

A DATA-DRIVEN HYBRID FEATURE-BASED MACHINE LEARNING APPROACH TO EARLY DIABETES PREDICTION

B. Mouli Chandra¹, K. Sudharshan², C.H. Silpa³, Jeyaprabhavathi Perumal⁴, D.V. Ashok⁵, Kommu Kishore Babu⁶ and M. Mohana Prasad⁷

^{1,2,5}Department of Electrical and Electronics Engineering, QIS College of Engineering and Technology, India

^{3,7}Department of Artificial Intelligence and Machine Learning, QIS College of Engineering and Technology, India

⁴Department of Computing and Technology, Arden University, United Kingdom

⁶Department of Computer Science Engineering, Vignan's Foundation for Science, Technology and Research, India

Abstract

Diabetes is a major global health concern that can lead to severe complications if not detected early. Type 2 diabetes often grows progressively, with many individuals unaware of their condition until significant health damage occurs. Early detection is therefore essential to enable timely medical intervention and lifestyle changes to improve treatment outcomes and reduce long-term risks. In this study, significant health factors, including insulin, blood pressure, age, glucose levels, BMI, and family history, are analysed to predict diabetes in early stages by applying machine learning (ML) models. Numerous techniques, including Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Trees (DT), Support Vector Machines (SVM), Random Forest (RF), and Extreme Gradient Boosting (XG Boost), have been employed for estimation. A hybrid model combining Principal Component Analysis, Random Forest, and XG Boost was developed to enhance prediction accuracy. Principal Component Analysis selects the features, Random Forest ensures model stability and reduces overfitting, while XG Boost improves performance by minimising prediction errors. The proposed hybrid model has achieved a higher accuracy of 99.25%, surpassing the other models included in our study. The model also proved strong recall and precision, confirming its effectiveness in categorising diabetic and non-diabetic cases. These results confirm that the proposed model can serve as an effective and reliable tool for early detection of diabetes and help healthcare professionals.

Keywords:

Type 2 Diabetes, Early Prediction, Machine Learning, Random Forest, XG Boost, Hybrid Model

1. INTRODUCTION

Millions of people worldwide suffer from diabetes, the most common chronic illness that presents serious problems for healthcare systems. It is a leading cause of severe health complications, including kidney failure, cardiovascular disease, nerve damage, and vision impairment. In 2024, diabetes caused 3.4 million deaths, one every 9 seconds. Diabetes accounted for at least USD 1 trillion in health spending, a 338% increase over the previous 17 years, as per the International Diabetes Federation. Diabetes facts and statistics demonstrate the growing global impact on individuals, families, and countries. According to the most recent IDF Diabetes Atlas [1], 11.1%, or one in every nine, of the adult population (20-79 years) has diabetes, with more than four in ten being ignorant of the condition. According to IDF projections, one out of every eight people, or approximately 853 million, would have diabetes by 2050, a 46% increase. More than 90% of diabetic individuals have type 2 diabetes, which is caused by socioeconomic, demographic, environmental, and genetic factors.

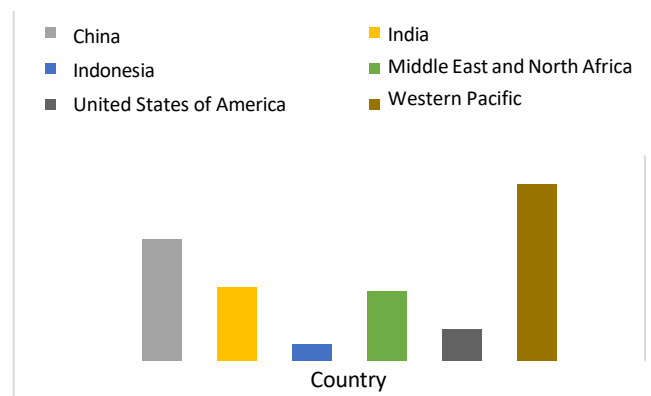


Fig.1. The top countries affected by diabetes in 2024.

The Fig.1 shows the countries most affected by diabetes in 2024. Type 2 diabetes, the most common form, often develops gradually, making early diagnosis critical in preventing long-term complications. Many individuals remain unaware of their condition until it has already caused significant damage, making early detection essential for better disease management and improved patient outcomes [2]. The primary contributors to the growth of type 2 diabetes are urbanisation, reduced levels of physical activity, an aged population, and rising rates of overweight and obesity. However, diabetes can be mitigated by introducing prevention measures for type 2 diabetes, as well as providing early identification and adequate treatment for all kinds. These precautions can support persons with the illness in avoiding or delaying problems.

