Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 09, September-2023

Diabetes Disease Prophecy : A Comprehensive Review of Machine Learning Approaches

Om Sohani
Dept. of Artificial Intelligence and Data Science
K.J. Somaiya Institute of Technology
Mumbai, India

Sakshi Zanjad
Dept. of Artificial Intelligence and Data Science
K.J. Somaiya Institute of Technology
Mumbai, India

Smita Prajapati
Dept. of Artificial Intelligence and Data Science
K.J. Somaiya Institute of Technology
Mumbai, India

Vraj Parekh
Dept. of Artificial Intelligence and Data Science
K.J. Somaiya Institute of Technology
Mumbai, India

Preksha Shah
Dept. of Artificial Intelligence and Data Science
K.J. Somaiya Institute of Technology
Mumbai, India

***Abstract*— High blood sugar levels are the hallmark of diabetes, a chronic metabolic disorder. Chronic Diabetes disease is a long-term medical condition characterized by persistently elevated blood sugar levels and the possibility of complications. This abstract provides a comprehensive summary of numerous research papers focusing on the application of machine learning techniques to the prediction and analysis of diabetes-related conditions. Collectively, these studies contribute to the development of accurate and timely predictive models for diabetes risk assessment and management. They employ a variety of machine learning approaches, such as ensemble learning, deep learning methods, and various algorithms, to predict chronic diabetes disease, diagnose the initial phase of diabetes, identify diabetes-related renal disease, and evaluate early diabetes risk. In addition, these papers investigate the analysis of diabetic prediction using machine learning on diverse datasets, also studying about the Type 1 and Type 2 diabetes Various Machine Learning algorithms like Random Forest (RF), K-Nearest Neighbors (K-NN), Linear Regression (LR), Adaptive Boosting (AdaBoost), Decision Tree (DT), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Convolutional Neural Network(CNN), Gradient Boost(GB)and Java48(J48) are used. Among these algorithms, the Random Forest and K-Nearest Neighbors (KNN) models have showcased outstanding performance, reaching impressive accuracy rates of 99.9% and 98.62%, respectively. This corpus of research highlights the significance of machine learning in enhancing our understanding of diabetes and improving patient outcomes through early detection and risk assessment.**

Keywords—Machine Learning, Diabetes, Random Forest, K-nearest neighbors, Type 1 Diabetes, Type2 Diabetes.

I. INTRODUCTION

Chronic diabetes, characterized by persistently elevated blood sugar levels, is a metabolic disorder that persists throughout life. It encompasses both Type 1 and Type 2 diabetes and poses serious health hazards if left untreated. Diabetes type 1 is an autoimmune disorder in which the body does not produce insulin. Insulin resistance is characteristic of type 2 diabetes, which is frequently associated with lifestyle factors such as diet and exercise. Both have an effect on glucose levels. Preventing complications such as cardiovascular disease, kidney problems, and nerve injury requires effective management, including lifestyle modifications and medications.

The following papers constitute a significant body of research devoted to harnessing the power of machine learning for the prediction, diagnosis, and management of diabetes, a chronic and pervasive disease that affects millions of individuals. In the first set of investigations, researchers investigate the development of chronic diabetes disease prediction models. These models aspire to provide precise and timely insights into the progression and management of diabetes, providing a potential lifeline for patients and healthcare professionals in search of effective intervention strategies.

Various machine learning approaches are used to investigate the first stage of diabetes prediction as we advance to the next frontier. These initiatives aim to identify early signs and symptoms, allowing for early interventions that can have a significant impact on the progression of disease and patient outcomes. In addition, the focus extends to the prognosis of chronic kidney disease in diabetic patients using ensemble learning techniques to improve the accuracy of such

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 09, September-2023

predictions. This intersection between diabetes and renal health is crucial, given that kidney complications are a common and severe complication of diabetes.

