

A REVIEW OF MACHINE LEARNING TECHNIQUES FOR DIABETES RISK PREDICTION

Limaje Milind Subhash*¹, DR. Rameshwar Singh Sikarwar *²

^{*1}Student, Dr. A.P.J. Abdul Kalam University, Indore, M.P., India

^{*2}Assistant Professor, Dr. A.P.J. Abdul Kalam University, Indore, M.P., India

Received: 17/12/2024

Revised: 15/01/2025

Accepted: 23/02/2025

ABSTRACT

Diabetes mellitus is one of the most common chronic diseases affecting millions of individuals worldwide and has become a major challenge for healthcare systems. The increasing prevalence of diabetes has created a strong demand for early prediction and preventive healthcare solutions. Traditional diagnostic procedures mainly rely on laboratory investigations and clinical expertise, which may not always support early identification of high-risk individuals. In recent years, machine learning techniques have emerged as promising tools for predictive healthcare due to their capability to analyze medical datasets and identify hidden disease patterns. Machine learning algorithms can assist healthcare professionals in predicting diabetes risk using clinical and physiological parameters such as glucose level, insulin concentration, body mass index, age, blood pressure, and hereditary factors. Several classification techniques, including Logistic Regression, Decision Tree, K-Nearest Neighbors, Random Forest, and deep learning models, have been applied for diabetes prediction. However, challenges related to prediction accuracy, feature relevance, computational complexity, overfitting, and interpretability continue to affect model performance. This review paper presents a detailed survey of machine learning techniques used for diabetes risk prediction. Existing studies are critically analyzed based on methodologies, datasets, predictive performance, strengths, and limitations. A comparative evaluation of commonly used classification algorithms is also provided to identify research trends and current limitations. The findings indicate that although intelligent techniques significantly improve prediction capability, there is still a need for computationally efficient, interpretable, and feature-optimized models suitable for practical healthcare applications. The study further highlights important research gaps and future directions for developing improved diabetes prediction systems.

KEYWORDS: Diabetes Risk Prediction, Machine Learning, Healthcare Analytics, Artificial Intelligence, Disease Prediction, Classification Algorithms, Predictive Healthcare.

DOI: <https://doi.org/10.64149/gjaets.12.2.21-26>

1. INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder that occurs when the human body fails to regulate blood glucose levels properly due to inadequate insulin production or ineffective insulin utilization. Over the last few decades, diabetes has become a major public health issue affecting people across different age groups. Increasing urbanization, unhealthy dietary habits, obesity, stress, sedentary lifestyles, and genetic predisposition have contributed significantly to the rising number of diabetic patients worldwide. If diabetes is not identified and controlled at an early stage, it may lead to serious complications including cardiovascular diseases, kidney dysfunction, vision loss, nerve damage, and other severe health conditions [10], [12].

Traditional diabetes diagnosis generally depends on blood tests, physician observations, and clinical expertise. Although these methods provide reliable outcomes, they often identify the disease only after symptoms become prominent. Delayed diagnosis can reduce treatment effectiveness and increase healthcare burden. Therefore, the healthcare sector has increasingly focused on intelligent computational systems that can support early disease prediction and preventive healthcare [26].

Machine learning, a major branch of artificial intelligence, has demonstrated considerable success in healthcare analytics and predictive diagnosis. Machine learning techniques can learn from historical medical records, identify hidden relationships among healthcare variables, and generate predictive outcomes without explicit programming [20]. These capabilities make machine learning highly suitable for diabetes prediction where multiple medical parameters influence disease occurrence.



Several machine learning approaches have been explored for disease prediction, including Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine, Random Forest, and deep learning models [24], [32]. Among these approaches, supervised classification methods have gained significant attention because diabetes prediction generally involves labeled healthcare datasets.

Despite the growing success of machine learning in healthcare applications, several challenges remain unresolved. Existing systems often focus mainly on prediction accuracy while neglecting model interpretability, feature relevance, and computational efficiency. Highly complex models frequently produce better outcomes but may not be suitable for real-world healthcare environments due to implementation complexity and poor explainability [9]. Therefore, a systematic review of machine learning approaches for diabetes risk prediction becomes necessary to understand recent developments, identify limitations, and establish future research opportunities.

The objective of this review paper is to examine existing machine learning techniques applied in diabetes risk prediction, compare traditional and advanced classification approaches, identify major research gaps, and establish future directions for developing reliable and interpretable diabetes prediction systems.

