International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Comprehensive Review on Diabetes Disease Prediction and Risk Modeling Using Machine Learning Algorithms

Abdul Aamir Khan[1*], Dr. B. K. Sharma[2]

Submitted: 17/10/2023 **Revised**: 20/11/2023 **Accepted**: 16/12/2023

Abstract: The increasing global prevalence of diabetes has intensified the demand for accurate and early diagnostic systems. Diabetes is a worldwide health issue that necessitates precise prediction techniques. This study examines research that uses clinical data and machine learning approaches to predict diabetes. Common preprocessing procedures include encoding categorical data, resolving missing values, and normalization. To improve model performance, dimensionality reduction techniques such as Principal Component Analysis (PCA) and feature selection are employed. Metrics like accuracy, precision, recall, F1-score, and AUC-ROC are used to compare supervised learning algorithms like Random Forest, Support Vector Machines (SVM), k-Nearest Neighbours (k-NN), Logistic Regression, and Decision Trees. Many research employ small datasets, which affects generalizability even while accuracy is excellent. The study emphasizes the necessity for diversified datasets and therapeutically relevant, interpretable models while highlighting shortcomings in model interpretability and validation.

Keywords: *Clinical Data, Diabetes Prediction, Healthcare, Logistic Regression, Machine Learning.*

1. INTRODUCTION

Chronic metabolic disorder leading to abnormal high levels of sugar [glucose] in blood, resulting in severe complications if diabetes mellitus goes became unmanaged. The rate of diabetes has been increasing said the World health Organization (WHO),to an estimated 422 million people with diabetes globally in 2014. This alarming trend emphasizes the need of effective prediction and management strategies to ensure diabetes' minimal adverse effect on public health. Typically, diabetes diagnosis and management utilise methods that involve clinical assessments and laboratory tests that could be cumbersome and may not be prompt.

This advent of machine learning (ML) in the recent years has led to the development of innovative ways of data analysis and predictive modelling in the field of healthcare. By analysing huge amount of data, machine learning algorithms can detect patterns, present predictions with a great accuracy. As one of the best tools for early detection and intervention in diabetes, this capability is of particular value. Using machine learning models, various datasets, such as Pima Indians Diabetes Database and UCI Diabetes Dataset, have been used for training and assessment of the diabetes risk prediction.

In recent years, many researchers have been engaged with applying machine learning techniques on diabetes prediction and explored different algorithms and methodologies. It has been proven that machine learning models including logistic regression, decision trees, random forests, support vector machines and neural networks can predict diabetes risk very well from clinical and demographic data. For example, in the studies using Pima Indians Diabetes Database it has been observed that models like Random Forests and Support Vector Machines have high accuracy rates and sometimes above 85%. Furthermore, feature selection methods, such as recursive feature elimination and correlation analysis, have been used to improve diabetes modeling with the feature selection techniques by selecting the most diabetes predictive feature. Nevertheless, while many studies seem to report promising results the methodologies, the evaluation metrics, and the datasets used vary significantly among researches that makes it difficult to make comparison among results from different researches. An emerging subject that uses cutting-edge computational approaches to improve the precision and effectiveness of diabetes risk assessment is the prediction of diabetic illness by machine learning.

[1*] *Department of Computer Science & Applications, Mandsaur University, Mandsaur,*
Madhya Pradesh, 458001, India
[2] *Department of Computer Science & Applications, Mandsaur University, Mandsaur,*
Madhya Pradesh, 458001, India
[*] *Corresponding author: kka68291@gmail.com[1]*
Corresponding authors: dr.balkrishnasharma@meu.edu.in[2]

Over 400 million people worldwide suffer from diabetes, and its incidence is still rising. As a result, there is an urgent need for novel approaches that can identify the development of this chronic illness before it shows clinical symptoms. Researchers want to find important risk variables and enhance early detection techniques by combining machine learning algorithms with sizable datasets. This would improve patient outcomes and healthcare administration.[1][2][3] Several machine learning approaches, such as ensemble methods, deep learning, and semi-supervised learning, are included in the diabetes prediction methodology. These methods build strong predictive models that can adjust to the unique characteristics of each patient by using both labeled and unlabeled data. This methodology's essential elements include feature selection, handling data imbalances, and evaluation metrics to guarantee the efficacy and Dependability of the model. Additionally, a comprehensive understanding of diabetes risk is made possible by the combination of lifestyle factors and demographic data [3] [4]. Despite its potential, the discipline has several obstacles to overcome, especially when it comes to data protection and machine learning models' interpretability. Discussions are sparked by worries about security and ethical use brought up by the dependence on private patient data. regarding the use of privacy-preserving technologies like federated learning and blockchain. Making machine learning models interpretable is also becoming more and more important in order to guarantee that medical professionals can use them in clinical settings [5] [6]. Going forward, improving diabetes prediction models requires interdisciplinary cooperation and the incorporation of cutting-edge technologies. The goal of ongoing research is to increase model applicability, interpretability, and accuracy across varied populations, thereby propelling advancements in diabetes preventive and care techniques. By encouraging collaborations between scientists, medical professionals, and tech specialists, the healthcare industry aims to provide solutions that address the intricacies of diabetes as well as the changing field of machine learning in healthcare [7] [8].

