INTRODUCTION

Diabetes is a chronic or lifelong illness that impacts the body's metabolism. Many people, even children, also suffered from this disease. What is diabetes? Glucose is necessary for energy generation in the body (cell). Due to insulin deficiency, blood sugar levels become high. Insulin is a hormone that regulates the level of sugar in blood. Diabetes, which is caused by hormonal deficiency. Diabetes is also called blood sugar. When increasing the level of glucose, lack of exercise, lifestyle, age, high blood pressure, poor diet, etc., can cause diabetes.

1.1 Symptoms of Diabetes are:

- Excessive elimination of urine
- Excessive thirst
- General weakness
- Loss of weight
- Frequent inflections
- Mood swings.

1.2 Types of diabetes are:

- Type 1(known as insulin-dependent diabetes).
- Type2(known as non-insulin-dependent diabetes).
- Gestational (which occurs during pregnancy).

If diabetes is predicted earlier, it can be managed. By following a regular checkup, a healthy lifestyle, regular exercise and proper medication, we can observe a healthy and long life.

1.3 Prevention:

- Avoid unhealthy food.
- Keep the habit of daily exercise like bicycling, swimming, housework, gardening, walking, etc.

This work gathered researchers to discuss how questions and answers can be implemented into ML for predicting diabetes. In this paper the researcher seeks to give an insight into diabetes as a chronic disease affecting millions of people globally and moving up in low and middle- incomes countries. If the disease is not diagnosed and adequately managed, care is not sought in a timely manner, serious complications are most likely, as are greater costs and poorer quality of life.

This research aims at developing sophisticated ML models to identify diabetes at an early stage using patient data; clinical, demographic and lifestyle information. That is why machine learning, which is a subfield of artificial intelligence, is effective in working with arrays of data, as well as identifying some patterns that might be overlooked in other cases. These patterns can allow for an ML model to give an accurate prediction of an individual's probability to develop diabetes in the future and help those at risk get a timely diagnosis.

The study incorporates other parameters of health status, presence of disease, blood sugar level, age, weight, dietary history and exercise regime, among others, to improve on the testimony of the model. When integrated, this information is useful in supporting healthcare providers in determining which patients are vulnerable and how best to manage patient care and apply preventative measures that are unique to an individual patient or group thereof through the ML-based approach.

1.4 Classification and prediction techniques

It is a data analysis tasks; it is a process of grouping data into discrete classes and labels and organized analysis and interpretation. Classification is the procedure to categorize records. One simple example of classification is to check the patient to see whether it is diabetes or not. The result will be either positive/negative. So, there are a particular number of choices. Prediction is another data analysis task. Prediction models predict numerical output.

Effectively identifying these problems is used to data mining and machine learning techniques. Effectively, identifying these problems are used to data mining and machine leaning technique. Classification is one of the significant methods for different diseases prediction. Classification is a technique used in data mining to group items or data points into different categories based on their characteristics. The goal is to assign each item in a collection to a specific category of class. Expected accuracy is achieved using diabetes patient datasets. Like K-nearest neighbor, naive bayes, logistic regression, support vector machine, decision tree, etc. There are several classification techniques that are used for diabetes prediction, including:

- **Logistic regression:** this is a technique that is used to test the probability of the binary nature of results like diabetes and non-diabetes.

- **Decision tree:** this method is like decision-making trees, which have flowchart characteristics: the decision made depends on certain factors. In this, it is reflected in the aim of predicting the result of a new patient by comparing him/her with other patients based on different characteristics.

- **Random forest:** Random forests are a form of bootstrapped aggregating that involves using several decision trees in an effort to have a better generalization of the results of the trees in order to effect more accurate forecast.

- **K- Nearest neighbor:** K-NN is one type of supervised machine learning algorithm that works in both regression and classification procedures.

- **Support Vector Machine:** Support Vector Machine (SVM) is a type of supervised learning technique for prediction in different kinds of problems. It is useful in both classification and regression.
- **Gradient boosting:** gradient boosting works exactly like a smart committee of decision trees. Every tree memorizes erroneous results obtained from the previous tree and works for better results progressively.
- **Ensemble**: This is a method that combines predictions from two or more base models using techniques such as boosting, bagging, stacking, and voting to improve performance.
- **Boosting**: This is one of the most popular ensemble methods. It combines two or more weaker learners to create a strong learner and works to reduce the error rate.
- **Bagging**: This is another popular ensemble method. It combines two or more weaker learners to create a strong learner and helps reduce overfitting.
- **Stacking**: This is another one of the most popular ensemble methods. Multiple base models are trained on the same dataset and then their prediction accuracies are combined to make a final prediction.

