Open Access International Journal of Science and Engineering

ISSN: 2456-3293

Volume 9 Issue 1s (2026) | Pages: 107-112

Interpretable Machine Learning Approach for Real-Time Diabetes Risk Prediction

Rohini G. Khalkar[1], Monali Rajendra Shinde[2], Priyanshu Raj[3], Atharva Tamhankar[4], Avdhoot Shinde[5], Mandeep Singh[6]

[1,3,4,5,6]Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, India
[2]Department of Information Technology, Matoshri College of Engineering and Research Centre, Nashik, India
[1]rgkhalkar@bvucoep.edu.in, [2]mj16111986@gmail.com, [3]rajpriyanshu1711@gmail.com, [4]atharva.tamhankar21@gmail.com, [5]avdhoot678@gmail.com, [6]mandeeprajput808@gmail.com

Peer Review Information

Type: *Article*
Received: *29 March 2026*
Revised: *26 April 2026*
Accepted: *30 May 2026*
Published: *19 June 2026*

Abstract

Diabetes mellitus is one of the most prevalent chronic metabolic diseases and represents a major global public-health challenge in the 21st century. It is characterized by persistently elevated blood glucose levels caused by insufficient insulin production, insulin resistance, or both mechanisms. Type 2 Diabetes Mellitus (T2DM) accounts for most cases and is strongly associated with lifestyle factors such as obesity, sedentary behaviour, unhealthy dietary habits, and genetic predisposition [1]. Long-term uncontrolled diabetes can lead to severe complications including cardiovascular diseases, neuropathy, retinopathy, kidney failure, and reduced life expectancy [2].

Traditional diagnostic methods such as fasting plasma glucose tests, oral glucose tolerance tests, and glycated haemoglobin (HbA1c) assessments are clinically reliable but often require laboratory infrastructure and repeated testing [3]. With the rapid growth of healthcare datasets and electronic health records, machine learning techniques have emerged as effective tools for early disease prediction and healthcare analytics [7], [8].

This paper proposes a machine learning-based diabetes prediction system using a Decision Tree classifier. The model analyses several clinical parameters including glucose level, body mass index (BMI), lipid profile indicators, kidney function markers, and age-related attributes to predict diabetes risk. Data preprocessing techniques such as normalization, missing value handling, and class imbalance correction are applied to improve model performance. The trained model is integrated into a Flask-based web application capable of classifying individuals into non-diabetic, pre- diabetic, and diabetic categories. The system demonstrates the potential of interpretable machine learning models for early risk detection and decision support in modern healthcare systems.

Keywords: Diabetes Prediction; Machine Learning; Decision Tree; Healthcare Analytics; Medical Data Mining.

How to Cite This Article

Khalkar, R. G., Shinde, M. R., Raj, P., Tamhankar, A., Shinde, A., & Singh, M. (2026). Interpretable Machine Learning Approach for Real-Time Diabetes Risk Prediction. *Open Access International Journal of Science and Engineering*, 9(1s), 107–112.

Introduction

Diabetes mellitus is one of the most widespread and fastest- spreading non-communicable diseases in the global populations affecting the general health and health care sector significantly [1]. This is a metabolic disorder where hyperglycemia is continuous because of defects in the insulin production or sensitivity or a combination of the two. Type 2 diabetes is experienced in a huge majority of the incidences of diabetes worldwide. There are three fundamental types of diabetes, which are Type 1 diabetes, Type 2 diabetes, and gestational diabetes. Close association has been made between the increased incidence of diabetes and urbanization, sedentary lifestyles, poor eating habits, increasing levels of obesity and genetic susceptibility [5], [6].

There are severe multifaceted complications of diabetes in the long term. Some of the issues that may be caused by the cumulative damage of vital organs by chronic hyperglycemia include peripheral neuropathy, renal failure, coronary artery disease, stroke, visual impairment, and delayed wound healing [2]. Such issues cause morbidity and mortality rates to rise greatly and reduce the general quality of life of the victims. Diabetes causes significant financial effects on individuals and health care systems due to the high cost of long-term care, hospitalization, and the management of complications besides the ill effects on health.

As such, there is need to have early detection and effective management to minimize the complications and healthcare expenses of diabetes. Nevertheless, the basis of conventional methods of diagnosis is laboratory-based clinical tests, including HbA1c tests, oral glucose tolerance tests, and fasting plasma glucose [3]. These procedures involve frequent testing, medical accessibility as well as professional medical personnel even though they are accurate and widely acceptable. Preventive diagnosis and screening are however, yet to come and particularly in the rural and resource-limited regions.