Despite advancements in medical research, current diagnostic methods still have limitations. Traditional screening approaches, such as fasting blood sugar tests and HbA1c tests, are effective but often require repeated testing over time to confirm a diagnosis. These methods can be slow, costly, and sometimes inaccessible in remote or resource-limited areas [3]. Additionally, by the time diabetes is diagnosed through conventional methods, significant health deterioration may have already occurred. This delay in diagnosis contributes to a higher risk of complications, increased medical costs, and a greater burden on healthcare systems worldwide.

Advancements in Artificial Intelligence (AI) have created new opportunities to transform early diagnosis and disease prediction. Large amounts of healthcare data can be processed and analysed by ML algorithms, which might then spot trends that conventional diagnostic methods might miss [4]. Early intervention and individualised treatment plans are made possible by ML's ability to make predictions more quickly and accurately by utilising data-driven decision-making. An ML-based approach to early

diagnosis not only enhances efficiency but also reduces healthcare costs by minimising the need for extensive laboratory testing and late-stage treatments. Furthermore, such models can be integrated into wearable devices and mobile health applications, enabling individuals to monitor their risk levels and take proactive measures to prevent diabetes [5]. In this study, ML techniques are explored for their potential to improve early diabetes detection. The study evaluates several methods, such as KNN, RF, DT, SVM, LR, and XG Boost. Unlike traditional diagnostic tests, ML models can analyse a combination of risk factors in real-time, providing valuable insights into an individual's likelihood of developing diabetes before symptoms become severe.

1.1 OBJECTIVES OF THE WORK

1. To develop a hybrid ML predictive model that is not only accurate but also cost-effective and easily accessible.
2. To transform the way diabetes is diagnosed and managed, making preventive care more efficient and widely available by integrating AI into healthcare.
3. To provide medical practitioners with a tool that helps them to spot those at risk early on, enabling prompt lifestyle changes, treatment, and improved disease control by leveraging the power of ML.

Additionally, this study advances the general objective of using AI in healthcare prevention. As the global burden of diabetes continues to rise, the need for innovative, technology-driven solutions becomes increasingly urgent.

Section 2 discusses the related work to the selection of the proposed model. Section 3 provides the theories and methods used in the proposed model. Section 4 describes the proposed model and its workflow. Section 5 evaluates the proposed model with comparative results. Section 6 concludes the work by providing insights into the topic for future studies.

2. RELATED WORK

With the rising prevalence of diabetes worldwide, researchers have explored various techniques to enhance early diagnosis and prediction. Traditional diagnostic methods, while effective, often require repeated testing and can be time-consuming. ML has revolutionised diabetes prediction by providing accurate, scalable, and cost-effective solutions. Simple models, including Decision Trees and KNN, offer interpretability and ease of use, while advanced models, such as XG Boost and deep learning approaches, deliver higher accuracy. Hybrid models further improve predictive power by combining the strengths of multiple techniques. As research continues, these ML methods hold great promise for enhancing early diagnosis, enabling timely interventions, and ultimately reducing the impact of diabetes on public health. Recent works on various ML models have been employed, each with unique pros and cons, to predict diabetes [6]. According to Belson [7], the simplest and most widely used medical diagnosis model is the Decision Tree Classifier (DTC). This model functions similarly to a flowchart, with

each leaf node denoting the decisive forecast and each internal node representing a decision based on a specific health indicator. Decision Trees are popular because they provide clear and interpretable results, helping healthcare professionals understand