2. LITERATURE REVIEW

Machine learning techniques have significantly transformed predictive healthcare systems by enabling efficient disease diagnosis, medical decision support, and patient monitoring. Researchers across the world have explored various machine learning methods to improve disease prediction accuracy and support preventive healthcare. Diabetes risk prediction has become an important research area because of the increasing prevalence of the disease and the necessity for early diagnosis.

Li and Zhang [1] proposed a machine learning framework for early diabetic retinopathy detection using retinal imaging data. Their study showed that intelligent algorithms can effectively detect diabetes-related complications at an early stage through healthcare image analysis. The authors highlighted the importance of feature extraction and classification performance in predictive healthcare systems.

Smith et al. [3] developed a predictive healthcare framework for forecasting patient readmission rates using machine learning techniques. Their research demonstrated that predictive analytics can improve healthcare management by identifying high-risk patients. The study identified data preprocessing and feature optimization as significant elements influencing predictive performance.

Kumar et al. [24] focused specifically on diabetes prediction using machine learning algorithms and demonstrated that supervised classification techniques can effectively predict diabetes outcomes. Their work highlighted the significance of preprocessing methods and healthcare feature analysis for improving prediction reliability.

Das et al. [32] performed a comparative analysis of machine learning algorithms in disease prediction systems. The study evaluated different classification models including Decision Tree, Logistic Regression, and KNN. Their findings suggested that prediction performance depends heavily on data quality, feature relevance, and model tuning.

Jain et al. [26] examined the role of artificial intelligence in healthcare and discussed how machine learning techniques assist healthcare professionals through predictive analysis and intelligent automation. However, the authors also identified issues related to transparency and ethical concerns in healthcare AI systems.

Lee et al. [12] developed a machine learning model for predicting cardiovascular risks in diabetic patients and concluded that combining multiple healthcare variables significantly improves predictive reliability.

Johnson et al. [10] investigated diabetic foot ulcer prediction using machine learning and demonstrated that predictive systems can help healthcare professionals identify high-risk patients before complications become severe.

Brown et al. [37] emphasized the significance of machine learning in clinical decision-support systems and highlighted how predictive healthcare systems reduce diagnosis time and improve treatment planning.

Deep learning methods have also been increasingly adopted in healthcare applications. Chen and Wang [7] proposed a deep learning-based approach for Alzheimer's disease detection using neuroimaging data and reported strong predictive performance. Similarly, Yousef et al. [8] applied deep learning for brain tumor segmentation and achieved improved classification accuracy in medical imaging.

Sharma *et al*. [21] integrated IoT and machine learning technologies to develop intelligent healthcare monitoring systems capable of real-time patient monitoring. Their study indicated that predictive healthcare can become more effective through continuous data acquisition and intelligent analytics.

The reviewed studies demonstrate that machine learning significantly improves healthcare prediction systems, particularly in diabetes diagnosis. However, several limitations such as overfitting, computational complexity, poor interpretability, and insufficient feature optimization continue to affect predictive reliability.

3. COMPARATIVE ANALYSIS OF MACHINE LEARNING TECHNIQUES

Machine learning algorithms have become important tools for diabetes risk prediction because of their ability to process healthcare data and identify relationships among medical variables. Different classification techniques have been explored by researchers to improve prediction accuracy and support healthcare professionals in early disease diagnosis. However, every machine learning algorithm has specific strengths and limitations, and their performance depends on dataset characteristics, preprocessing quality, feature selection, and model optimization.

Among traditional classification methods, Logistic Regression is one of the most commonly used algorithms for healthcare prediction. It is particularly suitable for binary classification problems where the target output contains two possible classes, such as diabetic and non-diabetic conditions. Logistic Regression estimates the probability of disease occurrence using a mathematical function and produces interpretable results. The simplicity and lower computational requirement of this method make it suitable for healthcare applications. However, the algorithm often struggles to handle highly complex or nonlinear relationships among features [30].

Decision Tree is another frequently used classification approach in predictive healthcare systems. The algorithm constructs a tree-like structure by splitting healthcare data according to feature values and decision rules. Decision Tree models are easy to interpret and visualize, which makes them highly suitable for clinical decision-support systems. Healthcare professionals can understand prediction logic more clearly compared to black-box models. Nevertheless, Decision Trees are often vulnerable to overfitting problems, especially when training data contains noise or high variability [32].

K-Nearest Neighbors (KNN) is a similarity-based classification method used in healthcare analytics. The algorithm classifies new patient data by comparing it with previously labeled samples available in the dataset. KNN is considered simple and effective for smaller datasets because it does not require complex training procedures. However, the prediction process becomes computationally expensive for large-scale healthcare datasets, and the algorithm is sensitive to irrelevant features and noisy data [20].

