



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Machine Learning for Diabetes Prediction

M. Krishna Kumar¹, P. Samayan², S. G. Abarna³, S.Varsha⁴, P.Meenadharshini⁵

Assistant Professor, Department of Mathematics, AAA College of Engineering and Technology, Sivakasi, Tamil Nadu, India¹

Assistant Professor, Department of Artificial Intelligence and Data Science, AAA College of Engineering and Technology, Sivakasi, Tamil Nadu, India^{2,5}

UG Student, Department of Artificial Intelligence and Data Science, AAA College of Engineering and Technology, Sivakasi, Tamil Nadu, India^{3,4}

ABSTRACT: Diabetes is one of the most common chronic health conditions affecting a large number of people worldwide, leading to serious health complications if detected and treated at a late stage. Conventional approaches for detecting diabetes include clinical tests and expert opinions, which are often time-consuming. Moreover, it is not always possible to ensure early detection. With recent advances in technology, machine learning has been found to be an effective method for predicting diabetes by processing large amounts of medical data and identifying underlying patterns and factors.

This paper aims to provide a comprehensive review of machine learning-based approaches for predicting diabetes, as reported in various studies between 2020 and 2025. Supervised learning approaches, including Logistic Regression, Support Vector Machines, Decision Tree, Random Forest, and various ensemble-based approaches, have been used to improve the accuracy of prediction. The approaches are based on various parameters, including glucose levels, body mass index, age, blood pressure, and family history of diabetes. The recent developments include the integration of deep learning and hybrid approaches for improving prediction accuracy.

However, challenges such as data imbalance, lack of quality datasets, overfitting, and model interpretability persist. The study demonstrates the prospect of machine learning in facilitating early diagnosis and preventive healthcare for diabetes patients. Future work in this area should be directed at creating precise, interpretable, and scalable machine learning models for practical applications in healthcare.

KEYWORDS: Machine Learning, Diabetes Prediction, Supervised Learning, Classification Algorithms, Logistic Regression, Random Forest, Support Vector Machine, Health Data Analysis, Predictive Modeling, Early Diagnosis, Artificial Intelligence, Healthcare Analytics

I. INTRODUCTION

Diabetes, a chronic metabolic disorder, causes high blood glucose levels, which can result in serious complications if not identified and controlled at an early stage. According to global health statistics, the number of people suffering from diabetes is rising day by day, making it a major concern for the global health care system. Early prediction and diagnosis of diabetes can help in controlling the complications caused by the disease.

Traditionally, the diagnosis of diabetes can be done by performing tests and manually evaluating the results by experts. However, these methods can sometimes be time-consuming and expensive. In the wake of the rapid increase in the amount of data available in the health care sector, there is a need for the implementation of intelligent systems.

Machine Learning (ML), a part of Artificial Intelligence, has gained much attention in recent times for its capabilities to analyze large amounts of data and identify intricate patterns that cannot be easily identified by human beings. In the context of predicting diabetes, ML algorithms have the capability to analyze patient data, including age, body mass index (BMI), glucose levels, insulin levels, blood pressure, family history, etc., to predict the chances of developing



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

diabetes. The most popular ML algorithms used to predict diabetes are Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, Ensemble Methods, etc.

The integration of ML in the healthcare domain has created new avenues for early diagnosis, preventive medicine, and efficient patient treatments. ML has provided an opportunity to healthcare providers to identify patients with higher chances of developing diabetes and take the necessary preventive steps to avoid the disease from progressing further. Moreover, ML has the capabilities to improve the efficiency of decision-making tasks for healthcare professionals.

However, ML-based diabetes prediction systems are faced with various challenges, including data quality, class imbalance, overfitting, lack of interpretability, etc. These need to be addressed to build efficient prediction models. This paper aims to give a detailed review of the application of the various machine learning techniques in the prediction systems, including the methodology, application, advantages, and disadvantages, as well as the future direction.

II. REVIEW OF EXISTING PAPER

1. Advances in Artificial Intelligence for Diabetes Prediction — Khokhar et al. (2025)

Khokhar et al. (2025) presented a systematic review of artificial intelligence techniques for diabetes prediction. The study highlights different machine learning models used with medical datasets for early diabetes detection. The authors point out that AI-based models can greatly improve prediction accuracy compared to traditional methods. However, they also discuss challenges like data quality, model interpretability, and the absence of standardized datasets, which limit real-world applications.