2. LITERATURE REVIEW

In this paper, discuss about the diabetes and their types then explain the diabetes symptoms. This study focused on diabetes prediction using PIMA Indian Diabetes dataset, implemented various machine learning techniques like ensemble and classification method. In this work, they evaluated the prediction results using various performance metrics or evaluation metrics like f1 score, recall, precision, confusion matrix and classification accuracy. In outliers' detection using and removal using three different method which is z-score, IQR and percentile method. This paper compared four different machine learning algorithm logistic Regression, random forest, support vector machine and gradient boosting then they found the best algorithm which is logistic regression algorithm gave 95.85% accuracy (Gupta et al., 2024).

This paper, work is diabetes prediction at an early stage. Diabetes dataset contains 2000 cases. K-Nearest Neighbors, SVM, Decision Tree, Logistic Regression, Random Forest classification algorithms are used to prediction. The Decision Tree gave the highest accuracy is 99% (Rani, 2020).

This paper, work is diabetes prediction at an early stage by using a machine learning classification algorithm and an ensemble learning method. SVM, Decision Tree, Random Forest, Decision Tree, Logistic Regression, KNN and Gradient Boosting classifiers algorithms are used. The Random Forest gave the highest accuracy of 77% (Soni et al., 2020).

The healthcare industry contains a lot of sensitive information, especially regarding diseases like type2 diabetes mellitus, which is a significant world health concern. Medical experts need a dependable predictive framework to analyze diabetes. This study advances artificial intelligence method that are helpful for analyzing large amounts of data and effective data mining methods can find patterns that can help in diagnosing diabetes more accurately. This study utilizes medical bioinformatics to predict diabetes, using R programming and the Pima Indian diabetes database. In this paper, the bootstrapping technique is used and after that applying decision tree, naïve bayes and k-nearest neighbor's (knn), logistic regression, ada boost, support vector machine compared. as a result, ada boost and support vector machine give the highest accuracy, 94.44%, as compared to the other two methods. The aim of the study is to develop the best prediction system to improve the accuracy of diagnosis and be helpful for healthcare professionals (Aada et al., 2019).

This paper, work is diabetes prediction with improved accuracy machine learning and pipeline model are applied for improving the accuracy of diabetes prediction. The Decision tree, Gaussian Naïve Bayes, Linear Discriminant

Analysis, Support vector classifier, Random Forest, Extra trees, Ada Boost Classifier, Bagging, KNN, Logistic Regression, Linear Support Vector Machine, Perceptron, Gradient Boosting algorithms are used to predict. This study was compared with various machine learning methods. The Logistic regression gave the highest accuracy 96%. The PIMA diabetes dataset and diabetes dataset are compared. The pipeline model is used for improving the accuracy of diabetes prediction. The Ada Boost algorithm gave the highest accuracy (Mujumdar et al., 2019).

The objective of this study is to design a model can predict diabetes with highest accuracy rate. There are three machine learning algorithms used namely Naive bayes Support vector machine and Decision tree are used in this experiment to predict diabetes at an early stage. In this paper PIMA Indian Diabetes Dataset are used. The performances of the entire three algorithms are evaluated on various measures like Accuracy, Recall, Precision and F-measure. The highest accuracy achieved 76.30% using Naïve Bayes classification algorithm (Sisodia et al., 2018).

The aim of research detecting diabetes using the PIMA Indian Diabetes Dataset, focuses on women's health. In this paper, two machine learning algorithms are used Support Vector Machine and Decision Tree algorithms. These algorithms help to predict diabetes diseases. Aim to simplify tasks like classification and clustering through automation. R studio framework is used for better analysis and visualizations. Support vector machine gives 82% accuracy (Rathore et al., 2017).