how a diagnosis is made. However, they can sometimes overfit the data, indicating they may not perform well when tested on new datasets. Breiman [8] developed the Random Forest model as an advanced bagging ensemble learning model to overcome the limitations of Decision Trees, such as overfitting. RF generates more accurate and consistent results by building numerous trees and combining their predictions rather than depending on just one. By employing a technique called Bootstrap Aggregating, it randomly selects different portions of the data to train each tree, thereby reducing the risk of overfitting. This approach enhances the reliability of Random Forest, making it particularly effective in medical applications where data can often be noisy or incomplete. Vladimir introduced the Support Vector Machine, effective for distinguishing between classes, such as diabetic and non-diabetic patients. SVM finds the optimal possible boundary (hyperplane) that separates the groups with maximum margin. This approach is useful when working with structured datasets that have clear patterns [9]. However, SVM is highly ignorant of computing large datasets. Therefore, choosing the right parameters can sometimes be challenging. Evelyn et al. presented the K-Nearest Neighbors algorithm, a powerful algorithm that categorises incoming data points based on similarity to pre-existing data. It examines the 'k' closest neighbors and assigns a class based on the majority vote. KNN performs well when data points naturally form clusters, such as groups of diabetic and non-diabetic individuals. However, its performance rests on the choice of 'k' and can slow down significantly when dealing with large datasets [10]. Another well-known technique is Naive Bayes, based on the 18th-century Bayes' theorem developed by Reverend Thomas Bayes [11]. This probabilistic classifier assumes that all features are independent. While this assumption is not always realistic in medical datasets, Naive Bayes remains effective in many cases due to its speed and simplicity. It is mostly helpful when working with data that has many dimensions, such as genetic or biochemical markers for diabetes prediction. Diabetes prediction has extensively utilised Artificial Neural Networks (ANNs). Nitesh et al. have conducted studies on prognosis and treatment prediction for Type 2 diabetes using ML classifiers and deep neural networks [12]. Convolutional Neural Networks (CNNs) have also been adapted for medical applications. In diabetes prediction, CNNs can help extract important patterns from complex datasets, such as analysing retinal scans to detect diabetic retinopathy. When combined with traditional models, CNNs can enhance accuracy and provide insights into disease progression. According to Pradeep Kumar Roy, CNNs have been successfully implemented for retinal images to detect diabetic retinopathy, thereby improving early diagnosis and treatment strategies [13]. Recurrent Neural Networks (RNNs) support in assessing time-based health data, including tracking blood glucose levels over time. These models are perfect for forecasting illness patterns and maintaining patient records since they are made to analyze sequential input and retain historical information. Rabby et al [14] presents a new method for predicting blood glucose levels that takes sensor errors into account. It employs a deep RNN model based on stacked Long Short Term Memory (LSTM). This method leverages LSTM's capability to capture long-term dependencies in healthcare

data, assisting doctors in monitoring diabetes progression and recommending timely interventions. Extreme Gradient Boosting is one of the most advanced ML techniques in medical research.

This model constructs multiple small decision trees sequentially, with each new tree correcting the errors of the previous ones. XG Boost is highly efficient and can manage missing data, making it an excellent choice for diabetes prediction, where patient records often contain incomplete information. Compared to Random Forest, XG Boost is more optimised for performance and frequently achieves better accuracy. Researchers such as Shuqi Liu have demonstrated the effectiveness of XG Boost in predicting diabetes, emphasising its accuracy compared to other models like SVM, RF, and KNN.

To improve outcomes, researchers have recently looked into hybrid models that combine two or more statistical approaches. For instance, combining error correction and variance reduction results in more accurate predictions [15]. Similarly, other 2-dimensional models can be integrated with 1-dimensional models to leverage both feature capabilities. Combining classification and regression models utilises necessary features to enhance accuracy. The optimisation also plays a role in the estimator's calculation, thereby affecting the ML models' prediction performance. These hybrid models enhance accuracy, making them valuable tools for early detection of diabetes [16]. Therefore, this research develops an efficient hybrid predictive ML model to close the gap between conventional diagnostic techniques and cutting-edge data-driven healthcare solutions. This research uses a feature selection model, an optimised model, and two ML models to develop a forecasting model for diabetes prediction.

3. METHODS AND THEORIES

This study proposes a hybrid ML model that increases the robustness, accuracy, and efficiency of diabetes prediction by combining the strengths of Random Forest and XG Boost. Although both models perform well individually, their integration allows for leveraging complementary advantages, resulting in a more dependable predictive system.

3.1 RANDOM FOREST MODEL

Random Forest, a popular and powerful ML method, is widely used in classification and regression tasks. It was crucial to the development of the proposed diabetes prediction model. In healthcare, particularly in the early detection of diabetes, Random Forest has garnered attention for its ability to handle complex, real-world datasets with high precision. During training, the model builds numerous decision trees and then combines their predictions to generate a final output. In classification tasks, the combined judgment of these trees undergoes a process called majority voting, which enhances the model's overall accuracy and stability [17]. The principle behind Random Forest is based on the idea that a group of varied and relatively weak learners can collectively create a strong learner. Each decision tree in the forest is trained on a random subset of the dataset, and at each split in the tree, a random collection of characteristics is examined. This fosters a high degree of diversity among the trees, thereby reducing the likelihood of all trees committing the same errors. Consequently, Random Forest is particularly robust against overfitting, a common issue with individual decision trees, and it maintains high accuracy even when the input data is noisy or incomplete, a common challenge in medical datasets [18].