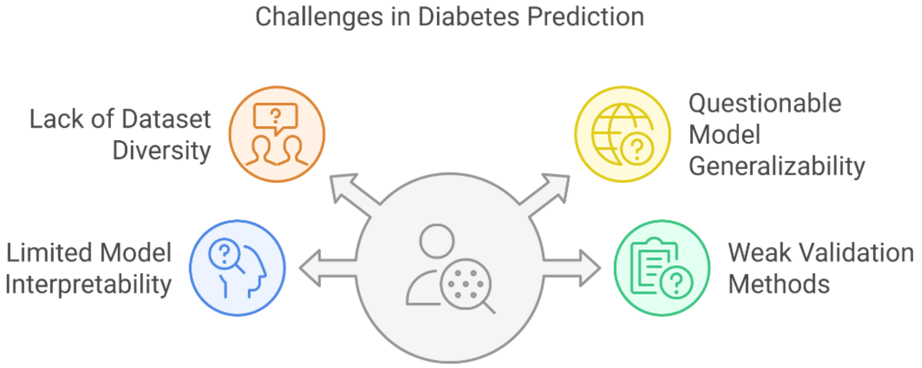

Challenges in Diabetes Prediction
Lack of Dataset Diversity
Questionable Model Generalizability
Limited Model Interpretability
Weak Validation Methods

Fig. 1 challenges in diabetes prediction

2. Literature Review

The use of machine learning algorithms to forecast the development of diabetes based on diagnostic criteria is examined in the work "Predictive Modelling for Diabetes Using Machine Learning" (2024) by Shriya Aishani Rachakonda et al. Random Forest, k-Nearest Neighbours (k-NN), Decision Trees, Support Vector Machines (SVM), and Logistic Regression are among the supervised learning classification methods used in this work. Out of all of these, the Random Forest algorithm performed the best, attaining 84% accuracy, 83% precision, 78% recall, 80% F1-score, and 0.86 AUC. The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) dataset, which supplied the clinical characteristics required for model training and testing, served as the basis for the study [9].

Machine Learning-Based Diabetes Prediction System, 2023, According to Kundan Kumar et al., machine learning methods like Support Vector Machine, Numpy, and Decision Trees can predict diabetes with 80% accuracy. The Support Vector Machine Classifier had an 80% accuracy rate in predicting diabetes.Machine learning modeling is an effective way to classify people with and without diabetes. 80% [10].

Predictive Modeling Using Machine Learning Techniques and Classifiers for Diabetes Mellitus

Disease Prediction and Type Classification, 2022, B. Ahamed and others The research provides machine learning techniques for precisely diagnosing and classifying diabetes mellitus. Several supervised learning classification techniques were used to create a trustworthy diabetes predicting model for this specific study: Random Forest, k-Nearest Neighbors (k-NN), Decision Trees, Support Vector Machines (SVM), and Logistic Regression The Random Forest algorithm was the best, achieving 84% accuracy, 83% precision, 78% recall, 80% F1-score, and 0.86 AUC across the Pima Indians Diabetes Dataset and survey [11].

Machine Learning-Based Diabetes Prediction: Examination of 70,000 Clinical Database Patient Records, 2022, According to Sony M. et al., one machine learning technique that can be used to precisely predict diabetes is called Random Forest. By 2040, it is predicted that 642 million people worldwide will have diabetes, demonstrating the disease's rising prevalence. The Pima Indian Diabetes Dataset and the Diabetes 130-US hospitals data collection for 1999–2008 [12].

A New Framework for Predicting Diabetes in Healthcare Using Machine Learning Methods, 2022 R. Krishnamoorthi and associates Based on ingenious machine learning, the study proposes a diabetes prediction paradigm with an accuracy of 83%. The study's findings can be useful to academics, stakeholders, students, and health professionals involved in diabetes prediction research and development. The proposed machine learning-based architecture achieved an accuracy score of 86% [13].

Predicting Diabetes Through Machine Learning Algorithms, 2022, V. Yamana Machine learning algorithms can predict up to 90% of diabetes cases. In order to understand why certain machine learning classifiers performed badly, the study examined model overfitting and underfitting. In order to predict diabetes with a consistent and optimal accuracy of 90%, the study employed machine learning algorithms [14].