In the field of early risk assessment, researchers investigate the use of deep learning techniques to predict the likelihood that a patient will develop diabetes. These techniques provide a robust and nuanced approach to identifying at-risk individuals, thereby facilitating preventative measures. In addition, the analysis of diabetic prediction using machine learning algorithms, particularly when applied to diverse datasets such as the BRFSS dataset, provides valuable insights into population-level trends and variations, thereby facilitating the development of public health strategies.

The pursuit of accurate prediction extends to differentiating Type 1 and Type 2 diabetes, two distinct forms of the condition. Achieving this differentiation is essential for individualizing patient treatment and care plans. In conclusion, these research endeavors represent a multidimensional exploration of machine learning's potential to revolutionize our understanding and management of diabetes, with the promise of enhanced patient care, early intervention, and public health strategies.

II. Literature Survey

In this study, Gurpreet Singh et al. (2022) investigate machine learning models for chronic diabetes prediction. It probably examines the various prediction models already in use and suggests a well-known predictive model for this disease utilizing machine learning methods. [1]

The study of Minhaz Uddin Emon et al. (2021) focuses on using machine learning to predict diabetes in its early stages. This article will discuss current strategies with a focus on the use of machine learning for diagnosing diabetes in its early stages. [2]

In this study, Md. Omar Faruque et al. (2021) explore the use of ensemble learning to anticipate chronic kidney disease in diabetic individuals. This research does a review of earlier work on using ensemble approaches to forecast renal disease in the setting of diabetes. [3]

A mini-review paper on machine learning techniques for early diabetes prediction was put up by Rouaa Alzoubi et al. in 2022. It is a summary of the body of research on the use of machine learning for early diabetes detection. [4]

Deep learning techniques for early diabetes risk prediction were introduced by Luyao Xu et al. in 2021. The goal of this work was to analyze earlier studies that used deep learning to identify diabetes risk factors early on. [5]

Using machine learning methods, Rumini et al. (2022) performed a prediction analysis of diabetes mellitus. In this article, the author conducted a survey of the literature on machine learning-based diabetes prediction models. [6]

On the basis of the BRFSS dataset, Dr. Lakshmi H.N. et al. (2023) examined diabetic prediction using machine learning methods. The BRFSS dataset and machine learning techniques will likely be used in this paper's review of pertinent literature on diabetes prediction. [7]

The objective of this study by Menaka V et al. (2022) is to reliably predict both Type 1 and Type 2 diabetes. In this, the author reviews existing literature on predictive models for distinguishing between these two diabetes types using machine learning. [8]

III. Methodology

A. *Random Forest (RF):* An ensemble approach that, in order to lessen overfitting and boost accuracy, builds numerous decision trees during training and integrates their predictions. It works well for tasks involving classification and regression.

B. *Support Vector Machine (SVM):* A technique for supervised learning that locates the best hyperplane in high-dimensional space to divide data points. It is appropriate for regression and linear and non-linear classification issues.

C. *k-Nearest Neighbors (K-NN):* A straightforward classification technique that labels data points in feature space according to the dominant class among the k-nearest neighbors. It is adaptable and simple to comprehend.

D. *Naive Bayes (NB):* A Bayes-based probabilistic classifier that relies on the "naive" assumption of feature independence. It is frequently employed in spam filtering and text classification.

E. *Adaptive Boost (ADA Boost):* An ensemble technique that combines a number of weak learners and builds a strong classifier by iteratively changing sample weights. In categorization tasks, it is frequently employed.

F. *Linear Regression (LR):* The statistical method of linear regression, which is frequently employed in prediction and correlation analysis, models the relationship between variables by fitting a linear equation to the data.

G. *Multilayer Perceptron (MLP):* A feed-forward neural network with many interconnected layers of neurons. Numerous machine learning tasks, particularly deep learning, are appropriate for it.

H. *Decision Tree (DT):* A classification and regression model based on trees. To create predictions at leaf nodes, it divides the data into subsets based on feature values.

I. *Convolutional Neural Networks (CNNs):* Deep neural networks made for handling data with a grid structure, like photographs. For feature extraction and hierarchical learning, they make use of convolutional layers, which are popular in computer vision.

