

Early Detection of Diabetes Using Machine Learning Algorithms and Performance Analysis

Muneer Ahemed^{1,*}, Medikonda Swapna², Medikonda Asha Kiran¹, Kavitha Guda³, S. Jayanth¹, Manyam Thaile¹, Medikonda Neelima⁴, Niteesha Sharma¹, and Ramesh Babu Pittala⁵

¹*School of Engineering, Anurag University, Hyderabad, India*

²*Nitte Meenakshi Institute of Technology, Bengaluru, India*

³*Geethanjali College of Engineering and Technology, Hyderabad, India*

⁴*GVP College of Engineering, Visakhapatnam, India*

⁵*C.R. Rao Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad, India*

*Corresponding author: muneernrpt@gmail.com

Abstract. Diabetes is a widespread metabolic disease and a major public health problem that affects millions of people worldwide. Early screening and risk prediction allow timely intervention for high-risk groups and significantly reduce the risk of severe complications. In recent years, machine learning (ML) has shown clear value in the analysis of large-scale medical data. This study reviews and compares mainstream models for diabetes prediction and proposes a new prediction framework. On the Pima Indians Diabetes Dataset, the proposed hybrid model reached an accuracy of 0.92, a precision of 0.91, a recall of 0.90, and an F1 score of 0.91, higher than the decision tree, support vector machine, random forest, and logistic regression models tested. Such a system may help medical staff with screening and clinical decision-making.

Keywords. Diabetes prediction; machine learning; healthcare analytics; disease diagnosis; classification algorithms; predictive modeling

I. INTRODUCTION

Diabetes is a chronic metabolic disorder characterized by persistently elevated blood glucose levels caused by inadequate insulin production, impaired insulin action, or both [1], [2]. Driven by obesity, unhealthy lifestyles, and insufficient physical activity, the prevalence of this chronic, high-burden disease has continued to rise in recent years, and there is an urgent clinical need for early diagnosis and management. Its complications, which include cardiovascular disease and renal failure, impose a heavy burden on both patients and healthcare systems [3]-[5].

The volume of digital medical data has grown rapidly in recent years. Machine learning (ML) can uncover hidden patterns in medical data to support clinical diagnosis, monitoring, and decision-making, and many real-world applications have already been deployed [6]-[8]. Previous researchers have used logistic regression, decision trees, support vector machines (SVMs), random forests, and neural networks to build accurate diabetes prediction models [9]-[12]. The widely used Pima Indians Diabetes Dataset covers core clinical attributes such as the number of pregnancies, body mass index (BMI), glucose concentration, blood pressure, insulin level, and age, and it supports model development and validation; artificial intelligence (AI) and related technologies have further improved the accuracy of medical prediction in recent years [2],

[13]. This study compares multiple ML algorithms for diabetes prediction and proposes a prediction framework intended to improve classification accuracy and support the early diagnosis of diabetes [14].

II. LITERATURE REVIEW

Many researchers have explored the potential of ML in diabetes-related medical applications and have confirmed that it can extract useful insights from medical data. Rajendra and Latifi [9] used logistic regression for prediction, and Barakat and colleagues [10] used an SVM to improve classification performance. Earlier ML research in medical AI progressed toward increasingly complex models: decision trees support interpretable medical decision-making [6], the random forest model developed by Kaur and colleagues [11] achieved higher accuracy and stability, and Zhang and Qie [12] reviewed deep learning models that can capture complex patterns in medical data.

Other studies [4], [7], [15] indicate that the importance of AI, data science, and predictive analytics in medicine continues to grow. Similarly, [5], [8], [16] show that intelligent decision support systems can improve disease diagnosis and treatment planning.

Research on ML applications in healthcare can be divided into three groups by focus. Chen and colleagues [14]

applied data mining and ML methods to medical prediction and classification tasks; Swapna and colleagues [2] showed that combining these methods with medical datasets is effective for early diabetes diagnosis; and [1], [3], [17]-[19] together confirm the potential of ML for diabetes prediction and medical analysis. Clinical modeling still faces three core challenges: improving prediction accuracy, selecting the best algorithm, and developing reliable models suited to real clinical settings. Table 1 summarizes existing methods.