Predicting Diabetes Using Machine Learning Algorithms, 2021 Arwatki Chen Lyngdoh et.al Five machine learning algorithms for predicting diabetes disease are evaluated in the study; the KNN classifier has an accuracy of up to 76%. The study identified the reasons behind some classifiers' inability to consistently achieve high accuracy by analyzing training and testing accuracy and looking for signs of overfitting or underfitting. 76% accuracy was attained in the diabetes dataset, which included data on risk factors and consequences associated with the disease [15].

Machine learning-based predictive modeling and analytics for diabetes, 2021 Kaur Harleen et al. Using the Pima Indian diabetes dataset, the study creates and evaluates five different machine learning models for diabetes prediction and classification. The study developed and analyzed five different machine learning models—SVM, k-NN, ANN, and MDR—to classify patients as either diabetic or non-diabetic. The Pima Indian diabetes dataset shows that the SVM-linear model has the best accuracy of 0.89 and precision of 0.88, the k-NN model has the best recall and F1score of 0.90 and 0.88, and the AUC values of the two models are 0.90 and 0.92 [16].

Diabetes Disease Prediction Using the 2021 Machine Learning Model, Sharma Amandeep et al. The study suggests a machine learning model that uses logistic regression, ANN, Naïve Bayes, and decision trees to predict diabetes. The study developed a diabetes prediction model using supervised machine learning algorithms like logistic regression, Naïve Bayes, decision trees, and artificial neural networks. The accuracy of the decision tree approach is 76.52%, the Naïve Bayes algorithm is 76.95%, and logistic regression is 80.43%.

The accuracy of the artificial neural network classifier on the Pima Indian diabetes dataset is 75.21% [17].

Table 1 Overview of Existing Studies on Diabetes Prediction Models:

Paper	Abstract summary	Main Findings	Accuracy	Dataset
Predictive Modelling For Diabetes Using Machine Learning,	The paper applies machine learning based algorithms to predict the	the study employed supervised learning	With an accuracy of 84%, precision of 83%, recall of 78%, F1-score of 80%, and	the National Institute of Diabetes and Digestive and

2024, Shriya Aishani Rachakonda et. al. [9]	occurrence of diabetes by using diagnostic characteristics.	classification techniques, such as Random Forest, k-Nearest Neighbours (k-NN), Decision Trees, Support Vector Machines (SVM), and Logistic Regression.	AUC of 0.86, the Random Forest algorithm performed the best.	Kidney Diseases (NIDDK) dataset.
Diabetes Prediction System Using Machine Learning, 2023, Kundan Kumar et.al [10]	Diabetes may be predicted with 80% accuracy using machine learning techniques like Decision Trees, Numpy, and Support Vector Machine.	Diabetes was predicted by the Support Vector Machine Classifier with 80% accuracy. People with and without diabetes can be efficiently categorised using machine learning modelling.	80%	Not mentioned
Diabetes Mellitus Disease Prediction and Type Classification Involving Predictive Modeling Using Machine Learning Techniques and Classifiers, 2022, B. Ahamed et. Al. [11]	The study offers machine learning methods for accurately predicting and categorising diabetes mellitus.	A reliable predictive model of diabetes for this particular study was made using several supervised learning classification methods: Logistic Regression, Support Vector Machines (SVM), Decision Trees, k-Nearest Neighbours (k-NN), Random Forest	The highest performance was achieved by the Random Forest algorithm with accuracy of 84%, precision = 83%, recall = 78%, F1-score = 80%, and AUC = 0.86.	the Pima Indians Diabetes Dataset, and survey.
Prediction of Diabetes Using Machine Learning: Analysis of 70,000 Clinical Database Patient Record, 2022, Sony M et.al. [12]	Random forest is one of the machine learning algorithms that can be used to accurately predict diabetes.	The number of patients suffering from diabetes is expected to reach 642 million globally by 2040, indicating the growing prevalence of the disease.	_	Diabetes 130-US hospitals for the years 1999-2008 Data Set and the Pima Indian Diabetes Dataset.