Random Forest has gained considerable attention in healthcare prediction because of its ensemble learning capability. Instead of relying on a single decision structure, Random Forest combines multiple decision trees and generates predictions based on majority voting. This mechanism reduces overfitting and improves prediction stability. Several studies have shown that Random Forest often performs better than individual classifiers in healthcare prediction tasks because of its capability to manage nonlinear relationships and complex datasets [35].

Deep learning models have also been increasingly applied in healthcare prediction systems. Neural networks and deep architectures automatically extract hidden relationships from healthcare records and often produce strong predictive performance. Despite their effectiveness, deep learning models generally require large training datasets, increased computational resources, and sophisticated optimization techniques. Furthermore, interpretability remains a major challenge because healthcare professionals may find it difficult to understand the reasoning behind predictions [31].

A comparative understanding of these algorithms is important because prediction accuracy alone cannot determine model suitability. Factors such as computational efficiency, interpretability, scalability, and implementation feasibility must also be considered when selecting predictive models for healthcare applications.

Algorithm	Advantages	Limitations	Suitability
Logistic Regression	Simple, interpretable, low computation	Weak nonlinear capability	Suitable for basic classification
Decision Tree	Easy visualization, interpretable	Prone to overfitting	Moderate suitability

K-Nearest Neighbors	Easy implementation	High computation cost	Effective for small datasets
Random Forest	High accuracy, reduced overfitting	Slightly increased complexity	Highly suitable
Deep Learning	Strong predictive performance	Complex and less interpretable	Suitable for large datasets

Table 1: Comparative Analysis of Machine Learning Algorithms

4. CONCLUSION AND FUTURE SCOPE

This review study examined various machine learning techniques used for diabetes risk prediction and analyzed their contribution to predictive healthcare systems. The findings from the literature indicate that machine learning has become an effective approach for assisting early disease diagnosis by analyzing patient medical data and identifying hidden relationships among healthcare attributes. Several classification techniques, including Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Random Forest, and deep learning models, have demonstrated promising outcomes in diabetes prediction and healthcare analytics [20], [24], [32].

The literature further reveals that traditional machine learning approaches remain highly useful because of their simplicity, lower computational cost, and better interpretability. Models such as Logistic Regression and Decision Tree allow healthcare professionals to understand prediction behavior more clearly, making them suitable for decision-support applications. On the other hand, advanced methods including Random Forest and deep learning frameworks generally provide improved predictive performance by handling nonlinear relationships and complex healthcare patterns more effectively [7], [31]. However, these methods often involve increased computational complexity, reduced interpretability, and higher implementation requirements.

From the review of previous studies, several important research gaps have been identified. Many existing diabetes prediction systems primarily focus on improving prediction accuracy while comparatively less attention is given to model simplicity, practical usability, and interpretability in healthcare environments. In many studies, all medical features are treated equally despite the fact that certain parameters such as glucose level, body mass index (BMI), insulin concentration, age, and hereditary influence contribute more significantly to diabetes prediction [12], [24]. Furthermore, some advanced predictive systems require extensive computational resources and large datasets, limiting their applicability in practical healthcare scenarios [31].

Another important limitation identified in the literature is poor generalization capability. Several prediction systems produce effective results only on specific datasets and may not perform consistently across different healthcare populations. In addition, many studies provide limited comparative analysis among traditional classification algorithms, creating challenges in selecting the most suitable model for diabetes prediction [30], [32].

Based on the findings of this review, future research should focus on developing balanced and intelligent classification models that achieve improved prediction performance while maintaining simplicity, interpretability, and computational efficiency. Instead of relying only on highly complex architectures, future diabetes prediction systems should emphasize feature optimization and meaningful attribute prioritization to improve predictive consistency. The integration of lightweight intelligent models with healthcare monitoring technologies may also improve real-time disease prediction and preventive healthcare services [21], [33].

In conclusion, machine learning continues to play an increasingly important role in predictive healthcare and diabetes diagnosis. Although existing studies have achieved encouraging results, there remains significant scope for developing improved and practically applicable diabetes prediction systems. Therefore, future research can focus on designing feature-optimized classification models capable of delivering accurate, interpretable, and computationally efficient predictions suitable for real-world healthcare environments.