Bagging (bootstrap aggregating) and random feature selection are the two primary methods used in the Random Forest construction process. In order to train various decision trees, bagging entails selecting several random samples from the dataset (with replacement). This ensures that each tree sees a slightly different view of the data, making them unique. Random feature selection, on the other hand, ensures that not all trees use the same features to make decisions, which further enhances diversity. Together, these techniques allow Random Forest to achieve a good balance between bias and variance, leading to a robust and reliable model. During the evaluation phase of our study, Random Forest consistently achieved high accuracy levels, reaching approximately 87% when tested independently. This performance is notably higher than many traditional models and aligns with, or surpasses, results reported in previous studies on diabetes prediction. Furthermore, the model's capacity to identify both diabetic and non-diabetic patients was known by its high precision-recall ratio. Its performance remained stable across various runs, and it showed minimal sensitivity to small changes in the dataset, highlighting its robustness [19].

Despite the advantages of Random Forest, it is important to know its complications. One drawback is that the model may become computationally demanding as the number of trees rises. Additionally, while it provides feature importance scores, the internal workings of the ensemble are still considered a "black box" to some extent. This means that although we can understand which features are important overall, it is challenging to trace how a specific prediction was made for an individual case, something critical in clinical decision-making. Despite these limitations, Random Forest has proven to be an ideal base model in our hybrid approach. Its ability to provide accurate and stable predictions made it the ideal candidate to generate initial predictions, which were then further refined using XG Boost.

3.2 XG BOOST MODEL

Gradient Boosting (GB) is a highly effective ML technique that has garnered significant attention for its capacity to produce robust predictive models, particularly in structured data problems such as healthcare. In this study, Gradient Boosting was employed to complement Random Forest in a hybrid model aimed at the early detection of diabetes. The fundamental concept of gradient boosting involves building models sequentially, with each new model learning from the errors of its predecessor. This iterative correction of mistakes enhances the overall system's accuracy over time, establishing Gradient Boosting as one of the most potent tools for classification problems. Unlike Random Forest, which builds several trees independently before aggregating their predictions, Gradient Boosting constructs decision trees one at a time. Each new tree focuses on addressing the residual errors from the prior trees. These residuals represent the discrepancy between actual and expected outcomes. By minimising these discrepancies, Gradient Boosting sharpens the predictions with every iteration. The method employs a technique known as gradient descent to optimise performance, which centres on minimising a loss function (such as log loss for classification) by adjusting the model in a direction that most effectively reduces error [20].

XG Boost (Extreme Gradient Boosting), a scalable and extremely effective variant of the conventional Gradient Boosting

technique, was used to implement Gradient Boosting. XG Boost offers several enhancements, including

regularisation to prevent overfitting, parallel computation for faster training, and native support for missing data. The BMI, age, blood pressure, glucose levels, and insulin indicators made it an ideal choice for processing the healthcare data used in this research. Gradient Boosting's effectiveness rests on its capacity to detect subtle patterns in data. It is particularly effective in datasets where relationships between features are not straightforward, which is often the case in medical datasets [21]. Additionally, because it builds on the weaknesses of previous models, it has a natural ability to fine-tune predictions, leading to high levels of accuracy. Gradient Boosting is not without its challenges. This method's main drawback is that it can be computationally demanding, particularly when building a lot of trees. Each tree is dependent on the previous one, meaning the process cannot be easily parallelised (although XG Boost addresses this to some extent). Additionally, because Gradient Boosting focuses intensely on minimising error, it can sometimes lead to overfitting, particularly if the model is not carefully tuned. To avoid this, regularisation strategies, including restricting tree depth, modifying the learning rate, and early stopping, are frequently employed [22].

4. PROPOSED MODEL

The flow diagram in Fig.2 depicts the essential phases involved in constructing the diabetes prediction model. It starts with data collecting and pre-processing, then trains using a hybrid strategy of Random Forest and XG Boost to increase accuracy.