A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques, 2022 R. Krishnamoorthi et.al [13]	The study suggests an 83% accurate paradigm for diabetes prediction based on clever machine learning.	Health professionals, stakeholders, students, and academics working on diabetes prediction research and development can all benefit from the study's conclusions.	The accuracy score of the suggested machine learning-based architecture was 86%.	Survey
Diabetes Disease Prediction By Using Machine Learning Algorithms,2022, V.Yamuna. [14]	Up to 90% of diabetes cases can be predicted using machine learning algorithms.	The study looked at model overfitting and underfitting to determine why some machine learning classifiers performed poorly.	The study used machine learning algorithms to predict diabetes with a consistent and best accuracy of 90%.	Not mentioned
Diabetes Disease Prediction Using Machine Learning Algorithms,2021, Arwatki Chen Lyngdoh et.al. [15]	The study assesses five machine learning algorithms for diabetic illness prediction; the KNN classifier achieves an accuracy of up to 76%.	By analysing training and testing accuracy and searching for indications of overfitting or underfitting, the study determined why certain classifiers failed to reach consistent and high accuracy.	Achieved the accuracy of 76%	diabetes dataset containing information on risk factors and outcomes related to diabetes.
Predictive modelling and analytics for diabetes using a machine learning approach,2021, Harleen Kaur et.al. [16]	The research develops and assesses five alternative machine learning models to predict and categorise diabetes in the Pima Indian diabetes dataset.	In order to categorise patients as either diabetes or non-diabetic, the study created and examined five distinct machine learning models: SVM, k-NN, ANN, and MDR.	SVM-linear model provides best accuracy of 0.89 and precision of 0.88, k-NN model provided best recall and F1scoreof0.90and0.88, AUC value of SVM-linear and k-NN model are 0.90 and 0.92	the Pima Indian diabetes dataset,
Prediction of Diabetes Disease Using Machine Learning	The research proposes a machine learning model to predict	Using supervised machine learning algorithms such as logistic regression,	logistic regression displays 80.43% accuracy, Naïve Bayes	the Pima Indian diabetes dataset,

Model,2021, Amandeep Sharma et.al. [17]	diabetes utilising methods including decision tree, Naïve Bayes, ANN, and logistic regression.	Naïve Bayes, decision trees, and artificial neural networks, the study created a model for diabetes prediction.	algorithm is 76.95%, decision tree algorithm has an accuracy of 76.52%, Artificial neural network classifier has 75.21% accuracy	

Research Gap:

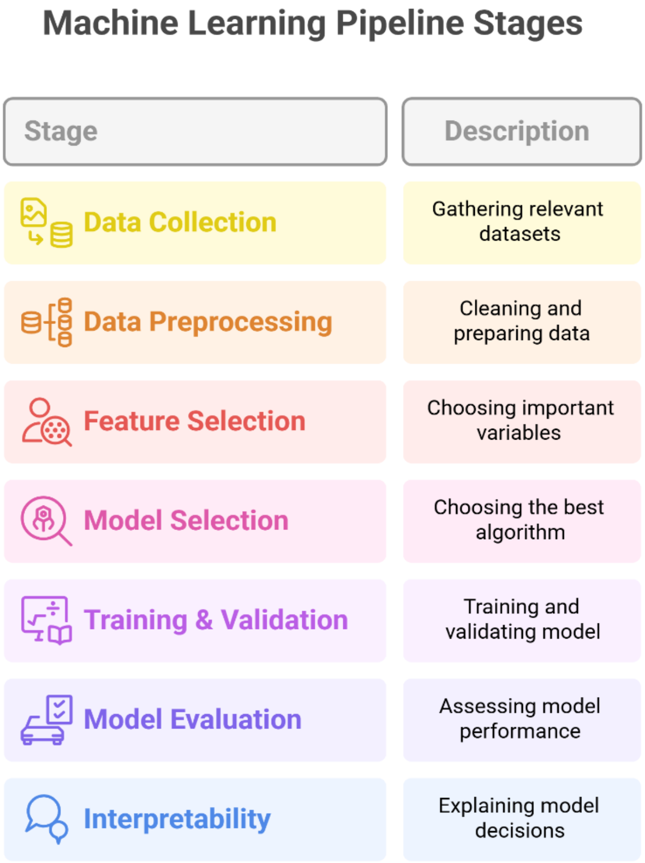
Although there have been tremendous strides taken in this area of machine learning, there are still gaps in literature pertaining to predicting diabetes. The first limitation is the simple fact that the datasets used are not diverse because most investigations heavily depend on the Pima Indians Diabetes Database and similar datasets, which might reflect only the subset of people who have diabetes and therefore may not be true to the entire population. Furthermore, the generalizability of the models developed is in question due to this. Further, while all algorithms achieve high accuracy, there is little focus on interpretability of the model, and clinical applicability to the predictions. Furthermore, the validation methods used in a lot of studies are not robust since they generally rely on small samples sizes or do not provide external validation, which is prone to over fit and degrade generalization capability in real world.

3. METHODOLOGY

The approaches used to prediction of diabetes, and hence the methodologies employed in the reviewed studies, are quite different. Such common practices typically include steps for data pre processing like normalization, missing values handling and categorical variable encoding. Reducing dimensionality using feature selection techniques is commonly carried out to improve model performance, and PCA and a multitude of statistical tests may be used. Studies also vary in the kinds of machine learning algorithms' choices, which sometimes consider comparing more than one algorithm to determine what the best training model is for their material hard drive. Typically, model performance is evaluated using evaluation metrics like accuracy, precision, recall, area under the receiver operating characteristic curve (AUC-ROC) to give insight to the best and worst algorithms.