The emergence of machine learning and data mining methods in healthcare analytics has become possible due to increasing access to digitized medical data and increasing computing power. Machine learning models can analyze big data on medical data and identify latent trends as well as predict the risk of diseases with high accuracy [7]. With the assistance of early risk identification and personalized healthcare plans, these models may serve as decision-support medical tools. Decision Tree classifiers are very widely used in healthcare applications among other machine learning algorithms because they are easy to interpret, transparent, and are rule- based in nature [8], [12]. This is unlike the black-box models that offer obscure decision paths that assist clinicians in determining how particular clinical parameters affect the outcome of predictions. The proposed use of a Decision Tree- based diabetes prediction system in this study will be to assign patients to the non-diabetic, pre-diabetic, and diabetic groups. It is very crucial to include the pre-diabetic stage since this would allow the intervention of the disease at an early stage before it advances to a critical point. The system is deployed as a web-based application based on Flask framework, which makes it more accessible and feasible in real life.

Literature Survey

The last ten years have been associated with a lot of research involving the use of data mining and machine learning methods to predict diabetes and diagnose the disease. The initial one relied primarily on statistical models of linear and logistic regression to investigate the relationship between the levels of glucose and the occurrence of diabetes. Though the said models were informative in terms of their information on risk factors of diabetes, their predictive power was not as high as the ability of the models to represent non-linear relationships in real health-care data [9].

As computers began growing more capable, there was a renewed interest in considering other methods of machine learning like Naive Bayes and K-nearest neighbors (KNN) that were created to perform simpler functions. Naive Bayes was a popular form of Bayesian classifier as it is easy to implement and it does not require a lot of computation, but they also do not capture the relationship between variables, which can be inaccurate in data with correlated variables. The K-Nearest Neighbors algorithm was quite successful on small size datasets but when using large amounts of data, it experienced some complexity in its input formats, slowness during large information processing, and failure when working with large databases in the health sector.

SVM have gained immense popularity in prediction of diabetes as they have a solid theoretical basis and can work on high-dimensional data. In several studies, SVM was found to be more accurate in comparison to conventional classifiers [10]. Along with these benefits, the SVM models are black- box systems and, thus, it is hard to understand the logic of prediction, which is a significant limitation of the healthcare decision-support system [12].

Deep learning and Artificial Neural Networks (ANN) have been used to predict diabetes as well. Such models can learn very complex non-linear associations and have been found to be highly predictive in a series of studies [7]. Neural networks, however, require massive amounts of data, extensive amounts of computational resources, and parameter optimization. In addition, they are non-transparent and explainable which poses a question of trust and adoption by clinics.

Random Forest, Gradient Boosting, and AdaBoost are ensemble learning methods that have been proved to have high predictive power when many models are used to minimize variance and overfitting [11]. In specific, random forest classifiers have demonstrated good performance in diabetes prediction. However, ensemble models make systems more complex and less interpretable, and this can restrict their use in real-life clinical settings where explanations and trust are critical factors [14].

It is to this end that Decision Tree classifiers have come out as a solution with moderation and high interpretation. As demonstrated by several studies, Decision Trees are at least effective in structured medical data and can work with missing values, categorical attributes, and feature interactions effectively [8], [12]. The hierarchical format of the Decision Trees enables medical workers to follow lines of prediction and interpret the role of clinical variables in the outcomes. One major shortcoming that has been found in the current literature is that most studies have focused on binary classification in which they have classified victims of diabetes as diabetic or non-diabetic. According to recent studies, it is crucial to consider the pre-diabetic stage as a distinct category so that early preventive intervention is possible [10]. Also, the issue of class imbalance is frequent in the datasets of diabetes, especially in multi-class cases. Such methods as the Synthetic Minority Over-sampling Technique (SMOTE) are demonstrated to increase minority classes prediction performance and decrease model bias considerably [13].

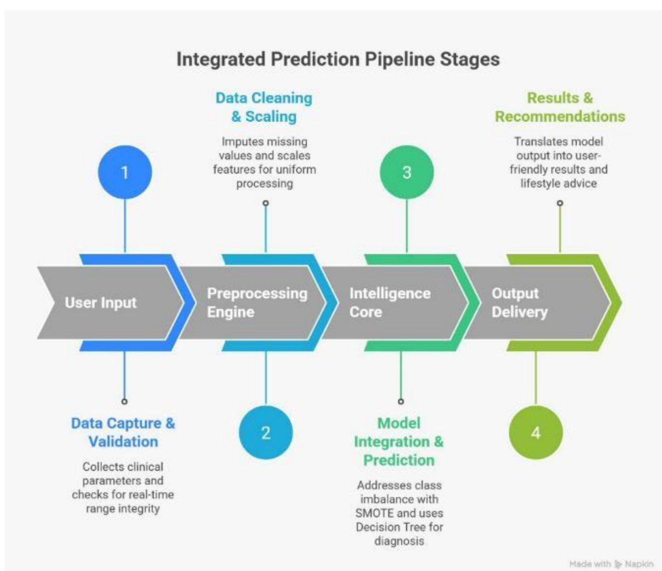
Relying on these findings, the current work contributes to the existing studies and suggests an explainable, multi-class diabetes prediction model that considers the complex data preprocessing procedures, the technique of dealing with the class imbalance, and the integration of real-time functionality by implementing a web-based application.