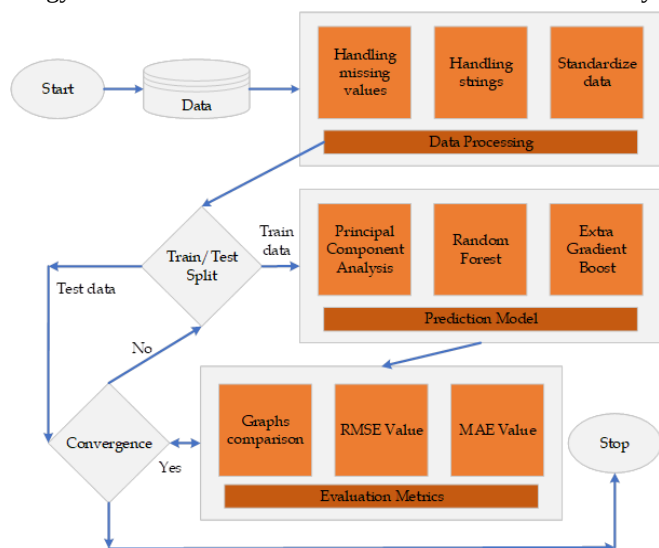


Fig.2. Flow chart for the Proposed model

The model is then assessed using performance measures to confirm its dependability. Finally, it is deployed for real-time diabetes prediction, aiding in early diagnosis and medical intervention.

4.1 DATA PREPROCESSING

Preparing the raw data for analysis is the first step in the procedure. This entails handling the most pertinent datatypes that influence diabetes prediction, adjusting or scaling, or altering

datatypes to guarantee consistency, and addressing missing values by imputing or eliminating incomplete data. These procedures are essential for ensuring the data consistent, and clean.

4.2 PRINCIPAL COMPONENT ANALYSIS

From the dataset, pertinent characteristics are chosen by principal component analysis, including age, BMI, blood pressure, and blood sugar levels. By removing superfluous or unnecessary features, this stage streamlines the model and enhances its interpretability and performance. This model helps to understand the important features for predicting diabetes [23].

4.3 RANDOM FOREST MODEL

After preprocessing and feature selection of the data, the model is trained using the Random Forest model. Random Forest is great at reducing overfitting and gives reliable baseline predictions. In our examination, we used Random Forest on a dataset that included important health markers such as blood pressure, insulin levels, age, BMI, glucose levels, and family history. These features are recognised as significant in assessing an individual's risk of developing diabetes. Random Forest not only handled this data efficiently but also provided insights into the importance of each feature, allowing us to understand which health factors contributed most to the prediction outcome.

4.4 XG BOOST MODEL

When used alongside Random Forest in a hybrid framework, Gradient Boosting adds depth and detail to predictions, allowing for more reliable and meaningful outcomes. Its error-correction capabilities, adaptability, and precision make it an excellent tool in healthcare applications, where accurate predictions are essential. Once the data has been learned by the Random Forest model, then passed on to XG Boost for further training. XG Boost, being a powerful boosting algorithm, builds on the strengths of Random Forest by focusing on the harder-to-predict cases and improving accuracy. It fine-tunes the predictions by correcting errors made in earlier steps.

By combining three models, first Principal Component Analysis for feature selection, Random Forest for stability, and then XG Boost for precision and refinement to ensure that the prediction for diabetes is both robust and highly accurate.

4.5 MODEL EVALUATION AND PARAMETER TUNING

Recall, accuracy, precision, and F1 score are used to assess the performance of the selected models. The procedure is repeated with changes in the train/test ratio made to the data if the model fails to fulfil the intended performance criteria.

4.6 TESTING AND COMPLETION

The model goes through a final validation to make sure it is prepared for deployment once it satisfies all performance requirements. With this, the development process is finished, and an ML model-based application that can precisely predict diabetes at an early stage has been developed.

5. RESULTS AND DISCUSSION

In this study, the proposed model is compared with various models, including LR, DT, SVM, RF, XG Boost, and hybridised random forest and extreme gradient boost models, as shown in TABLE I. The proposed hybrid ML model achieved outstanding performance in predicting diabetes, as demonstrated by the evaluation metrics. With a high accuracy of 99.25%, the model accurately classified the majority of cases in the test dataset. It achieved a precision of 0.99 for the

non-diabetic class (0) and 1 for the diabetes class (1), successfully reducing false positives, as shown in Fig.3. Similarly, recall values were 1 and 0.98 for classes 0 and 1, respectively, indicating an effective reduction of false negatives. This is critical in medical diagnostics because a missed diagnosis might have catastrophic implications. The model’s overall robustness and dependability were demonstrated by the F1-score, which balances precision and recall, reaching 0.99 for both classes. These metrics, calculated over a total of 400 test samples, validate that the model is not only accurate but also consistent across various evaluation aspects. Such high-performance results position the model as a promising tool for real-world early diabetes detection, potentially enabling healthcare providers to identify at-risk individuals more efficiently.