The following are the main steps that make up the suggested methodology:
Data The cleaning: Taking care of missing values and making sure the data types are right to enable efficient analysis. Data processing :One of the most important steps in using the Pima Indians Diabetes Database is data preprocessing. Effectively managing missing values is one of the main responsibilities. Certain factors in this study, such as blood pressure, skin thickness, insulin, glucose, and body mass index.

Normalization and standardization are done to get the data ready for modeling.
In order to guarantee that every feature contributes equally to the model's performance, the data must be scaled.The models' accuracy rates differed greatly from one another. One model, for example, demonstrated a strong performance across a variety of data sources with an accuracy of 93.22% on the Pima Indian dataset and 98.95% on the Mendeley dataset. Generally speaking, 78% accuracy is regarded as good, but 81% accuracy is regarded as exceptional, especially when verified using 10-fold cross-validation.



Machine Learning Pipeline Stages

Stage	Description
Data Collection	Gathering relevant datasets
Data Preprocessing	Cleaning and preparing data
Feature Selection	Choosing important variables
Model Selection	Choosing the best algorithm
Training & Validation	Training and validating model
Model Evaluation	Assessing model performance
Interpretability	Explaining model decisions

Fig. 2 Stages of ML model

In given Fig2 The standard methodology used in studies predicting diabetes using machine learning is depicted in the flowchart. It describes important procedures that lead to the final prediction, such as data pretreatment, dimensionality reduction, algorithm selection, model training, and evaluation. The accuracy and dependability of the model are enhanced by this methodical approach.

Several algorithms that use both labeled and unlabeled data are part of the machine learning-based methodology for diabetic illness prediction. The main emphasis is on semi-supervised learning techniques, which are especially beneficial in situations where obtaining labeled data is expensive or limited. This improves learning effectiveness by enabling models to be trained on bigger datasets that contain unlabeled data. For example, significant amounts of unlabeled data have been integrated with labeled instances using Laplacian SVM (LapSVM), which has enhanced accuracy and generalization capacities [1]. The literature analysis, which focused on a carefully selected dataset of 2,351 publications, was organized in accordance with the standards laid forth by Marcus et al. (45). Using Term Frequency-Inverse Document Frequency (TF-IDF) in a methodical manner was used in the examination [2][3]. A systematic four-point scale was created for both qualitative and quantitative evaluations in order to preserve objectivity and decrease prejudice. Several factors were deemed crucial in the context of diabetes prediction.

Lifestyle characteristics and demographic information were emphasized as important predictors. According to studies, age and gender were found to be important variables. affecting metabolic health and insulin sensitivity [2]. Furthermore, the models incorporated modifiable lifestyle characteristics including physical activity and smoking, highlighting the necessity for dynamic prediction models that can adjust to shifting health behaviors [2]. More and more people are following the trend of merging data from many sources, like lifestyle and genetic data, to create more thorough risk assessments [2].

Several strategies were used to solve data imbalance, a prevalent problem in classification problems, with the goal of rebalancing the distribution of data. Techniques like cost-sensitive learning, ensemble methods, and resampling were used to improve enhance generalization ability and model performance [1][2]. These tactics are essential for ensuring that minority groups are fairly represented in model predictions and reducing the bias caused by unequal sample sizes across categories [1]. Accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC) are the main metrics used to assess the performance of the model [1][8][18]. Together, these measures provide a thorough evaluation of the model's effectiveness. Because in Precision, precision and recall offer information about the model's capacity to accurately identify positive examples, whereas accuracy quantifies the percentage of true findings among all cases [19][20].

The Pima Indians Diabetes Database, which includes medical diagnostic records from 768 female patients of Pima Indian heritage who are 21 years of age or older, is a well-known dataset used in machine learning for diabetes prediction. It has nine characteristics that support binary classification tasks linked to diabetes diagnosis, including eight predictors and one target variable. The dataset is a standard for machine learning research because of its use in creating and assessing different machine learning algorithms [20].

Numerous clinical and biometric parameters necessary for diabetes diagnosis are included in the collection. Among other things, important characteristics include blood pressure, age, body mass index (BMI), and glucose levels. The purpose of this dataset is to use these measurements to diagnose and forecast if a patient has diabetes. [21]. With 500 cases being diabetic negative and 268 being diabetic positive, the dataset is very frequently imbalanced, which makes predictive modeling difficult because of possible biases in model results.

An explanation of the main machine learning techniques frequently used to predict diabetes.