TABLE 1. COMPARISON OF EXISTING METHODS

Ref.	Method	Advantage	Accuracy
[9]	Logistic regression	Simple and easy to implement	76%
[10]	SVM	Better classification performance	82%
[6]	Decision tree	Easy to interpret	79%
[11]	Random forest	High stability and accuracy	85%
[12]	Neural network	Handles complex data patterns	Improved
[4]	Hybrid ML models	Better decision-making	Improved
[7]	Medical AI systems	Improves diagnostic accuracy	High
[16]	AI in medicine	Assists diagnosis	Reliable
[5]	Clinical models	Supports prognosis	Accurate
[8]	Decision support system	Helps medical decisions	Improved
[14]	ML prediction	Evaluates performance	High
[20]	ML classification	Compares algorithms	Good
[2]	Diabetes ML models	Early detection of diabetes	High

III. PROPOSED SYSTEM

This study aimed to develop a diabetes risk prediction system using ML techniques. Impaired insulin production or utilization can cause persistent hyperglycemia, which may lead to severe complications if it is not detected and managed early. Timely diabetes risk assessment and early screening can support preventive interventions, improve health outcomes, and reduce the disease burden; this is the clinical motivation for the study. The architecture of the proposed system is shown in Fig. 1: data collection, feature extraction and selection, a

split into training and testing data, and the application of a light gradient boosting machine (LGBM) algorithm to build the prediction model.

This study used the Pima Indians Diabetes Dataset, a widely recognized benchmark for diabetes prediction. As described in [13], the subjects of this dataset are Pima Indian women, and it includes eight core health features, such as the number of pregnancies. Each subject has a label indicating whether she has diabetes. The dataset allows ML models to learn patterns of diabetes onset and to predict the probability of diabetes in new patients.

A. Data Preprocessing

Data preprocessing is a necessary prerequisite for an effective diabetes prediction model. Flaws in raw medical datasets reduce model performance, so all preprocessing operations must be completed before model training begins. First, mean and median imputation were used to handle missing and invalid values in the dataset. Next, feature normalization was applied to put the attributes on a common scale and so improve the efficiency and convergence of the algorithms. After relevant attributes had been selected through feature selection, the dataset was split into training and test subsets, which were used for model development and for verifying the predictive performance and generalization ability of the model, respectively.

B. Algorithms Used

This study used several ML algorithms to predict diabetes risk and compared their performance to identify the most effective model for accurate disease prediction.

Decision tree (DT): A learning algorithm, trained on labeled data, that classifies by building a hierarchical tree structure from feature-based decision rules.

Random forest (RF): An ensemble learning algorithm that combines the predictions of multiple decision trees to improve classification accuracy, increase model stability, and reduce overfitting.