Table.1. Proposed Models’ vs Other Models’ Performance For Diabetes Prediction

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.78	0.79	0.76	0.77
Decision Tree	0.82	0.83	0.81	0.82
SVM	0.85	0.86	0.84	0.85
KNN	0.86	0.87	0.86	0.86
Gradient Boost	0.94	0.95	0.94	0.94
Random Forest	0.96	0.96	0.95	0.95
RF + XGB	0.982	0.98	0.98	0.98
RF + XGB + PCA	0.992	1.00	0.98	0.99

The model’s classification performance is visualised through the confusion matrix, as shown in Fig.3. The program accurately identified 263 non-diabetic instances out of 263 actual non-diabetic patients, misclassifying none as diabetic. Similarly, only three cases were incorrectly classified as non-diabetic out of 137 actual diabetes instances, with 134 cases correctly identified.

These findings demonstrate the model’s remarkable ability to distinguish between individuals with and without diabetes. The model’s precision and recall are underscored by the low rate of misclassifications, especially in medical settings, where minimising false positives and negatives is crucial for effective diagnosis and treatment planning, as shown in Fig.4. Overall, the confusion matrix affirms the model’s reliability and robustness in real-world scenarios. The classification report heatmap depicts the model’s performance on the three primary evaluation criteria of precision, recall, and F1-score. The model’s exceptional precision of 0.99, recall of 1, and F1-score of 0.99 for class 0 (non-diabetic) show that it nearly always accurately classifies people

without diabetes. It demonstrated its efficacy in precisely identifying diabetes cases by maintaining a strong precision of 1, a recall of 0.98, and an F1-score of 0.99 for class 1 (diabetic). The model’s overall high accuracy of 0.9925 indicates that it performed consistently in both classes. The model’s well-rounded and trustworthy forecasts are further highlighted by the weighted and macro averages of 0.99 for all metrics. The model’s high sensitivity and specificity make it a good fit for real-world diabetes detection, as shown in the heatmap in Fig.4. The feature importance chart in Table.2 highlights the important health parameters that had the most influence in predicting diabetes. Glucose levels stood out as the most significant factor, with the highest importance score, underscoring its direct link to diabetes diagnosis.

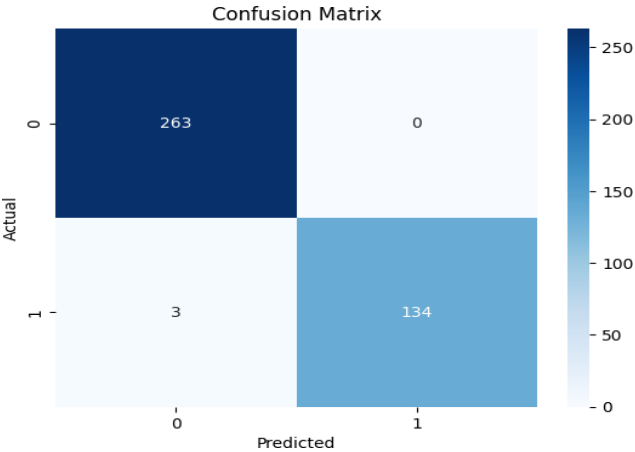


Fig.3. Confusion Matrix for Diabetes Classification

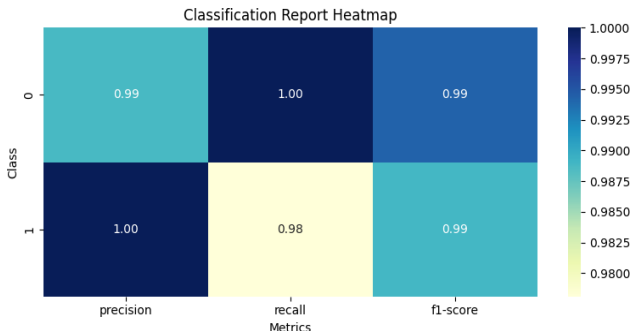


Fig.4. Classification Report Heatmap for Diabetes classification

Table.2. Correlation of Feature Selection to the Target

Age	Insulin	BP	Glucose	ST	BMI	DPF
0.14	0.075	0.08	0.28	0.06	0.17	0.125

Table.3. Entry of Features into the Web-Based Application

Age	Insulin	BP	Glucose	ST	BMI	DPF
56	250	138	168	38	38	0.345

5.1 PREDICTION RESULT

The person is likely to have diabetes. Following this, BMI (Body Mass Index) and Age also played major roles in determining diabetes risk, reflecting how body weight and age are crucial indicators. Other features like Diabetes Pedigree Function

(DPF), Blood Pressure (BP), and Insulin levels contributed moderately, suggesting their relevance in supporting the prediction. Factors such as Pregnancies, Skin Thickness (ST), and Cluster had relatively lower importance but still added value to the model's overall accuracy. This analysis provides meaningful insights into how different health metrics influence diabetes prediction, helping clinicians and researchers prioritise key indicators for early diagnosis.