J. *Gradient Boosting (GB):* A common ensemble learning technique for classification and regression applications that builds decision trees progressively to repair mistakes in prior trees.

K. *Java 48(J48):* In machine learning, the classification algorithm J48 is employed. It is useful for tasks like data categorization and prediction since it creates decision trees to categorize data based on attribute values.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 09, September-2023

IV. DISCUSSIONS

A. Confusion Matrix

When describing the performance of a classification model on a collection of data for which the true values are known, a confusion matrix is a table that is frequently employed. It helps you understand how well a binary classification algorithm works, in particular. TP, FP, TN, and FN are the four entries in the matrix, which stands for true positive, false positive, and true negative. In the context of diabetes prediction using a machine learning model, Fig. 1. represents the confusion matrix in context of a diabetic person.

	Actual	
Predicted	**Diabetic (P)**	**Non Diabetic (N)**
Diabetic (P)	TP	FP
Non Diabetic (N)	FN	TN

Fig. 1. Confusion Matrix

- True Positive (TP): The model correctly predicted that a person has diabetes.
- True Negative (TN): The model correctly predicted that a person does not have diabetes.
- False Positive (FP): The model incorrectly predicted that a person has diabetes when they do not (Type I error).
- False Negative (FN): The model incorrectly predicted that a person does not have diabetes when they actually do (Type II error).

Choosing evaluation criteria that are appropriate for the unique features of the problem is crucial when applying machine learning models to predict diabetes. Here are some appropriate evaluation measures given that predicting diabetes is a binary classification task (diabetic or not):

1. Accuracy

Evaluates the model's overall accuracy. It is the proportion of cases that were successfully predicted to all instances.

$$\text{Accuracy} = \frac{TN + TP}{TP + TN + FP + FN} \quad (1)$$

2. Precision

Measures the accuracy of total positive predictions. It is the ratio of correctly predicted positive observations to the total predicted positives.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

3. Recall(Sensitivity)

It determines whether the model is able to include every positive outcome. It measures the proportion of observations in the actual class that were correctly predicted as positive to all other observations.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

4. F1-value

Precision and recall are summed to form the F1-Score. It balances precision and recall also its useful when you want to strike a balance between false positives and false negatives.

$$\text{F1-value} = \frac{2(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \quad (4)$$

5. Specificity

This metric assesses the model's ability to correctly identify negative cases.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

V. RESULTS AND ANALYSIS

Various machine learning techniques have been employed in studies aiming to effectively predict diabetes at an early stage, including classical and advanced methods like Deep learning and Convolutional neural networks. Table I provides an overview of the algorithms utilized in these research efforts, along with their corresponding performance metrics. It's noteworthy that all of these algorithms have demonstrated commendable accuracy in their predictions. However, the Random forest and K-nearest neighbors (KNN) models have demonstrated superior performance, with an impressive accuracy rate of 99.9% and 98.62% for early diabetes prediction. Additional metrics like precision, recall, and F1 score are essential for a more nuanced assessment, especially when identifying true positive cases. Some applications prioritize higher precision, minimizing false positives. The K-nearest neighbors model has the highest recall value, while the random forest model excels in precision. Both random forest and KNN models have shown remarkable performance, but the Random forest model is more significant for identifying actual positive cases of the disease.

TABLE I. ALORITHMS AND THEIR ACCURACY

Sr. No.	Algorithms	Accuracy (%)
1.	Random Forest	99.9%
2.	K-Nearest Neighbors	98.62%
3.	Support Vector Machine	96%
4.	Linear Regression	94%
5.	AdaBoost	94%
6.	Decision Tree	93.88%
7.	Multi Layer Percepetron	92%
8.	Gradient Boost	87.31%
9.	Naïve Bayes	79.22%
10.	J48	79%
11.	CNN	76%

The following Fig. 2 shows the graphical analysis of all of machine learning techniques on comparing with each other according to their performance ranging them from highest to lowest accuracy.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 09, September-2023