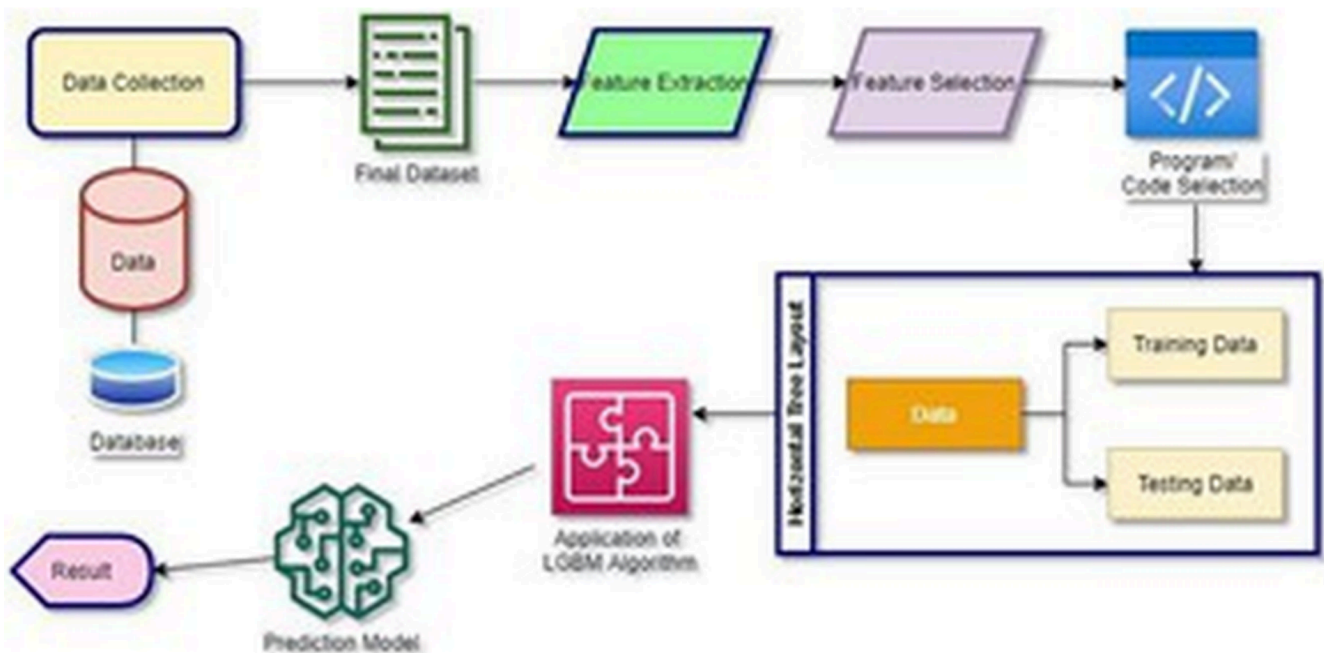


Fig. 1. System architecture of the proposed model.

SVM: A classification algorithm, trained on labeled data, that finds an optimal hyperplane to separate the classes. Its ability to handle high-dimensional feature spaces and complex data distributions makes it widely used for predictive modeling and classification.

Logistic regression (LR): A statistical classification technique, used in this study to estimate the probability of the binary disease outcome in diabetes prediction.

Proposed hybrid model: The proposed model integrates multiple ML algorithms within the workflow shown in Fig. 1.

C. Performance Comparison

Table 2 compares the proposed model with the existing ML algorithms to verify its usefulness in supporting early disease diagnosis. Accuracy, precision, recall, and the average of these three scores were used to assess and compare predictive effectiveness.

Table 3 lists the evaluation metrics of the proposed hybrid model, which scored higher than the other algorithms tested on every metric in Table 2. Fig. 2 compares the accuracy of the algorithms.

TABLE 2. PERFORMANCE COMPARISON OF MACHINE LEARNING ALGORITHMS

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	Average (%)
Decision tree	82	81	80	81
SVM	85	84	83	84
Random forest	88	87	86	87
Logistic regression	84	83	82	83
Proposed model	92	91	90	91

TABLE 3. PERFORMANCE EVALUATION METRICS OF THE PROPOSED MODEL

Metric	Value
Accuracy	0.92
Precision	0.91
Recall	0.90
F1 score	0.91

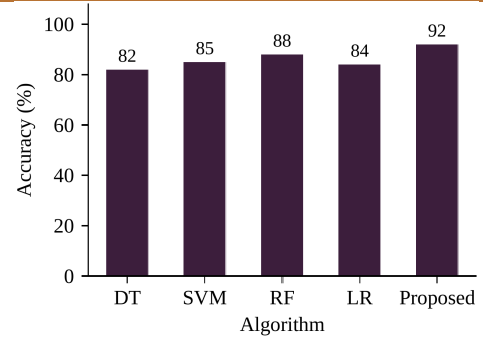


Fig. 2. Accuracy comparison of the machine learning algorithms.

IV. RESULTS AND DISCUSSION

The results indicate that ML models trained on this dataset can predict diabetes risk. The comparative performance of all tested algorithms is presented using the four evaluation metrics in Table 2.

The receiver operating characteristic (ROC) curve of the proposed hybrid model in Fig. 3 gives an area under the curve (AUC) of approximately 0.96.

All models were compared on a common evaluation benchmark. The decision tree model reached an accuracy of 82%, and its interpretability suits medical needs, but it is prone to overfitting on limited datasets. The logistic regression model reached an accuracy of 84%; its outputs are stable, but it struggles to model complex nonlinear relationships.

The SVM outperformed the decision tree and logistic regression with a classification accuracy of 85%; it can identify the optimal decision boundary but depends on parameter tuning and the choice of kernel function. Random forest was the best standalone algorithm, with an accuracy of 88%, owing to the performance gains of its ensemble of multiple decision trees.

