An Analytical Study of Machine Learning Models for Early Diabetes Risk Prediction

OPEN ACCESS

Volume: 13

Special Issue: 3

Month: February

Year: 2026

P-ISSN: 2321-788X

E-ISSN: 2582-0397

Citation: Rajarajeswari, V. "An Analytical Study of Machine Learning Models for Early Diabetes Risk Prediction." *Shanlax International Journal of Arts, Science and Humanities,* vol. 13, no. 3, 2026, pp. 335–39.

DOI: https://doiorg/10.34293/sijash.v13iS3-i2-Feb.10301

Mrs. V. Rajarajeswari
Assistant Professor, Department of Computer Science
Thiruthangal Nadar College, Chennai.

Abstract

Diabetes mellitus is a chronic metabolic disorder that has become a major global health concern due to its increasing prevalence and severe long-term complications. Early prediction and diagnosis of diabetes play a crucial role in preventing disease progression and improving patient outcomes. In recent years, machine learning (ML) techniques have been widely applied to diabetes prediction using clinical and demographic data. This study presents a comprehensive analysis of existing machine learning models used for diabetic prediction. Commonly employed algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbours, Naïve Bayes, and Artificial Neural Networks are examined based on their prediction accuracy, sensitivity, specificity, interpretability, and computational complexity. Publicly available datasets, particularly the Pima Indians Diabetes Dataset, are frequently used for model evaluation. The analysis highlights that ensemble models like Random Forest and Gradient Boosting generally outperform traditional classifiers in terms of accuracy, while simpler models such as Logistic Regression offer better interpretability for clinical decision-making. However, challenges such as data imbalance, overfitting, lack of explainability, and limited real-world clinical validation remain significant. This analytical study provides insights into the strengths and limitations of existing machine learning approaches and identifies research gaps to guide the development of more robust, interpretable, and clinically applicable diabetes prediction systems.

Keywords: Diabetes Prediction, Machine Learning, Classification Algorithms, Ensemble Learning, Healthcare Analytics

Introduction

Diabetes mellitus has emerged as a significant global health challenge due to its high prevalence as a non-communicable disease affecting millions of individuals and placing a significant burden on healthcare systems. The disease is characterized by chronic hyperglycaemia resulting from defects in insulin secretion, insulin action, or both. If not detected and managed at an early stage, diabetes can lead to severe complications such as heart disease, kidney failure, nerve damage and vision loss.

Traditional diagnostic methods rely on laboratory tests and clinical assessments, which may be time-consuming and often detect the disease only after significant progression. With the rapid growth of healthcare data and advancements in computational intelligence, machine learning techniques have emerged as powerful tools for

early diabetes risk prediction. ML models can analyse large volumes of clinical and demographic data to identify hidden patterns and support timely clinical decision-making.

This paper provides an analytical study of widely used machine learning models for early diabetes risk prediction, comparing their performance, advantages, and limitations. The objective is to help researchers and practitioners understand current trends and identify future research directions.

Related Work

Several studies have explored the application of machine learning techniques in diabetes prediction. Logistic Regression has been commonly used due to its simplicity and interpretability. Decision Tree and Naïve Bayes classifiers have also been applied for rule-based and probabilistic prediction.

More recent studies demonstrate that ensemble methods such as Random Forest and Gradient Boosting significantly improve prediction accuracy by combining multiple weak learners. Support Vector Machines and Artificial Neural Networks have shown strong performance in handling nonlinear relationships within the data. However, many existing works rely heavily on benchmark datasets and lack real-world clinical validation.

Machine Learning Models for Diabetes Prediction

This section discusses commonly used ML algorithms for diabetes risk prediction.

Logistic Regression (LR)

Logistic Regression is a statistical model widely used for binary classification problems. It offers high interpretability and helps clinicians understand the contribution of each feature to the prediction outcome.

Decision Tree (DT)

Decision Trees use a hierarchical structure to make decisions based on feature values. They are easy to interpret but prone to overfitting.

Random Forest (RF)

Random Forest is an ensemble-based learning technique that aggregates the predictions of multiple decision trees. It improves prediction accuracy and reduces overfitting.

Support Vector Machine (SVM)

SVM constructs an optimal hyperplane to separate classes and is effective in high-dimensional spaces. However, it requires careful parameter tuning.

K-Nearest Neighbours (KNN)

KNN classifies instances based on the majority class of nearest neighbours. Its performance depends on the choice of distance metric and value of K.

Naïve Bayes (NB)

Naïve Bayes is a probabilistic classification algorithm grounded in Bayes' theorem.. It is computationally efficient but assumes feature independence.

Artificial Neural Networks (ANN)

ANNs are inspired by the human brain and can model complex nonlinear relationships. Despite high accuracy, they lack interpretability.

Dataset Description

Most diabetes prediction studies use publicly available datasets, particularly the Pima Indians Diabetes Dataset from the UCI Machine Learning Repository. The dataset contains attributes such as glucose level, blood pressure, BMI, insulin level, age, and diabetes outcome.

Challenges associated with this dataset include class imbalance, missing values, and limited population diversity.

Performance Evaluation Metrics

The performance of ML models is commonly evaluated using the following metrics:

- Accuracy
- Sensitivity (Recall)
- Specificity
- Precision
- F1-score
- Area Under the ROC Curve (AUC)

These metrics provide a comprehensive understanding of model effectiveness in clinical contexts.