2. Machine Learning and Artificial Intelligence in Type 2 Diabetes Prediction — Kiran et al. (2025)

Kiran et al. (2025) conducted a thorough bibliometric and literature analysis on machine learning techniques for predicting Type 2 diabetes. The study shows the increasing use of algorithms like ensemble learning and deep learning. The authors conclude that although these models boost prediction performance, problems such as overfitting and limited generalization still pose significant challenges.

3. Supervised Machine Learning for Diabetes Prediction — Ansari et al. (2025)

Ansari et al. (2025) proposed supervised machine learning methods for predicting diabetes using feature selection techniques. The study shows that choosing relevant features improves model accuracy and lowers computational complexity. Algorithms like Random Forest and Support Vector Machine did well in prediction tasks. However, the study points out the need for stronger models to manage imbalanced datasets.

4. Diabetes Prediction Based on Machine Learning — Sui (2024)

Sui (2024) examined various machine learning techniques for diabetes prediction, with a special emphasis on classification. The author compared various machine learning models and proved that Logistic Regression and Decision Tree are effective for predicting diabetes. The author also stressed the need for data preprocessing and feature engineering.

5. Prediction of Diabetes Mellitus Progression — Chauhan et al. (2023)

In a study by Chauhan et al. (2023), supervised machine learning models were used for predicting the progression of diabetes mellitus. The study shows the need for early prediction in order to avoid complications. From the results, it is evident that machine learning models can attain a high degree of accuracy. The accuracy is, however, dependent on the quality of data used.

6. Secure Machine Learning for Diabetes Prediction — Hennebelle et al. (2022)

The research paper by Hennebelle et al. (2022) specifically dealt with privacy-preserving machine learning approaches for diabetes prediction. In this paper, the authors introduced secure models for diabetes prediction, which ensure the confidentiality of the patient's data while maintaining the accuracy of the predictions. The paper also discussed the significance of data security in healthcare applications, but it also added complexity to the computations.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

7. Comparative Study of Classifiers for Diabetes Prediction — Hasan & Yasmin (2025)

Hasan and Yasmin (2025) have done a comparative analysis of various machine learning classifier models in predicting diabetes. The models analyzed in the study include Naive Bayes, K-Nearest Neighbors, and Random Forest. The results indicate that ensemble models perform better in terms of accuracy compared to other models. However, the interpretability of the models is a problem.

8. Machine Learning Techniques for Early Diabetes Prediction — Alzboon et al. (2025)

Alzboon et al. (2025) compared various machine learning techniques in the early prediction of diabetes. The paper is focused on improving the accuracy of the prediction using advanced machine learning techniques and feature selection. The results of the paper show that hybrid models perform better than traditional models but demand more computational power.

9. AI and Machine Learning in Prediabetes — Lalani et al. (2025)

In a study conducted by Lalani et al. (2025), the application of AI and ML in predicting prediabetes-related conditions was explored. The study emphasizes the significance of early detection of the condition to prevent the advancement of diabetes. The researchers assert that AI systems can be used to support early intervention, but a large amount of data is required.

10. Machine Learning for Gestational Diabetes Prediction — Zhao et al. (2025)

Zhao et al. (2025) conducted a systematic review with a meta-analysis of various studies concerning the application of machine learning models for predicting gestational diabetes mellitus. The article indicates that the models are effective in predicting patients with a higher risk of developing the disease. However, the authors indicate some limitations, including a lack of external validation and inconsistency in the data used.

III. COMPARISON WITH EXISTING PAPER

The comparison of the existing research works on the implementation of the concept of machine learning for the prediction of diabetes reveals that a wide variety of machine learning algorithms are being utilized to improve the early diagnosis and accuracy of the predictions. The most commonly used traditional machine learning algorithms, such as Logistic Regression, Decision Trees, and Support Vector Machines, are being widely accepted and implemented for the early diagnosis of diabetes, as these models are easy to implement and can achieve high accuracy when appropriate preprocessing and feature selection techniques are utilized. In addition, research works such as those of Sui (2024) and Chauhan et al. (2023) emphasize the importance of selecting appropriate features such as glucose levels, BMI, and age.