To forecast the likelihood of diabetes, a set of health parameters was input into the proposed hybrid machine learning model, which had been trained. According to Table.3, the person was 56 years old, had two pregnancies, a blood pressure of 138, a glucose level of 168, a skin thickness of 38, an insulin level of 250, a BMI of 38, and a Diabetes Pedigree Function of 0.345. The proposed algorithm estimated that the individual is likely to have diabetes based on these statistics, particularly the elevated BMI and glucose levels. This prediction demonstrates the practical application of the ML model in real-time risk assessment, providing a quick and data-driven approach to early diagnosis. Such predictive tools can be valuable for both patients and healthcare providers in initiating timely medical advice and interventions.

6. CONCLUSION

Improving patient outcomes and minimising serious health problems largely depend on early diabetes diagnosis. Therefore, in this study, Principal Component Analysis and Gradient Boosting were combined with Random Forest to create a hybrid model that balances accuracy, speed, precision, and robustness. The analysis of feature importance indicates that Glucose Level, BMI, Age, Blood Pressure, and Insulin Levels are key factors in predicting diabetes. Random Forest provided strong, stable initial predictions due to its ensemble nature and resistance to overfitting, while Gradient Boosting refined these predictions by correcting errors and capturing subtle patterns in the data. This sequential improvement resulted in a final accuracy of 99.25% in our hybrid model, surpassing both individual and hybrid models. The ability of Gradient Boosting to enhance prediction quality made it a vital component in achieving these results. The proposed model outperformed logistic regression, decision trees, support vector machines, random forest, extreme gradient boost, hybridised random forest, and extreme gradient boost models in identifying individuals at risk for diabetes. This study highlights the increasing role of AI in healthcare, demonstrating that machine learning can offer fast, accurate, and cost-effective solutions for early diabetes detection. By combining the advantages of various models, we move closer to a more proactive healthcare system, where early intervention can significantly lessen the impact of diabetes and improve overall patient well-being. Further optimisation of hyperparameters suggests that fine-tuning hyperparameters or incorporating additional health indicators could enhance accuracy. Future research can focus on integrating real-time monitoring with wearable devices, exploring personalised risk assessments, and improving scalability for use in larger, more diverse populations. To continuously track user data, including heart rate, blood sugar levels, and physical activity, the model can be paired with wearable medical technology.

REFERENCES

- [1] "IDF Diabetes Atlas 2025 Global Diabetes Data and Insights", Available at: <https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/>, Accessed in 2025.
- [2] M.A. Rahman, L.F. Abdulrazak, M.M. Ali, I. Mahmud, K. Ahmed and F.M. Bui, "Machine Learning-based Approach for Predicting Diabetes Employing Socio-Demographic Characteristics", *Algorithms*, Vol. 16, No. 11, pp. 1-10, 2023.
- [3] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh and G. Stiglic, "Early Detection of Type 2 Diabetes Mellitus using Machine Learning-based Prediction Models", *Scientific Reports*, Vol. 10, No. 1, pp. 1-11, 2020.
- [4] O. Iparraguirre-Villanueva, K. Espinola-Linares, R.O. Flores Castaneda and M. Cabanillas-Carbonell, "Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes", *Diagnostics*, Vol. 13, No. 14, pp. 1-7, 2023.
- [5] O.N. Ergün and H.O. Ilhan, "Early Stage Diabetes Prediction using Machine Learning Methods", *European Journal of Science and Technology*, Vol. 8, pp. 1-8, 2021.
- [6] O.P. Suresh, B.M. Chandra, C.S. Reddy and B.V. Prasanth, "Economic Analysis of Grid-Tied Hybrid Solar Wind System with Electric Vehicle CS using COA-QNN Approach", *Journal of Energy Storage*, Vol. 100, pp. 1-15, 2024.
- [7] P. Durga and T. Sudhakar, "An Analysis of Various Machine Learning Techniques for Predicting Diabetes in its Early Stages", *Journal of Pharmaceutical Negative Results*, Vol. 13, No. 1, pp. 1-9, 2022.
- [8] C.K. Tiwari, K. Mittal and K. Dinesh, "An Insight into 'Decision Tree Analysis'", *World Wide Journal of Multidisciplinary Research and Development*, Vol. 3, No. 12, pp. 111-115, 2017.
- [9] P. Dutta, S. Paul and A. Kumar, "Comparative Analysis of Various Supervised Machine Learning Techniques for Diagnosis of COVID-19", *Electronic Devices, Circuits and Systems for Biomedical Applications: Challenges and Intelligent Approach*, Vol. 8, pp. 521-540, 2021.
- [10] C. Cortes, V. Vapnik and L. Saitta, "Support-Vector Networks Editor", 1995.
- [11] B.M. Chandra, P. Thenmozhi, S. Padma, Y. Shailaja and E.M. Kumaran, "LTECP: Experimental Analysis of Medicinal Plant Recognition System based on Learning Assisted Texture Evaluation and Classification Principle", *Proceedings of International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems*, Vol. 8, pp. 1-6, 2024.
- [12] B. Qu, "Bayes' Theorem in Machine Learning: A Literature Review", *Theoretical and Natural Science*, Vol. 86, No. 1, pp. 26-31, 2025.
- [13] N. Biswas, K.M.M. Uddin, S.T. Rikta and S.K. Dey, "A Comparative Analysis of Machine Learning Classifiers for Stroke Prediction: A Predictive Analytics Approach", *Healthcare Analytics*, Vol. 2, pp. 1-16, 2022.
- [14] P.B.K. Chowdary and R.U. Kumar, "An Effective Approach for Detecting Diabetes using Deep Learning Techniques based on Convolutional LSTM Networks", *International*