Data mining got a lot of attention for extracting important information from vast amounts of datasets, especially in fields like medical science. Data mining algorithms include naive bayes, association, classification, clustering and decision tree which are well known. These methods help to predict diabetes diseases at early stage for future diagnosis. In this paper, they aim to compare various data mining algorithms using the PID dataset to improve the early prediction of diabetes diseases. In this study, the PIMA dataset is used. The decision tree gives the best accuracy, 75.65% (Azar et al., 2018).

Machine learning is an application of artificial intelligence that is helping medical experts diagnose diseases better. Diabetes, a common health problem, is growing because of bad habits like unhealthy eating and not exercising. It helps to predict diabetes at an early stage because it can lead to many serious health issues like nerve damage, kidney damages, etc. This research focuses on classifying individual people with diabetes using classification algorithms such as Random Forest, Support Vector Machine, Logistic Regression, KNN, Gradient Boosting and Decision Tree. The experiment found that the KNN algorithm gave better results than when compared with other classified algorithms. The KNN gave high accuracy, 85% (Jeevan Nagendra Kumar et al., 2019).

The aim of this paper, work is proposed a deep neural network approach for diagnosis of diabetes, trained using ten-fold and five-fold cross validation fashion. In this experiment PIMA Indian diabetes dataset from the UCI repository. In this paper, implementation has been done by the deep neural network. Investigation has involved deep neural network technique to identify of diabetes diseases based on the several medical factors like pregnancy, blood pressure, age, body mass etc. The result five-fold cross-validation give highest accuracy is 98.35% (Ayon et al., 2019).

This research, work on diabetes disease was examined using a machine learning algorithm. The aim of this research is to create an effective model with high accuracy to predict diabetes diseases. Kaggle datasets are used; it consists of 9 columns and 768 rows. Logistic regressions, K-nearest neighbor, support vector machine algorithm, gradient boosting classifier algorithm are used. KNN algorithm is used to reduce the processing time. Support vector algorithm more efficiently handles a large number of features for the dataset. The gradient boosting classifier performs well on this type of dataset, giving a maximum accuracy of 81% (Panda et al., 2022).

Diabetes is a serious health problem worldwide, especially in developed nations. In year 2035, it's predicted that there will be about 592 million peoples worldwide who have diabetes. Nigeria is facing a big challenge with diabetes too. In this paper, the dataset of diabetes mellitus (DM) type2 was collected from the Murtala Mohammed Specialist Hospital, Kano and used to develop predictive machine learning models (supervised) based on random forest, logistic

regression, naïve bayes, k-nearest neighbor's, support vector machine and gradient boosting algorithm. Random forest was the best model found among the developed models with 88.76% in terms of accuracy, which is the highest accuracy. In terms of the receiver operating characteristic curve, both random forest algorithm and gradient boosting algorithm determined the highest predictive learning models, both around 86.28%. The model helps medical personnel by analyzing patient data to diagnose and predict type2 diabetes (Muhammad et al., 2020).

In this paper, discuss diabetes and symptoms. Information about diabetes prediction model. The PIMA dataset is used. The Dataset contains 8 factors like pregnancies, BMI, diabetes pedigree, insulin, age, blood pressure, BMI, diabetes pedigree, glucose and skin thickness. Four algorithms are used such as k-nn, logistic regression, random forest and support vector machine. The random forest gives the highest accuracy 92.18% (Punya et al., 2021).

Designed a diabetes prediction model with improved accuracy. The Pima Indian Diabetes Dataset and the Vanderbilt Dataset were used in this study. The Pima dataset contains nine features/variables, while the Vanderbilt dataset includes 16 features/variables: Patient Number, Cholesterol, Glucose, HDL, CHOL/HDL, Age, Gender, Height, Weight, BMI, Systolic BP, Diastolic BP, Waist, Hip, Waist-to-Hip Ratio, and Diabetes status.

In the paper, the datasets were split into training and testing sets with a ratio of 70:30. Various machine learning algorithms were employed, including logistic regression, decision tree, support vector machine (SVM), naïve bayes, k-nearest neighbor (KNN), and ensemble models.

To improve performance, ensemble methods such as boosting, bagging and stacking were utilized. The ensemble methods achieved better results compared to logistic regression. Additionally, k-fold cross-validation was applied to improve accuracy and ensure the model's reliability (Rajendra et al., 2021).