Proposed Methodology

Data Refining and Feature Engineering

The integrity of the predictive model relies on a rigorous data preparation phase executed by the preprocessing engine shown in Stage 2 of [Figure 1].

- Data Cleaning and Imputation: Clinical features containing zero values--which are physiologically inconsistent for metrics like Glucose, BMI, and Cholesterol--are treated as missing data.
- Normalization and Standardization: To prevent features with large numeric ranges from disproportionately influencing the model, StandardScaler is applied, centering data at a zero mean and scaling it to unit variance.
- Categorical Encoding: Qualitative variables, specifically gender, are transformed into binary numerical formats to ensure algorithmic compatibility.

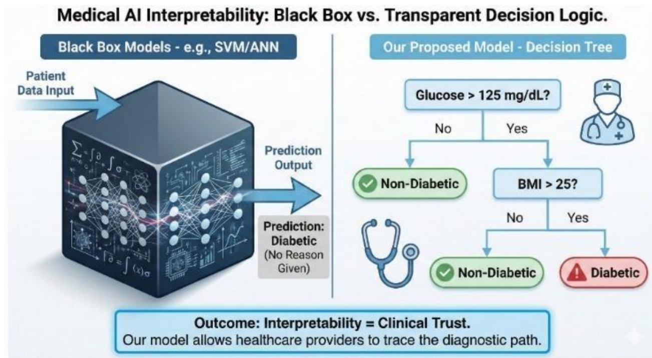


Algorithmic Selection: The Interpretability Factor

Although there are several ensemble approaches, the primary intelligence core that was chosen is the Decision Tree Classifier. The major benefit of such a method, as illustrated in comparison to conventional black-box models (such as SVM or ANN) in [Figure 2], is its natural interpretability.

- Rule-Based Reasoning: The Decision Tree is executed by a hierarchical tree of logical branching (e.g., verification of Glucose thresholds initially); this is illustrated on the right of [Figure 2]. This resembles clinical logic of professionals, which promotes trust with health practitioners.

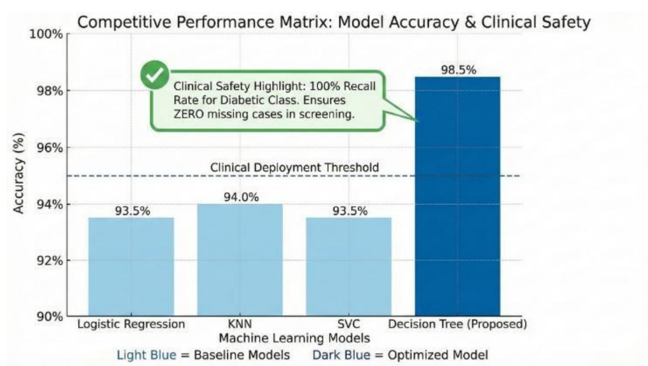
- Multi-Class Strategy: The approach does not merely make binary outcomes but has a pre-diabetic group. This can be used as an early-detection system of preventive action.
- Class Balancing: To address dataset imbalance, Synthetic Minority Oversampling Technique (SMOTE) is incorporated before the model is being trained to ensure that the model is not biased in favour of the majority.



Comparative Performance Analysis

Empirical evaluation against baseline classifiers confirms the superior efficacy of the proposed Decision Tree model, as detailed in the performance matrix in [Figure 3].

- Accuracy Dominance: While baseline models like Logistic Regression and SVC achieved respectable accuracies of 93.5%, the optimized Decision Tree reached a peak accuracy of 98.5%, surpassing the clinical deployment threshold.
- Clinical Safety Highlight: Most critically, as emphasized by the callout in [Figure 3], the model demonstrated a 100% Recall rate for the Diabetic class. In a healthcare context, this is paramount, as it ensures that zero actual diabetic cases are missed during screening.



