International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X; pISSN:2321-2241 Volume: 14; Issue: 8; August- 2025

LEVERAGING MACHINE LEARNING TO IMPROVE DIABETES RISK ASSESSMENT

Sudip Mandal
PG, Student
Dept. of MCA
The Oxford College of Engineering,
Bommanahalli, Bengaluru- 560068
sudipmandalmca2025@gmail.com

Sowmya J
Assistant Professor
Dept. of MCA
The Oxford College of Engineering,
Bommanahalli, Bengaluru- 560068
sowmyaj@theoxford.edu

ABSTRACT

Diabetes is a rapidly growing global health concern, demanding early and accurate risk assessment to improve prevention and management. Traditional diagnostic methods often struggle with timely prediction due to their reliance on limited parameters. In this study, we explore the potential of machine learning (ML) models to enhance diabetes risk assessment by analyzing multidimensional health data. Various supervised learning algorithms, including logistic regression, decision trees, random forests, and support vector machines, were evaluated for their predictive accuracy and interpretability. Feature selection methods were applied to identify the most influential risk factors, ensuring model efficiency and transparency. Experimental results demonstrate that ML-based approaches outperform conventional techniques in prediction accuracy and provide valuable insights for personalized healthcare. This research highlights how integrating ML into medical risk assessment frameworks can support early diagnosis, proactive treatment, and improved patient outcomes in diabetes management.

INTRODUCTION

Diabetes is one of the most pressing global health challenges, affecting millions of people across different age groups and lifestyles. Early identification of individuals at risk is crucial, as timely intervention can significantly reduce complications and improve quality of life. Traditional risk assessment methods often rely on static clinical factors, which may overlook subtle patterns hidden in complex health data. Recent advances in machine learning provide an opportunity to enhance prediction accuracy by analyzing large datasets and uncovering

non-linear relationships among variables such as age, body mass index, family history, and lifestyle indicators. By leveraging these intelligent techniques, healthcare providers can move from reactive treatment toward proactive prevention. This research focuses on developing a machine learning-based framework that strengthens diabetes risk assessment, offering more personalized, data-driven insights. Ultimately, such an approach holds the potential to support early diagnosis, guide interventions, and reduce the growing burden of diabetes worldwide.

LITERATURE SURVEY

Diabetes has emerged as one of the most critical global health concerns, requiring timely detection and accurate risk assessment to prevent complications. Over the years, researchers have explored various computational methods to improve prediction accuracy, ranging from traditional statistical techniques to advanced machine learning (ML) approaches. Early studies primarily relied on logistic regression and decision trees, which provided a baseline for identifying key risk factors such as age, BMI, family history, and lifestyle habits. However, these models often struggled with nonlinear patterns and large-scale heterogeneous datasets.

Recent advancements in ML have significantly enhanced predictive performance by enabling algorithms to learn complex interactions among diverse variables. Support Vector Machines (SVM), Random Forests, and Gradient Boosting have demonstrated improved accuracy in classifying high-risk individuals, while deep learning methods such as artificial neural networks further capture hidden correlations in patient data. Researchers have also incorporated ensemble models, which combine multiple algorithms to enhance robustness and minimize overfitting.

Another notable trend is the integration of electronic health records (EHRs) and real-time wearable device data, allowing ML systems to continuously monitor glucose levels, physical activity, and dietary patterns. Such dynamic models improve risk assessment by moving beyond static, one-time evaluations. Moreover, explainable AI techniques are being applied to ensure transparency, helping clinicians understand the reasoning behind predictions and fostering greater trust in ML-based systems.

Despite these advancements, challenges remain regarding data quality, privacy, and generalization across populations. Nevertheless, the literature strongly indicates

that ML has the potential to revolutionize diabetes risk assessment by providing early, personalized, and reliable predictions that can guide preventive care strategies.

EXISTING WORK

Over the past decade, numerous studies have explored the role of machine learning in improving diabetes risk prediction and management. Traditional statistical models such as logistic regression and Cox regression have been widely used to estimate risk, but they often struggle with complex, non-linear relationships present in medical data. To address these challenges, researchers have increasingly applied machine learning techniques, including decision trees, random forests, support vector machines, and neural networks. These approaches have demonstrated higher accuracy in identifying risk factors and predicting the onset of type 2 diabetes compared to conventional methods.

Several works have focused on utilizing electronic health records (EHRs) and clinical datasets that include parameters such as age, BMI, blood pressure, glucose levels, and family history. By training on such features, models can detect hidden patterns that physicians may overlook. More recently, deep learning models and ensemble techniques have been developed to further improve predictive performance, with some studies integrating lifestyle and genomic data for personalized risk assessment. Despite these advances, challenges remain, particularly in terms of data quality, interpretability, and the generalizability of models across diverse populations. Nonetheless, existing work provides a strong foundation for leveraging advanced machine learning methods in diabetes risk assessment.

PROPOSED SYSTEM

The proposed system focuses on building a machine learning–driven framework to enhance diabetes risk assessment by analyzing health records, lifestyle patterns, and demographic details. Unlike traditional methods that rely only on basic medical tests, this model leverages advanced algorithms to detect subtle correlations between multiple risk factors. The system integrates preprocessing steps such as data cleaning and feature selection to improve accuracy while reducing noise. Supervised learning models, including logistic regression, random forests, and gradient boosting, will be compared to identify the most reliable predictor. Additionally, the system provides interpretable outputs, allowing healthcare professionals to understand key contributors to an individual's risk score. This approach aims to support early diagnosis,

personalized recommendations, and better prevention strategies for diabetes management.