REFERENCES

1. H. Li and Y. Zhang, "Machine Learning Framework for Early Detection of Diabetic Retinopathy Using Retinal Imaging Data," *IEEE Access*, vol. 9, pp. 73396–73406, 2021, doi: 10.1109/ACCESS.2021.307284.
2. S. Ghalib, A. Kasem, and A. Ali, "Analytical Study of Wireless Ad-Hoc Networks: Types, Characteristics, Differences, Applications, Protocols," in *Proceedings of the Second International Conference on Futuristic Trends in Networks and Computing Technologies*, Singapore: Springer, 2019, pp. 22–40.
3. A. Smith, M. Johnson, R. Williams, and D. Clark, "Predictive Healthcare Framework Utilizing Machine Learning Algorithms for Forecasting Patient Readmission Rates," *IEEE Access*, vol. 10, pp. 14288–14302, 2022, doi: 10.1109/ACCESS.2022.3288798.
4. X. Du, Y. Chen, H. Zhao, and L. Wang, "Convolutional Bidirectional Long Short-Term Memory Neural Network for Driver Emotion Recognition," *IEEE Access*, vol. 9, pp. 28579–28589, 2021.
5. W. Wang, J. Liu, and M. Chen, "Driver Emotion Recognition Using Electrocardiogram (ECG) Features," *IEEE Access*, vol. 10, pp. 29433–29442, 2022.
6. S. Sachdeva and A. Ali, "Machine Learning with Digital Forensics for Attack Class in Cloud Network Environment," *International Journal of System Assurance Engineering and Management*, vol. 13, pp. 156–165, 2022.
7. Y. Chen and Q. Wang, "Deep Learning-Based Approach for Early Detection of Alzheimer's Disease Using Multimodal Neuroimaging Data," *IEEE Access*, vol. 9, pp. 73212–73222, 2021.
8. M. Yousef, A. Hassan, and K. Mahmood, "Brain Tumor Segmentation Using Deep Learning With Bridged U-Net," *IEEE Access*, vol. 9, pp. 139565–139579, 2021.
9. J. Fernández-Alemán, P. Sánchez-Henarejos, and A. Toval, "Ethical Considerations in the Use of Machine Learning for Diabetic Healthcare," *IEEE Access*, vol. 10, pp. 37159–37171, 2022.
10. M. Johnson, R. Thomas, and S. Brown, "Predictive Modeling of Diabetic Foot Ulcers Using Machine Learning Techniques," *IEEE Access*, vol. 10, pp. 136759–136770, 2022.
11. Y. Wu, Z. Li, and H. Zhang, "Real-Time Monitoring of Asthma Exacerbations Using Machine Learning-Based System," *IEEE Access*, vol. 10, pp. 18592–18604, 2022.
12. H. Lee, J. Kim, and R. Wilson, "Predicting the Risk of Cardiovascular Events in Diabetic Patients Using Machine Learning Algorithms," *IEEE Access*, vol. 10, pp. 18615–18628, 2022.
13. P. Roberts, A. Walker, and M. Thompson, "Transformer Models for Natural Language Understanding in Diabetic Patient Interactions," *IEEE Access*, vol. 9, pp. 86495–86507, 2021.
14. R. Gupta, S. Sharma, and V. Mehta, "Personalized Medication Dosage Recommendation in Chronic Disease Management Using Machine Learning," *IEEE Access*, vol. 10, pp. 51279–51293, 2022.
15. Y. Zhang, T. Liu, and P. Chen, "Predictive Analytics Framework for Sepsis Detection in ICU Patients Using ML," *IEEE Access*, vol. 10, pp. 88263–88278, 2022.
16. M. Ahmad, S. Khan, and A. Hussain, "Emotion Recognition of Driver Based on KLT Algorithm and ShuffleNetV2," *Signal, Image and Video Processing*, vol. 18, pp. 3643–3660, 2024.
17. A. Garg, P. Sharma, and R. Kumar, "Shadow Preservation Framework for Content-Aware Image Retargeting," *Journal of Autonomous Intelligence*, vol. 6, no. 3, pp. 1–20, 2023.
18. H. I. Hamid, A. Rahman, and M. Khan, "Writer Identification Using CNN," *Journal of Autonomous Intelligence*, vol. 7, no. 1, pp. 1–14, 2024.
19. M. Rahmani, S. Gupta, and N. Singh, "Secure Routing Framework Using Hybrid Optimization in IoT WSN," *Journal of Nanoelectronics and Optoelectronics*, vol. 18, pp. 338–346, 2023.
20. K. Patel, R. Sharma, and P. Mehta, "A Survey on Machine Learning Techniques in Healthcare," *International Journal of Computer Applications*, vol. 177, no. 5, pp. 10–15, 2019.
21. R. Sharma, A. Verma, and S. Kumar, "IoT-Based Smart Healthcare Monitoring System Using ML," *International Journal of Engineering Research and Technology*, vol. 9, no. 6, pp. 115–122, 2020.
22. S. Verma, N. Gupta, and R. Kumar, "Cloud-Based Healthcare System Using Deep Learning," *International Journal of Advanced Research in Computer Science*, vol. 12, no. 2, pp. 45–52, 2021.
23. P. Singh, V. Sharma, and A. Yadav, "A Review on Deep Learning in Medical Imaging," *International Journal of Scientific Research in Computer Science*, vol. 8, no. 3, pp. 25–31, 2020.