3. MATERIAL AND METHODOLOGY

3.1 Understanding the Goal

The objective is to employ machine learning (ML) techniques to diagnose defects, determine if a person has diabetes, or predict the likelihood of developing the disease in the future. This involves identifying patterns within data, such as blood sugar levels, age, blood pressure, weight, and other related activities.

3.2 Collecting Data

To train our machine learning algorithm, we require relevant datasets.

- Examples include free datasets like the Pima Indians Diabetes Dataset and those available on Kaggle. Many researchers have used the Pima dataset and Kaggle datasets, which contain various features and attributes. These attributes include a target variable, typically represented as {0, 1}, indicating the outcome used for prediction.

- medical record sheet or health survey. Some researchers used hospital datasets for training and testing the model for increasing the quality of data.

- It is to use machine learning (ML) to predict whether a person has diabetes or might develop it in the future. This involves analyzing patterns in data like blood sugar levels, age, blood pressure, weight, and lifestyle.

3.3 Data Preprocessing

Before using the data in ML models:

- Cleaning the Data: This step involves removing errors and handling missing values. Null values can be checked using methods like isnull() and duplicated(). Missing values can be replaced with appropriate values, such as the average, and the data should be normalized to ensure consistency.

- Standardizing: Standardization ensures that larger values (e.g., age) do not dominate smaller ones (e.g., blood sugar levels) by scaling the numbers to a uniform range, allowing the model to treat all features equally.

3.4 Split the Data: Divide it into2 sets are:

- **Training Set:** It is utilized for training the ML model.
- **Testing Set:** Used to evaluate the model's performance and ensure it produces accurate results.

Typically, the data is split in a ratio such as 80% for training and 20% for testing. This helps the model generalize well to unseen data.

3.5 Selection of Model

It however should be noted that as pointed out above there are actually many kinds of ML models that one may choose. Popular ones for diabetes prediction are:

- **Logistic Regression:** For binary classification that is for yes/no answers.
- **Decision Trees:** To make predictions based on questions (e.g., Is blood sugar high?).
- **Random Forest:** Combines many decision trees for better accuracy.
- **Support Vector Machine (SVM**): Finds the best way to separate people with and without diabetes.
- **K-Nearest Neighbor (KNN)**

3.6 Training the Model

- Fit the training data into the ML model.
- The model learns patterns (e.g., high blood sugar + obesity = higher diabetes risk).

3.7 Testing the Model

- Use the testing data to check if the model can correctly predict diabetes.
- Compare predictions with actual outcomes to measure performance.

3.8 Measuring Performance

In this phase, we check the performance by using the work model:

- **Accuracy:** Percentage of correct predictions.

$$\textbf{Accuracy= (TP + TN)/(TP+TN+FP+FN)}$$

- **Precision and Recall:** How well the model identifies people with diabetes.

$$\textbf{Precision=TP/(TP+FP)}$$

$$\textbf{Recall=TP/(TP+FN)}$$

- **F1 Score:** It described the measure test accuracy. The f1 score range is [0,1]

$$\textbf{F1=2*(Precision * Recall)/ (Precision + Recall)}$$

- **ROC Curve:** Shows how good the model is at distinguishing between people with and without diabetes.
- **Cross Validation:** Cross-validation is a method used to evaluate the performance of a model by dividing the dataset into training and validation sets multiple times like K-folds (K= 5 or 10 folds).

3.9 Improving the Model

- Adjust the model settings (called hyperparameters) like random search, grid search.
- Try different models or combine them.
- Use more or better-quality data for training.
- Using different ensemble method for improving the performance of the model.

3.10 Using the Model

Once the model is accurate enough, you can use it to predict diabetes in real-world cases, such as:

- Health screening tools.
- Apps or websites that provide early warnings for at-risk individuals.

4. RESULT AND ANALYSIS

Table 1 displays the performance results of previous work on machine learning methods used for diabetes prediction. It includes the comparison of various algorithms, such as logistic regression, decision trees, and ensemble methods, along with their accuracy, precision, recall, and f1 score.