On the contrary, recent research has emphasized advanced techniques such as ensemble learning, hybrid, and artificial intelligence-based techniques to achieve even higher accuracy in prediction outcomes. For instance, the research works published by Ansari et al. (2025) and Alzboon et al. (2025) have shown the effectiveness of hybrid techniques in achieving even higher accuracy in comparison to the use of a single classifier. However, some common problems have been found in the majority of the research works, such as the problem of data imbalance, overfitting, and the need for a good dataset. There is a need to develop more robust, efficient, and effective machine learning models, which can be practically implemented in the real world.

Author & Year	Method / Approach	Key Contribution	Advantages	Limitations
Khokhar et al. (2025)	AI-based diabetes prediction review	Comprehensive analysis of AI techniques	High prediction accuracy	Lack of standard datasets, low interpretability
Kiran et al. (2025)	Bibliometric + ML analysis	Study of ML trends in diabetes prediction	Identifies advanced techniques	Overfitting, limited generalization
Ansari et al. (2025)	Supervised ML with feature selection	Improved accuracy using selected features	Reduces complexity	Struggles with imbalanced data
Sui (2024)	Classification	Comparison of ML	Simple and	Requires proper



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

	algorithms	models	effective models	preprocessing
Chauhan et al. (2023)	Supervised learning models	Prediction of diabetes progression	High accuracy in controlled data	Data quality dependency
Hennebelle et al. (2022)	Privacy-preserving ML	Secure diabetes prediction	Data security ensured	High computational cost
Hasan & Yasmin (2025)	Comparative ML study	Evaluates multiple classifiers	Ensemble methods perform well	Lack of explainability
Alzboon et al. (2025)	Hybrid ML models	Improved prediction using hybrid approach	Better performance	High resource requirement
Lalani et al. (2025)	AI for prediabetes	Early detection of prediabetes	Preventive healthcare	Requires large datasets

IV. RESEARCH GAP

However, there are still some gaps in research on machine learning for diabetes prediction, despite significant advances in machine learning techniques. Firstly, many machine learning-based research on diabetes prediction is conducted on small-scale data. The performance of machine learning algorithms is dependent on the availability of data. The performance of machine learning algorithms on a specific data set is poor compared to their performance on different data sets.

Secondly, data imbalance is a major issue in machine learning-based research on diabetes prediction. The number of people who are non-diabetic is much higher compared to people who are diabetic. Lastly, there is a need for machine learning-based research on diabetes prediction to be interpretable. Most machine learning algorithms are not interpretable. For instance, ensemble learning and deep learning are complex machine learning algorithms. The performance of machine learning algorithms is poor compared to classical machine learning algorithms. Overfitting is a common problem with machine learning-based research on diabetes prediction.

Moreover, the studies tend to emphasize the accuracy of the models, but other key issues, such as implementation in real-time, scalability, and integration, may be overlooked. Additionally, there is a lack of emphasis on the privacy and security of the patient data, which is a key concern in a health context. Lastly, the models tend to lack real-world clinical validation, as they are usually validated on benchmark data and not on real-world data. Therefore, in the future, studies should emphasize developing robust, interpretable, and scalable machine learning models that can effectively handle a wide range of data and can be integrated into a real-world health context for early and accurate prediction of diabetes.

V. CONCLUSION

Machine learning has proved to be an effective and reliable method for the early prediction of diabetes, which has greatly aided in the diagnosis of the disease as well as preventive healthcare. The review of existing studies from 2020 to 2025 indicates that various machine learning algorithms, including Logistic Regression, Decision Trees, Support Vector Machines, Random Forest, and Ensemble, have been used to analyze patient data to predict the occurrence of the disease. The models are able to identify hidden patterns in patient data, which aids in the early prediction of the disease, thus preventing complications from the disease.