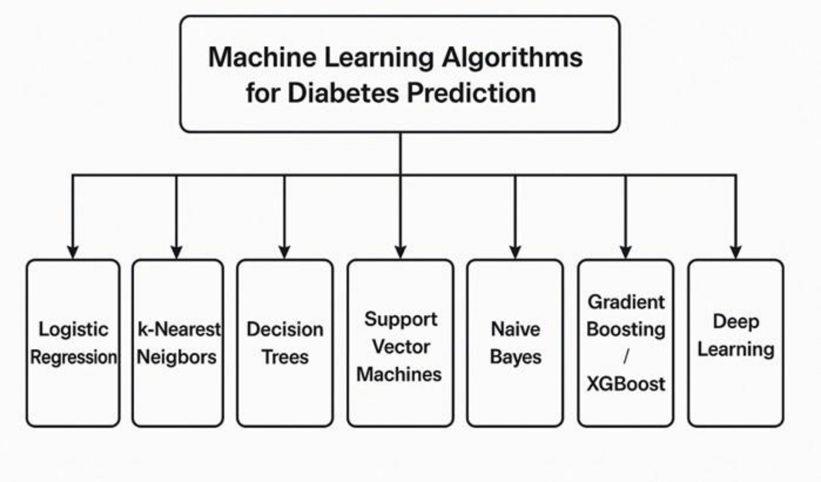


Fig. 3 Algorithms for diabetes prediction

There are extensive use of Machine learning algorithms for diabetes prediction as machine learning algorithms are able to discover complex patterns from clinical data. Logistic Regression (LR) is one of the most popular statistical model for binary classification, which estimates what the probability that the given input belongs to which class, such as diabetic or not diabetic. On the other hand, k against Nearest Neighbours (k against NN) is a simple yet useful instance based algorithm that classify new datasets based on the majority class such that its closest neighbour. Decision Trees (DT) are approach that combing them to a tree in which the internal node is the decision, the branches represent the outcomes and the leaves are the class label. Random Forest (RF) combine multiple decision trees into ensemble method to increase accurancy and live off overfitting. While in high dimension these things are good shots SVM's can find the hyperplane that separates the classes with the largest margin. Steeped in Bayes' theorem and relying on features to be independent, the simplicity of Naive Bayes (NB) has led it to deliver good performance. Gradient Boosting Machines (GBM) and XGBoost are two powerful ensemble technique that build models one by one sequentially, and each model is correcting error of previous model, and they are particularly useful for structured data. Deep Learning (DL) with artificial neural networks with several layers is highly suitable for finding complex relations in large data and it has demonstrated high accuracy in diabetes prediction.

To achieve a full evaluation, a lot of performance measures were used, such as Accuracy, Precision and Recall, and F-Measure based on the confusion matrix. Further, more advanced metrics were used including the Sensitivity, Specificity, the Area Under the ROC Curve (AUC), and the Matthews Correlation Coefficient (MCC). These diverse movements provide depth to understand the cost balance of different elements of model performance (e.g., how to evaluate that there are more false positives than negatives). This is necessary in order to select the best model for practical and clinical applications of the diabetes prediction.

Given how much machine learning depends on data, having accessible and varied datasets is essential to improving model performance. Future initiatives should concentrate on compiling extensive statistics from diverse demographics, taking into consideration various Factors, medical problems, and lifestyle choices that affect the risk of diabetes. Sharing anonymised patient data through collaborations between institutions can aid in the development of more broadly applicable models that can be used with a variety of populations.

4. CONCLUSION

Drawing insights from range of recent studies critically reviewed the application of various machine learning (ML) algorithms in diagnosis of diabetes based on a critical analysis of various research works carried out in the recent times. One can certainly see that ML techniques, such as Random Forest, Support Vector Machines (SVM), k-Nearest Neighbours (k-NN), Logistic Regression, and Deep Learning, are very good to predict. These models, when trained over the structured clinical sets such as Pima Indians Diabetes Database have performance of high accuracy, precision and AUCscores commonly. It is noteworthy that ensemble and deep learning models have time and again outperformed conventional classifiers in predictive ability.

Although these promising results show, various research limitations remain. Limitations in generalizability ensuing from the reliance of the findings on benchmark datasets which are rarely diverse. In addition, if interpretability is not addressed, which is critical for clinical acceptance of these models, they are generally under addressed in this regard. Further, although relatively few studies employ utilization or external validation strategies that can help insure that the models are robustly applicable across populations and clinical settings, this potential vulnerability is not addressed by most current publications.