- Journal of Advanced Computer Science and Applications*, Vol. 12, No. 4, pp. 519-525, 2021.
- [15] M.F. Rabby, Y. Tu, M.I. Hossen, I. Lee, A.S. Maida and X. Hei, "Stacked LSTM based Deep Recurrent Neural Network with Kalman Smoothing for Blood Glucose Prediction", *BMC Medical Informatics and Decision Making*, Vol. 21, No. 1, pp. 1-14, 2021.
- [16] K. Abnoosian, R. Farnoosh and M.H. Behzadi, "Prediction of Diabetes Disease using an Ensemble of Machine Learning Multi-Classier Models", *BMC Bioinformatics*, Vol. 24, No. 1, pp. 1-19, 2023.
- [17] R.F. Albadri, S.M. Awad, A.S. Hameed, T.H. Mandeel and R.A. Jabbar, "A Diabetes Prediction Model using Hybrid Machine Learning Algorithm", *Mathematical Modelling of Engineering Problems*, Vol. 11, No. 8, pp. 2119-2126, 2024.
- [18] C.N. Noviyanti and A. Alamsyah, "Early Detection of Diabetes using Random Forest Algorithm", *Journal of Information System Exploration and Research*, Vol. 2, No. 1, pp. 1-7, 2024.
- [19] C.S. Reddy, B.V. Prasanth, O.P. Suresh and B.M. Chandra, "A Hybrid YDSE–THDCNN Approach based Multi-Objective Optimization of Energy Management for Renewable Energy Sources with Electric Vehicles", *Journal of Energy Storage*, Vol. 100, pp. 1-7, 2024.
- [20] Q. Saihood and E. Sonuç, "A Practical Framework for Early Detection of Diabetes using Ensemble Machine Learning Models", *Turkish Journal of Electrical Engineering and Computer Sciences*, Vol. 31, No. 4, pp. 722-738, 2023.
- [21] O.H. Abdulganiyu, T. Ait Tchakoucht, A.E.H. Alaoui and Y.K. Saheed, "Attention-Driven Multi-Model Architecture for Unbalanced Network Traffic Intrusion Detection via Extreme Gradient Boosting", *Intelligent Systems with Applications*, Vol. 26, pp. 1-11, 2022.
- [22] V.K. Singh, J. Barman and M.H. Kolekar, "Extreme Gradient Boosting Algorithm for Automatic Cardiac Arrhythmia Classification", *Procedia Computer Science*, Vol. 17, pp. 1466-1475, 2025.
- [23] N.S.K. Kalyan Chakravarthy, M.C. Balapanur, A.N. Pandian, J. Pulikanti, D. Prasad and J.A.I.S. Masood, "Optimizing Medical Imaging: High-Performance Hardware for Image Processing and Machine Learning", *Proceedings of International Conference on Digital Medicine and Image Processing*, Vol. 8, pp. 112-116, 2024.
- [24] T. Srikanth, A.S. Kannan and B. Moulichandra, "Power Quality Improvement in Distribution Generation with Renewable Energy Sources using Hybrid GBDT-BCMO Approach", *Australian Journal of Electrical and Electronics Engineering*, Vol. 22, No. 1, pp. 45-65, 2025.