24. A. Kumar, R. Singh, and D. Mishra, "Diabetes Prediction Using Machine Learning Algorithms," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 6, pp. 215–221, 2019.
25. R. Mehta, S. Gupta, and V. Sharma, "Heart Disease Prediction Using ML Techniques," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 4, pp. 110–118, 2020.
26. S. Jain, N. Agarwal, and P. Verma, "Artificial Intelligence in Healthcare: A Review," *International Journal of Recent Technology and Engineering*, vol. 8, no. 6, pp. 2250–2257, 2020.
27. N. Gupta, R. Sharma, and A. Tiwari, "Predictive Analysis of Cancer Using Machine Learning," *International Journal of Innovative Technology and Exploring Engineering*, vol. 10, no. 5, pp. 89–96, 2021.
28. M. Khan, S. Ali, and A. Hassan, "Data Mining Techniques for Healthcare Applications," *International Journal of Data Mining*, vol. 6, no. 2, pp. 45–53, 2019.
29. P. Roy, S. Kumar, and A. Tiwari, "AI-Based Disease Prediction System," *International Journal of Computer Science Trends and Technology*, vol. 9, no. 4, pp. 78–84, 2021.
30. T. Ali, M. Hassan, and A. Raza, "Machine Learning Approaches for Medical Diagnosis," *International Journal of Computer Engineering*, vol. 5, no. 2, pp. 60–68, 2020.
31. V. Reddy, M. Rao, and P. Kumar, "Deep Neural Networks for Healthcare Applications," *International Journal of Advanced Computer Science*, vol. 11, no. 1, pp. 15–24, 2021.
32. S. Das, P. Sharma, and R. Kumar, "A Comparative Study of ML Algorithms in Disease Prediction," *International Journal of Computer Applications*, vol. 176, no. 18, pp. 12–19, 2020.
33. A. Nair, V. Menon, and R. Nambiar, "IoT and Machine Learning Integration for Smart Health," *International Journal of Engineering Science*, vol. 14, no. 3, pp. 95–104, 2021.
34. [34] D. Paul, M. Verma, and H. Sharma, "Healthcare Big Data Analytics Using Machine Learning," *International Journal of Big Data*, vol. 4, no. 2, pp. 50–61, 2020.
35. K. Verma, A. Sharma, and N. Singh, "Disease Prediction Using Hybrid Machine Learning Models," *International Journal of Emerging Technologies*, vol. 12, no. 4, pp. 101–109, 2021.
36. J. Thomas, D. Roberts, and W. Scott, "Application of AI in Medical Diagnosis Systems," *International Journal of Computer Science and Mobile Computing*, vol. 9, no. 7, pp. 35–42, 2020.
37. L. Brown, M. Green, and K. Adams, "Machine Learning for Clinical Decision Support Systems," *International Journal of Medical Informatics*, vol. 129, pp. 320–330, 2019.
38. S. Kulkarni, R. Joshi, and S. Patil, "Early Disease Detection Using Artificial Intelligence," *International Journal of Computer Science Research*, vol. 9, no. 3, pp. 85–93, 2021.
39. A. Bose, P. Das, and V. Roy, "Healthcare Monitoring Using Wearable Devices and ML," *International Journal of Smart Systems*, vol. 5, no. 2, pp. 44–52, 2020.
40. R. Iyer, S. Menon, and A. Kapoor, "AI-Based Medical Image Classification Techniques," *International Journal of Image Processing*, vol. 14, no. 1, pp. 11–20, 2019.
41. S. Kapoor, M. Gupta, and A. Sethi, "Machine Learning in Telemedicine Applications," *International Journal of Telemedicine*, vol. 6, no. 2, pp. 70–79, 2021.
42. M. Arora, N. Bansal, and P. Singh, "Predictive Analytics in Healthcare Using Data Science," *International Journal of Data Science*, vol. 3, no. 1, pp. 22–31, 2020.

CITE AN ARTICLE

Subhash L. M., Sikarwar. R. S., (2025). A REVIEW OF MACHINE LEARNING TECHNIQUES FOR DIABETES RISK PREDICTION. *Global Journal of Advanced Engineering and Technology Studies*, 13(2), 21-26.
<https://doi.org/10.64149/gjaets.12.2.21-26>