Table 2 table summarizing the algorithms, datasets used, and reported accuracy rates

Algorithm(s)	**Dataset**	**Accuracy (%)**
Random Forest, k-NN, Decision Trees, SVM, Logistic Regression	NIDDK Dataset	84%
SVM, Numpy, Decision Trees	Not Specified	80%
Random Forest, k-NN, Decision Trees, SVM, Logistic Regression	Pima Indians Diabetes Dataset + Survey	84%
Random Forest	Pima Indians Dataset, Diabetes 130-US hospitals (1999–2008)	Not Specified
Not specified (general ML-based architecture)	Not Specified	83%, proposed model: 86%
Not specified	Not Specified	90%
k-NN and others	Diabetes dataset with risk factors and consequences	76% (k-NN)
SVM (Linear), k-NN, ANN, MDR	Pima Indian Diabetes Dataset	SVM: 89%, k-NN: 90%
Logistic Regression, ANN, Naïve Bayes, Decision Trees	Pima Indian Diabetes Dataset	LR: 80.43%, NB: 76.95%, DT: 76.52%, ANN: 75.21%

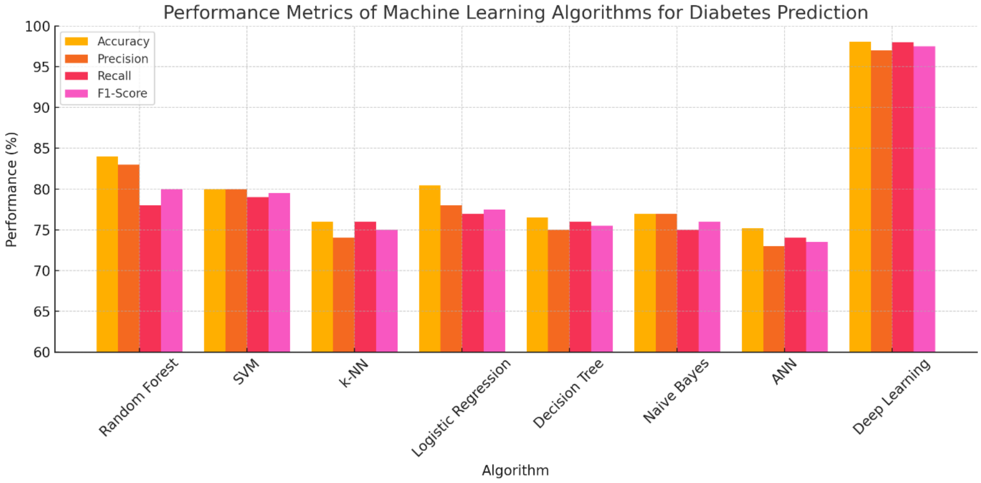


Fig shows the Accuracy, Precision, Recall, and F1-Score are the key performance metrics for which a visual comparison is made between different machine learning algorithms for diabetes prediction.

5. FUTURE SCOPE

To increase the clinical applicability and relevance of machine learning (ML) models for diabetes prediction, future studies need to address a number of important points. In the first place, diversification of datasets is required. Subsequent research should seek to leverage large, multi-center, and real-world datasets such as electronic health records and longitudinal data to make the models widely applicable and to reduce possible biases that come from using limited or homogeneous datasets. Interpretability is also of great importance. The use of explainable AI (XAI) methods will be crucial in making ML predictions transparent and interpretable, which is critical to winning the confidence of clinicians and enabling the clinical uptake of these models. Additionally, better validation methods need to be put in place to guarantee the robustness and generalizability of the models. This entails the application of strict validation procedures, i.e., external dataset verification and k-fold cross-validation, which serve to minimize the risk of overfitting and give a better assessment of the performance of the models. In conclusion, the incorporation of these ML platforms into clinical systems should be the order of the day. Future products should be built such that integration into current electronic health record systems and clinical decision support systems is seamless, such that real-time stratification of risk and diagnosis happens in alignment with prevailing clinical workflows.

References

[1] C.-Y. Chou, D.-Y. Hsu, and C.-H. Chou, “Predicting the onset of diabetes with machine learning methods,” *Journal of Personalized Medicine*, vol. 13, no. 3, p. 406, Feb. 2023, doi: 10.3390/jpm13030406.

[2] D. Kalla, N. Smith, F. Samaah, and K. Polimetla, “Enhancing early diagnosis: Machine learning applications in diabetes prediction,” *Journal of Artificial Intelligence &amp; Cloud Computing*, pp. 1–7, Mar. 2022, doi: 10.47363/jaicc/2022(1)191.

[3] Y. Qin *et al.*, “Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type,” *International Journal of Environmental Research and Public Health*, vol. 19, no. 22, p. 15027, Nov. 2022, doi: 10.3390/ijerph192215027.

[4] M. Talebi Moghaddam *et al.*, “Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm,” *BMC Medical Research Methodology*, vol. 24, no. 1, Sep. 2024, doi: 10.1186/s12874-024-02341-z.

[5] K. Abnoosian and R. Farnoosh, “Prediction of Diabetes Disease Using an Ensemble of Machine Learning Multi-Classifier Models,” *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4179050.