Comparative Analysis

Ensemble models such as Random Forest and Gradient Boosting generally achieve higher accuracy compared to individual classifiers. However, simpler models like Logistic Regression and Decision Trees offer better transparency and interpretability, which are essential for clinical adoption.

A trade-off exists between model accuracy and explainability, highlighting the need for explainable AI techniques in healthcare.

Results and Sample Performance Comparison

Table 1 presents a sample comparative analysis of commonly used machine learning models for diabetes prediction based on reported results in existing literature using the Pima Indians Diabetes Dataset. The values are representative and provided for analytical and illustrative purposes.

Table 1 Sample Performance Comparison of ML Models for Diabetes Prediction

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	AUC
Logistic Regression	78.5	76.0	80.2	75.4	0.83
Decision Tree	74.2	72.5	75.8	71.6	0.76
Naïve Bayes	76.8	74.9	78.1	73.8	0.81
K-Nearest Neighbours	77.4	75.6	79.0	74.2	0.82
Support Vector Machine	80.1	78.8	81.3	77.9	0.85
Artificial Neural Network	81.6	80.4	82.7	79.8	0.87
Random Forest	83.9	82.5	85.1	81.7	0.90
Gradient Boosting	85.2	84.0	86.3	83.1	0.92

The results indicate that ensemble-based models such as Random Forest and Gradient Boosting consistently outperform single classifiers in terms of overall accuracy and AUC. However, interpretable models like Logistic Regression still demonstrate competitive performance and are often preferred in clinical environments.

Table Discussion (Clinical Relevance): From a clinical perspective, higher sensitivity is particularly important in diabetes risk prediction, as it ensures that a greater number of at-risk patients are correctly identified at an early stage. As shown in Table 1, ensemble models achieve superior sensitivity and AUC values, making them suitable for large-scale screening and population-level risk assessment. However, their complex internal structures limit transparency, which can reduce clinician trust and hinder adoption in real-world healthcare settings. In contrast, simpler models such as Logistic Regression and Decision Trees, while slightly less accurate, offer better interpretability and enable clinicians to understand how individual risk factors contribute to predictions. Therefore, the results suggest that a hybrid or explainable ensemble approach may provide an optimal balance between predictive performance and clinical usability.

Challenges and Research Gaps

Despite promising results, several challenges remain:

- Data imbalance and quality issues
- Overfitting in complex models
- Lack of explainability
- Limited real-world clinical validation
- Ethical and privacy concerns

Addressing these issues is crucial for developing reliable and clinically applicable diabetes prediction systems.

Explainable Artificial Intelligence (XAI) for Diabetes Prediction

While machine learning models demonstrate promising performance in diabetes risk prediction, their adoption in real-world clinical environments is often limited by the lack of transparency and explainability. Explainable Artificial Intelligence (XAI) aims to address this challenge by providing human-understandable explanations for model predictions, thereby improving trust and interpretability.

Model-agnostic explanation techniques such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) have gained significant attention in healthcare applications. LIME explains individual predictions by approximating complex models locally with interpretable surrogate models, allowing clinicians to understand why a specific patient is classified as high or low risk. SHAP, based on cooperative game theory, assigns contribution scores to each feature, offering both global and local interpretability of model behavior.

In diabetes prediction, XAI techniques help identify clinically relevant risk factors such as glucose level, BMI, age, and insulin concentration, and quantify their influence on prediction outcomes. Integrating SHAP or LIME with high-performing ensemble models like Random Forest or Gradient Boosting can bridge the gap between predictive accuracy and clinical interpretability. Therefore, incorporating XAI methods is essential for developing trustworthy, transparent, and clinically acceptable diabetes prediction systems.

Future Directions

- Future research should focus on:
- Explainable machine learning models

- Hybrid and ensemble approaches
- Use of real-world clinical datasets
- Integration with clinical decision support systems
- Federated learning for privacy-preserving healthcare analytics

Conclusion

This paper presented an analytical study of machine learning models for early diabetes risk prediction. The comparative analysis shows that while ensemble and deep learning models offer high accuracy, simpler models provide better interpretability. A balanced approach combining accuracy, transparency, and clinical validation is essential for effective diabetes prediction systems.

References

1. World Health Organization, Diabetes Fact Sheet.
2. UCI Machine Learning Repository, Pima Indians Diabetes Dataset.
3. Breiman, L., "Random Forests," Machine Learning Journal.
4. Bishop, C.M., Pattern Recognition and Machine Learning.
5. Esteva, A. et al., "A Guide to Deep Learning in Healthcare," Nature Medicine.
6. G. Revathy, D. Ravikumar, K. Lakshmi and M. M. Santhosh Kumar, FEDMAP: Federated Learning for Dynamic Vehicle Traffic Mapping,2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2025, pp. 1440-1444, doi: 10.1109/ICSCDS65426.2025.11167623.
7. Ravikumar, D., V. Devi, C. Sharanya, P. Vijayalakshmi, and P. Radhakrishnan. Deploying an Innovative Routing Algorithm to Enhance Data Security via Internet Protocol Security Measures. Science and Technology-Recent Updates and Future Prospects Vol. 3 (2024): 167-180.
8. Suresh, G., D. Ravikumar, P. Prakasam, and V. Thirunavukkarasu. "Energy Reinforcement Model to Detect Selfish Node and Cluster Management in MANETs. In 2023 International Conference on Next Generation Electronics (NEleX), pp. 1-5. IEEE, 2023.