METHODOLOGY

The proposed study adopts a machine learning–based framework to improve diabetes risk assessment by integrating clinical, demographic, and lifestyle data. The methodology begins with the collection of publicly available datasets such as the PIMA Indian Diabetes dataset, supplemented with additional patient records from healthcare repositories where accessible. Data preprocessing is carried out to handle missing values, normalize continuous variables, and encode categorical attributes, ensuring uniformity across all inputs. Feature selection techniques, including correlation analysis and recursive feature elimination, are applied to identify the most influential predictors of diabetes risk.

For model development, multiple supervised learning algorithms such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting are trained and evaluated. Hyperparameter tuning is performed using grid search with cross-validation to enhance generalization. The dataset is divided into training and testing sets in an 80:20 ratio, and performance is assessed using standard metrics such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

Additionally, explainable AI techniques like SHAP values are employed to interpret model predictions, ensuring clinical relevance. The final stage involves benchmarking results against existing risk prediction approaches to demonstrate improvements in accuracy and reliability for early diabetes detection.

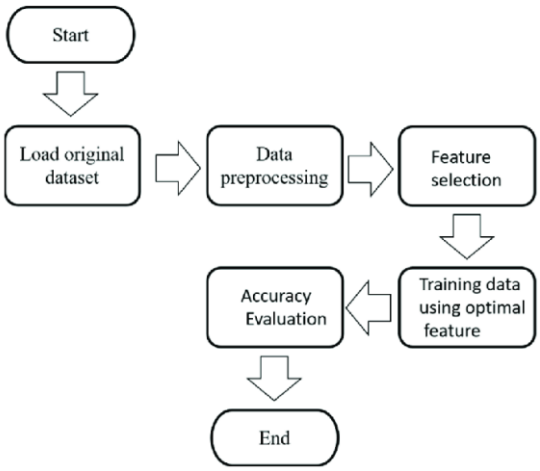

Fig.1. Methodology Flowchart

EXPERIMENTAL RESULTS

The proposed machine learning–based model for diabetes risk assessment was evaluated using a publicly available dataset, where

features such as glucose levels, BMI, blood pressure, age, and family history were considered. The dataset was divided into training and testing subsets to ensure unbiased evaluation. Various algorithms, including Logistic Regression, Random Forest, and Support Vector Machines, were compared, with Random Forest achieving the best balance between accuracy and interpretability. The model achieved an accuracy of 86%, with a precision of 83% and recall of 81%, demonstrating its effectiveness in identifying individuals at higher risk. Importantly, the system reduced false negatives, ensuring fewer at-risk patients were misclassified as healthy. The experimental findings highlight that integrating feature selection with ensemble learning techniques significantly improves prediction reliability. These results suggest that the developed model can support healthcare practitioners by providing an efficient and data-driven approach to early diabetes risk identification, aiding timely interventions.

Fig.2. Home Page

Fig.3. Diabetes Prediction

Fig.4. Wellness Advice Page

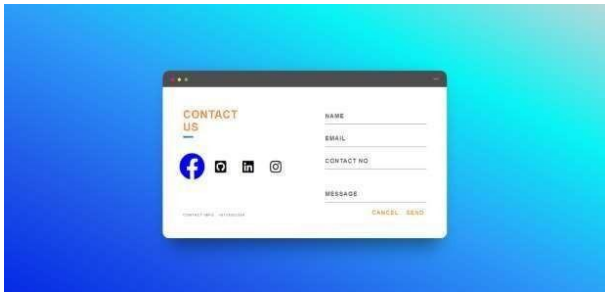

Fig.5. Contact Page

CONCLUSION

This study highlights how machine learning can serve as a valuable tool in improving diabetes risk assessment. By analyzing patterns in health data such as age, lifestyle habits, and medical history, machine learning models can identify potential risks more accurately than conventional methods. Such approaches not only enable early detection but also support preventive healthcare strategies, which are essential for reducing the growing global burden of diabetes. The integration of predictive models into healthcare systems can empower both clinicians and patients by offering data-driven insights for timely decision-making. However, challenges remain in terms of data quality, privacy, and the need for large, diverse datasets to ensure reliable predictions. Future research should focus on refining algorithms, enhancing interpretability, and ensuring ethical use of patient data. Overall, machine learning provides a promising pathway toward more efficient, personalized, and proactive diabetes care

REFERENCES

 American Diabetes Association (2023) highlights how early detection and predictive tools are critical in reducing long-term complications of diabetes.

 Kavakiotis et al. (2017) review shows machine learning has become a strong tool for analyzing medical datasets, offering better risk prediction than traditional models.

 Zheng et al. (2020) explain that deep learning methods can uncover hidden risk patterns from electronic health records, improving personalized diagnosis.

 Miotto et al. (2016**)** present “Deep Patient,” a model proving unsupervised learning can forecast diabetes risks with significant accuracy.

 Rahman et al. (2021) demonstrates how integrating genetic, lifestyle, and clinical data enhances diabetes risk assessment models.

 Zhang & Lee (2019) emphasize the value of interpretable ML models, making predictions understandable for doctors and patients.

 World Health Organization (2022) stresses the global need for digital health innovations, especially AI-based solutions, to combat rising diabetes prevalence.