[6] P. Alagumariappan *et al.*, “Optimized hybrid machine learning framework for early diabetes prediction using electrogastrograms,” *Scientific Reports*, vol. 15, no. 1, Mar. 2025, doi: 10.1038/s41598-025-93495-3.

[7] N. E. Costea, E. V. Moisi, and D. E. Popescu, “Comparison of Machine Learning Algorithms for Prediction of Diabetes,” in

2021 16th International Conference on Engineering of Modern Electric Systems (EMES), IEEE, Jun. 2021, pp. 1–4. Accessed: Apr. 23, 2025. [Online]. Available: https://doi.org/10.1109/emes52337.2021.9484116

[8] Tasin, T. U. Nabil, S. Islam, and R. Khan, “Diabetes prediction using machine learning and explainable AI techniques,” Healthcare Technology Letters, vol. 10, no. 1–2, pp. 1–10, Dec. 2022, doi: 10.1049/htl2.12039.

[9] S. A. Rachakonda, S. Pudipedi, and T. S. S. Angel, “PREDICTIVE MODELLING FOR DIABETES USING MACHINE LEARNING,” *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 08, no. 008, pp. 1–16, Aug. 2024, doi: 10.55041/ijsrem37149.

[10] K. Kumar and A. Tomar, “Diabetes Prediction System Using Machine Learning,” in *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, IEEE, Nov. 2023, pp. 286–291. Accessed: Apr. 23, 2025. [Online]. Available: https://doi.org/10.1109/icaiccit60255.2023.10466034

[11] B. S. Ahamed, M. S. Arya, S. K. B. Sangeetha, and N. V. Auxilia Osvin, “Diabetes Mellitus Disease Prediction and Type Classification Involving Predictive Modeling Using Machine Learning Techniques and Classifiers,” *Applied Computational Intelligence and Soft Computing*, vol. 2022, pp. 1–11, Dec. 2022, doi: 10.1155/2022/7899364.

[12] S. M. Kuriakose, P. Basa Pati, and T. Singh, “Prediction of Diabetes Using Machine Learning: Analysis of 70,000 Clinical Database Patient Record,” in *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, Oct. 2022, pp. 1–5. Accessed: Apr. 23, 2025. [Online]. Available: https://doi.org/10.1109/icccnt54827.2022.9984264

[13] R. Krishnamoorthi *et al.*, “A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques,” *Journal of Healthcare Engineering*, vol. 2022, pp. 1–10, Jan. 2022, doi: 10.1155/2022/1684017.

[14] V.Yamuna, “V.Yamuna, D.Ushanthi, Chaitanya, B., sri, Y., & T.Jagadish (2022). Diabetes Disease Prediction By Using Machine Learning Algorithms.,” *Semanticscholar*, 2022.

[15] C. Lyngdoh, N. A. Choudhury, and S. Moulik, “Diabetes Disease Prediction Using Machine Learning Algorithms,” in 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), IEEE, Mar. 2021, pp. 517–521. Accessed: Apr. 23, 2025. [Online]. Available: https://doi.org/10.1109/iecbes48179.2021.9398759

[16] H. Kaur and V. Kumari, “Predictive modelling and analytics for diabetes using a machine learning approach,” *Applied Computing and Informatics*, vol. 18, no. 1/2, pp. 90–100, Jul. 2020, doi: 10.1016/j.aci.2018.12.004.

[17] Sharma, A., “Prediction of Diabetes Disease Using Machine Learning Model.,” *Semanticscholar*, 2021.

[18] J. Kaliappan *et al.*, “Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets,” *Frontiers in Artificial Intelligence*, vol. 7, Aug. 2024, doi: 10.3389/frai.2024.1421751.

[19] Abousaber, H. F. Abdallah, and H. El-Ghaish, “Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets,” Frontiers in Artificial Intelligence, vol. 7, Jan. 2025, doi: 10.3389/frai.2024.1499530.

[20] N. Halder, “Exploring the Pima Indians Diabetes Dataset: Advanced Data Analysis Techniques in Python,” *Medium*, Jan. 03, 2024. Accessed: Apr. 23, 2025. [Online]. Available: https://medium.com/@HalderNilimesh/exploring-the-pima-indians-diabetes-dataset-advanced-data-analysis-techniques-in-python-f02cba6f9f35

[21] S. Erzurumlu, “Optimizing Healthcare Predictions with CatBoost: A Study on the Pima Indians Diabetes Dataset,” LinkedIn, Sep. 16, 2024. Accessed: Apr. 23, 2025. [Online]. Available: https://www.linkedin.com/pulse/optimizing-healthcare-predictions-catboost-study-pima-erzurumlu-ipuvf/