Table 1: Summary of different Diabetes prediction model

Reference	Best algorithm	Accuracy	Precision	Recall	F1 score
Rathore et al. (2017)	SVM	82%	-	-	-
Mujumdar et al. (2019)	ADABOOST	98.80%	-	-	-
Azar et al. (2018)	DT	75.65%	-	-	-
Sisodia et al (2018)	NB	76.30%	0.759	0.763	0.76
Rani (2020)	DT	99%	-	-	-
Muhammad et al. (2020)	RF	88.76%	-	-	-
Soni et al. (2020)	RF	77%	-	-	-
Panda et al. (2022)	GB	81%	0.76	0.61	0.68
Aada et al. (2019)	SVM & ADABOOST	94.44%	-	-	-
Gupta et al. (2024)	LR	95.85%	-	-	-
Ayon et al. (2019)	Deep NN	98.35%	-	-	0.99
Punya et al. (2021)	RF	92.18%	-	-	-
Jeevan Nagendra Kumar et al. (2019)	KNN	85%	-	-	-

5. CONCLUSION

This study focused on machine learning algorithms for predicting diabetes disease with more accuracy. The literature study in the decision tree algorithm achieved the highest accuracy. The different performance values of machine learning algorithms are calculated on various measures. Train and test the data like Kaggle, Pima Indian Diabetes Dataset. The classification algorithm's maximum testing accuracy has been achieved. The study has collected various types of machine learning algorithms and ensembled them for the purpose of improving accuracy, sensitivity and specificity. The main goal of this study is to examine different methods to improve prediction accuracy. The survey focuses on exploring various models and techniques, especially ensemble methods, which showed better performance than individual models.

REFERENCES

Aada, A., Tiwari, S. (2019). Predicting Diabetes in Medical Datasets Using Machine Learning Techniques. *International Journal of Scientific Research & Engineering Trends, 5(2).*

Ayon, S.I., Islam, M.M. (2019). Diabetes Prediction: A Deep Learning Approach. *I.J. Information Engineering and Electronic Business, 2, 21-27.*

Azar, A. Ali, Y., Awais, M., Zaheer, K. (2018). Data Mining Models Comparison for Diabetes Prediction. *International Journal of Advanced Computer Science and Applications, 9(8).*

Gupta, P., Sindhu, R. (2024). Diabetes Prediction Using Machine Learning. *Journal of Electrical Systems*, 2244-2257.

Jeevan Nagendra Kumar, Y., Kameswari Shalini, N., Abhilash, P.K., Sandeep, K., Indira, D. (2019). Prediction of Diabetes using Machine Learning. *International Journal of Innovative Technology and Exploring Engineering, 8(7),* 2278-3075.

Kumari, V. A., Chithra, R. (2013). Classification of Diabetes Disease Using Support Vector Machine. *International Journal of Engineering Research and Applications, 3(2), 1797-1801*, www.ijera.com.

Mujumdar[a], A., V[b], D. V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Science Direct, Procedia Computer Science, 165, 292–299*, www.sciencedirect.com.

Muhammad, L. J., Algehyne, E. A., Usman, S. S. (2020). Prediction Supervised Machine Learning Models for Diabetes Mellitus.*SN Computer Science Springer Nature journal*, http://doi.org/10.1007/s42979-020-00250-8.

Panda, M., Mishra, D. P., Patro, S. M., Salkuti, S. R. (2022). Prediction of diabetes disease using machine learning algorithms. *International Journal of Artificial Intelligence, 11(1), 284-290*, 2252-8938.

Punya, V. P., Sindhu, M. N., Poornima, M. E., Brindashree, Rajeshwari, J. (2021) DIABETES PREDICTION USING MACHINE LEARNING TECHNIQUES. *INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS, 9(6),* 2320-2882.

Rajendra, P., Latifi, S. (2021). Prediction of diabetes using logistic regression and ensemble techniques. *Computer Methods and Program in Biomedicine Update 1*, 100032, http://doi.org/10.1016/j.cmpbup.2021.100032.

Rani, K. M. J. (2020). Diabetes Prediction Using Machine Learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2456-3307.

Rathore, A., Chauhan, S., Gujral, S. (2017). Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women, 8(5), 0976-5697,www.ijarcs.info, https://doi.org/10.26483/ijarcs.v8i5.3674.

Sisodia, D., Sisodia, D. S. (2018). Prediction of Diabetes Using Algorithm. *Procedia computer science, 132, 1578-1585*,www.elsevier.com/locate/procedia.

Soni, M., Varma, S. (2020). Diabetes Prediction using Machine Learning Techniques. *International Journal of Engineering Research & Technology, 9(9),* http://www.ijert.org, 2278-0181.