The recent advancements in artificial intelligence, including hybrid models, feature selection, and other techniques, have greatly aided in the accuracy of the models. However, some of the challenges facing the application of machine learning algorithms in the early prediction of diabetes include data imbalance, overfitting, lack of interpretability, and reliance on good quality data. Moreover, most of the models have not been validated in practical scenarios.

To address the challenges associated with the current machine learning models, it is important to develop effective machine learning models that can generalize well across different datasets. Additionally, there is a need to ensure data



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

security, transparency, and real-time implementation of machine learning models. Overall, machine learning has a huge potential to transform the field of diabetes prediction and early diagnosis.

REFERENCES

1. Khokhar, P. B., Gravino, C., Palomba, F., et al. "Advances in Artificial Intelligence for Diabetes Prediction: Insights from a Systematic Literature Review," 2025. DOI: <https://doi.org/10.1016/j.artmed.2025.103132>
2. Kiran, M., Xie, Y., Anjum, N., et al. "Machine Learning and Artificial Intelligence in Type 2 Diabetes Prediction: A Comprehensive 33-Year Bibliometric and Literature Analysis," 2025. DOI : <https://doi.org/10.3389/fdgth.2025.1557467>
3. Ansari, G. A., Shafi, S., Ansari, M. D., et al. "Advanced Supervised Machine Learning Methods for Precise Diabetes Mellitus Prediction Using Feature Selection," 2025. DOI: <https://doi.org/10.3389/fmed.2025.1620268>.
4. M. Jansi Rani, and M. Prabha, An Efficient Resource Allocation Mechanism Using Intelligent Scheduling for Managing Energy in Cloud Computing Infrastructure, Information and Communication Technology for Competitive Strategies (ICTCS 2021), Lecture Notes in Networks and Systems 401 (2023), 81-86.
5. Metaheuristic Feature Selection for Diabetes Prediction with P-G-S Approach, Karuppasamy M, Jansi Rani M, Poorani K, 4th International Conference on Evolutionary Computing and Mobile Sustainable Networks, Procedia Computer Science 252 (2025) 165–171.
6. M. Prabha, and M. Jansi Rani, Future Worth: Predicting Resale Values with Machine Learning Techniques, Inventive Communication and Computational Technologies, Lecture Notes in Networks and Systems 757 (2023), 1101-1112.
7. M. Jansi Rani, M. Karuppasamy, M. Prabha, and K. Pooran, Detection of COVID-19 CoronaVirus Using ResNet Deep Learning Technique, Signal Processing, Telecommunication and Embedded Systems with AI and ML Applications, Lecture Notes in Electrical Engineering 1281 (2025), 71-83.
8. Sui, Y. "Diabetes Prediction Based on Machine Learning," 2024. DOI: <https://doi.org/10.54254/2753-8818/2025.18035>
9. Chauhan, A. S., Varre, M. S., Izuora, K., et al. "Prediction of Diabetes Mellitus Progression Using Supervised Machine Learning," 2023. DOI: <https://doi.org/10.3390/s23104658>
10. Hennebelle, A., Ismail, L., Materwala, H., et al. "Secure and Privacy-Preserving Automated Machine Learning for Diabetes Prediction," 2022. DOI: <https://doi.org/10.48550/arXiv.2211.07643>
11. Hasan, M., & Yasmin, F. "Predicting Diabetes Using Machine Learning: A Comparative Study of Classifiers," 2025. DOI: <https://doi.org/10.48550/arXiv.2505.07036>
12. Alzboon, M. S., Al-Batah, M., Alqaraleh, M., et al. "A Comparative Study of Machine Learning Techniques for Early Prediction of Diabetes," 2025. DOI: <https://doi.org/10.48550/arXiv.2506.10180>
13. Lalani, B., Herur, R., Zade, D., et al. "Applications of Artificial Intelligence and Machine Learning in Prediabetes: A Scoping Review," 2025. DOI: <https://doi.org/10.1177/19322968251351995>
14. Zhao, M., Yao, Z., Zhang, Y., et al. "Predictive Value of Machine Learning for the Progression of Gestational Diabetes Mellitus to Type 2 Diabetes: A Systematic Review and Meta-Analysis," 2025. DOI: <https://doi.org/10.1186/s12911-024-02848-x>



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details