The proposed hybrid ML model achieved 92% accuracy, 91% precision, 90% recall, and a 91% F1 score (Table 3), 4 percentage points above the accuracy of random forest,

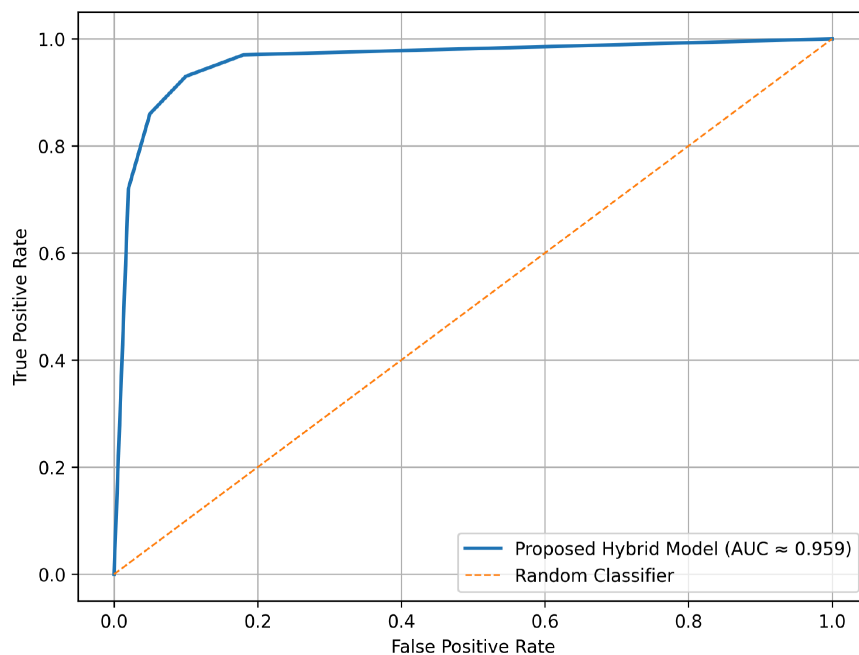


Fig. 3. ROC curve of the proposed diabetes prediction model.

the best single model (Table 4). Its stability was not evaluated separately.

TABLE 4. ACCURACY COMPARISON OF MACHINE LEARNING ALGORITHMS

Algorithm	Accuracy (%)
Decision tree (DT)	82
Support vector machine (SVM)	85
Random forest (RF)	88
Logistic regression (LR)	84
Proposed model	92

V. CONCLUSION

This study developed an ML framework for diabetes risk prediction. Using the publicly available Pima Indians Diabetes Dataset, the study carried out four preprocessing procedures and tested four types of algorithms. Random forest outperformed all other single classifiers, and the proposed hybrid model achieved an accuracy of 92% (Table 3). Its clinical value remains to be established on independent data.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and preparation of this work and approved the final version of the manuscript.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

DECLARATION OF GENERATIVE AI USE

The authors declare that no generative artificial intelligence tools were used in the preparation of this manuscript. The authors take full responsibility for the content of the publication.

DATA AVAILABILITY

This study used the publicly available Pima Indians Diabetes Dataset (Kaggle).