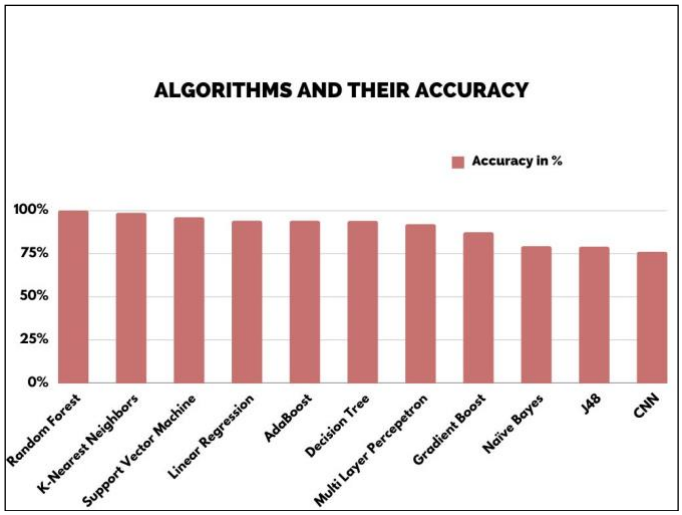

Fig. 2. Comparison of Algorithms

VI. CONCLUSION

This compilation of studies highlights the importance of early diabetes prediction in mitigating its impact. Various machine learning algorithms, including Random Forest, K-Nearest Neighbors, Support Vector Machines, etc. have been applied and compared for their effectiveness. Random Forest and KNN models have shown exceptional accuracy, with rates as high as 99.9% and 98.62% respectively. Deep learning techniques and optimization strategies show potential for further accuracy. The research emphasizes the benefits of leveraging machine learning models under medical guidance for early diagnosis, with future refinement through ensemble techniques and feature selection methods. Ultimately, these advancements have the potential to significantly improve patient outcomes and reduce the burden on healthcare systems worldwide.

REFERENCES

[1] L. H.N., A. S. Reddy and K. Naidu, "Analysis of Diabetic Prediction Using Machine Learning Algorithms on BRFSS Dataset," 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2023, pp. 1024-1028, doi: 10.1109/ICOEI56765.2023.10125804.

[2] M. V, L. V, S. M and R. Pari, "Accurate Prediction of Type 1 and Type 2 Diabetes," 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, 2022, pp. 1117-1121, doi: 10.1109/ICICICT54557.2022.9917938.

[3] L. Xu, J. He and Y. Hu, "Early Diabetes Risk Prediction Based on Deep Learning Methods," 2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Yibin, China, 2021, pp. 282-286, doi: 10.1109/PRAI53619.2021.9551074.

[4] R. Alzoubi and S. Harous, "Machine Learning Algorithms for Early Prediction of Diabetes: A Mini-Review," 2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, United Arab Emirates, 2022, pp. 401-405, doi: 10.1109/ICECTA57148.2022.9990240.

[5] Rumini, S. N. Wahyuni, B. Sudaryatno and A. Pramudyantoro, "Prediction Analysis of Diabetes Mellitus Based on Machine Learning Algorithm," 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2022, pp. 209-214, doi: 10.1109/ISRITI56927.2022.10052794.

[6] G. Singh, Mamta, J. Singh and M. Gahlawat, "Prominent Prediction Model for Chronic Diabetes Disease Using Machine Learning," 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 2022, pp. 1099-1105, doi: 10.1109/ICECA55336.2022.10009399.

[7] M. U. Emon, M. S. Keya, M. S. Kaiser, M. A. islam, T. Tanha and M. S. Zulfiker, "Primary Stage of Diabetes Prediction using Machine Learning Approaches," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 364-367, doi: 10.1109/ICAIS50930.2021.9395968.

[8] M. O. Faruque, S. Hossain and A. A. Marouf, "Predicting Chronic Kidney Disease of Diabetes Patients using Ensemble Learning," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 2021, pp. 1743-1747, doi: 10.1109/ICCES51350.2021.9489137.