ETHICS STATEMENT

This study used publicly available, de-identified data and did not involve direct interaction with human participants or animals; ethical approval was therefore not required.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- [1] KM Jyoti Rani, "Diabetes prediction using machine learning," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 6, no. 4, pp. 294-305, 2020, doi: 10.32628/CSEIT206463.
- [2] G. Swapna, R. Vinayakumar, and K. P. Soman, "Diabetes detection using deep learning algorithms," *ICT Express*, vol. 4, no. 4, pp. 243-246, 2018, doi: 10.1016/j.icte.2018.10.005.
- [3] Satyanarayana Nimmala, Maragoni Mahendar, Pinnapureddy Manasa, H. N. Lakshmi, Medikonda Asha Kiran, and C. Raghavendra, "Fuzzy-enhanced XGBoost model for classifying kidney disease severity," in *Proc. 2024 4th Int. Conf. Soft Comput. Secur. Appl. (ICSCSA)*, 2024, pp. 21-26, doi: 10.1109/ICSCSA64454.2024.00011.
- [4] Adurupalli Sony, Manyam Thaile, Medikonda Asha Kiran, Ramesh Babu Pittala, Busa Nithin Kumar, Gandham Sai Chandu, and Muntha Raju, "Predictive health monitoring system for patient risk assessment using machine learning," in *Proc. 2025 5th Int. Conf. Intell. Technol. (CONIT)*, 2025, pp. 1-7, doi: 10.1109/CONIT65521.2025.11167397.
- [5] Ewout W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, 2nd ed., Statistics for Biology and Health. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-16399-0.
- [6] Vili Podgorelec, Peter Kokol, Bruno Stiglic, and Ivan Rozman, "Decision trees: An overview and their use in medicine," *J. Med. Syst.*, vol. 26, no. 5, pp. 445-463, 2002, doi: 10.1023/A:1016409317640.
- [7] Tegote Roshan, Vaddarapu Tarun Sai, Medikonda Asha Kiran, Sowjanya Yerramaneni, Lakshmi Prasanna Byrapuneni, and A. B. Pavani, "Real time health monitoring system using IoT," in *Proc. 2025 5th Int. Conf. Intell. Technol. (CONIT)*, 2025, pp. 1-6, doi: 10.1109/CONIT65521.2025.11167103.
- [8] Ibrahim Adedeji Adeniran, Christianah Pelumi Efunniyi, Olajide Soji Osundare, and Angela Omozele Abbulimen, "Data-driven decision-making in healthcare: Improving patient outcomes through predictive modeling," *Int. J. Sch. Res. Multidiscip. Stud.*, vol. 5, no. 1, pp. 59-67, 2024, doi: 10.56781/ijrms.2024.5.1.0040.
- [9] Priyanka Rajendra and Shahram Latifi, "Prediction of diabetes using logistic regression and ensemble techniques," *Comput. Methods Programs Biomed. Update*, vol. 1, Art. no. 100032, 2021, doi: 10.1016/j.cmpbup.2021.100032.
- [10] Nahla Barakat, Andrew P. Bradley, and Mohamed Nabil H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 4, pp. 1114-1120, 2010, doi: 10.1109/TITB.2009.2039485.
- [11] Pavleen Kaur, Ravinder Kumar, and Munish Kumar, "A healthcare monitoring system using random forest and internet of things (IoT)," *Multimedia Tools Appl.*, vol. 78, no. 14, pp. 19905-19916, 2019, doi: 10.1007/s11042-019-7327-8.
- [12] Huanhuan Zhang and Yufei Qie, "Applying deep learning to medical imaging: A review," *Appl. Sci.*, vol. 13, no. 18, Art. no. 10521, 2023, doi: 10.3390/app131810521.
- [13] Aishwarya Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Comput. Sci.*, vol. 165, pp. 292-299, 2019, doi: 10.1016/j.procs.2020.01.047.
- [14] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869-8879, 2017, doi: 10.1109/ACCESS.2017.2694446.
- [15] J. Manoranjini Kiran, Vijayalaxmi Bindla, G. V. R. Sai Lavanya, and Satyanarayana Nimmala, "A study article on concepts, applications and accomplishments of artificial intelligence: Net App AI," *Grenze Int. J. Eng. Technol.*, vol. 10, no. 2, 2024. [Online]. Available: <https://thegrenze.com/pages/servej.php?fn=303.pdf&id=3282&journal=GIJET>
- [16] Giovanni Briganti and Olivier Le Moine, "Artificial intelligence in medicine: Today and tomorrow," *Front. Med.*, vol. 7, Art. no. 27, 2020, doi: 10.3389/fmed.2020.00027.
- [17] Bomma Ramakrishna, Konduri V. N. A. S. M. Prasad, J. E. N. Abhilash, S. Suman, M. S. V. K. V. Prasad, and Medikonda Asha Kiran, "Enhanced emotion recognition via multi-sensor data fusion and machine learning," in *Proc. 2025 3rd World Conf. Commun. Comput. (WCONF)*, 2025, pp. 1-7, doi: 10.1109/WCONF64849.2025.11233351.
- [18] Ravi Raju Bandlamudi, Medikonda Asha Kiran, G. Sekhar Reddy, Katika Sushmitha, Nalla Madhuri, and Nadukuda Sanjana, "OsteoPredict: Osteoporosis risk prediction using machine learning," in *Proc. 2025 5th Int. Conf. Intell. Technol. (CONIT)*, 2025, pp. 1-6, doi: 10.1109/CONIT65521.2025.11167375.
- [19] Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang, "Predicting diabetes mellitus with machine learning techniques," *Front. Genet.*, vol. 9, Art. no. 515, 2018, doi: 10.3389/fgene.2018.00515.
- [20] Patlolla Akshaya, Katta Preethi, Medikonda Asha Kiran, Manyam Thaile, Ramesh Babu Pittala, and P. Nagamani, "Multi-sensor fusion for emotion recognition using machine learning," in *Proc. 2025 5th Int. Conf. Intell. Technol. (CONIT)*, 2025, pp. 1-6, doi: 10.1109/CONIT65521.2025.11167520.