Comparative Study with Existing Results

Within the last several years, several machine learning methods are actively used predicting diabetes based on structured clinical and demographic data. Some of the algorithms used are Logistic Regression, Support Vector machine (SVM), Random Forest, k-Nearest neighbors (k- NN), Naive bayes or Decision Tree. Empirical investigations and systematic reviews suggest that these conventional machine learning models remain reliable in using tabular healthcare data because they are robust, less costly to compute, and simpler to execute than deep learning models [15], [16].

Majority of the available studies are limited to binary classification and classify subjects as diabetic or non- diabetic. Nevertheless, recent research points to the clinical significance of multi-classification, in which case pre- diabetic patients should be treated as a separate group. The strategy will result in the detection of risks at an earlier stage and help preventive healthcare. Although these have these benefits, multi-class classification has the disadvantages of class imbalance, higher model complexity, and decreased prediction stability especially in cases where real-world data are not very large [17].

The reported predictive performance in the literature is highly disparate based on the characteristics of the datasets, methods in feature selection and strategies of validation. The accuracy values are usually in the 75 to 88 percent range with ensemble type models like the Random Forest and the Gradient Boosting model being rated towards the 88 percent mark. Nevertheless, there is often a trade-off of such improvements in terms of higher model complexities, less transparency, and lower interpretability, which can limit their use in clinical decision support settings where explainability is necessary [18].

However, Decision Tree classifiers have been popular with medical prediction and diagnosis problems because of their clear rule-based format. The top-down decision-making of the Decision Trees enables healthcare practitioners to have a clear trace of the prediction routes and to have an idea of the role played by major clinical parameters that include glucose concentration, body mass index (BMI), blood pressure, and lipid profiles. According to several recent studies, interpretable models such as Decision Trees are still competitive with more complex algorithms when clinical trust, usability, and explainability are put in higher priority than marginal improvement of accuracy [19].

Another factor that has been widely documented as problematic with diabetes prediction datasets is class imbalance, i.e. the presence of multiple outcome classes i.e. non-diabetic, pre-diabetic and diabetic. Existing literature shows that the data-level methods, including Synthetic Minority Oversampling Technique (SMOTE) and controlled resampling techniques, show excellent predictive power on minority classes and reduce bias towards the majority-class results. This approach to imbalance handling has now become a common method of medical research in machine learning [20].

Though a few recent studies suggest slightly increased accuracy with deep learning architecture or sophisticated ensemble models, most of them are restricted to offline experimental scenarios and cannot be applied in practice in a real-time clinical application. Conversely, the current project takes a balanced approach that focuses on predictive reliability, model interpretability, good classes imbalance control, and real-time implementation in the form of Flask- based web application. This technique is consistent with the new developments in applied healthcare analytics that propose practical, explainable, and deployable systems instead of purely accuracy-oriented models [21].

Conclusion

In this research, machine learning can be used to predict diabetes on time using well-structured clinical datasets. Utilization of Decision Tree Classifier model is done that predicted patients with three classes Non-Diabetic, Pre- Diabetic and Diabetic. Using multi-classification makes this solution able to show risk level compared to binary models allowing for predictions that can be acted on with early prevention. Preprocessing techniques such as handling invalid and missing values, scaling, encoding categorical features, balancing data using resampling techniques are used. This produced a well-trained model that predicts high and unbiased accuracy for each class.Our Decision Tree model was able to capture non-linear relationships between features such as Glucose, BMI, Lipid and Age. Producing good enough results while still being able to explain individual feature impact on the predictions. This ability is crucial when it comes to models' purpose of being used in healthcare-based applications.Deployment of the model using flask web application was able to take user input of necessary features and return the predicted result. Validations and well-defined output classes also help enhance the trust of the system allowing the model to be used outside of a lab environment. Conclusively, machine learning can be used as a Diabetes risk screening assistant. Early detection of such lifestyle diseases can allow users to take the necessary actions towards preventing Diabetes.

Future Scope

The diabetes prediction system has few opportunities for improvement and extension despite the attained satisfactory performance and usability.

The next step of the work could be the improvement of the model using and comparing the advanced algorithms like Random Forest, Support Vector Machines, Gradient Boosting, and Neural Networks. Ensemble methods may enhance generalization and accuracy of forecasting, especially when trained using larger data.

System expansion can also enhance its effectiveness through expansion of the datasets. To be more robust and minimize bias, the inclusion of the data of various healthcare institutions and various groups of the population should be considered. Also, it is possible to add more clinical and lifestyle parameters, including HbA1c levels, blood pressure, insulin levels, family history, and physical activity patterns, which would make it more clinical.

The other opportunity is the incorporation of explainable AI methods like the feature importance analysis or rule-based visualization. Such approaches would enhance transparency and trust by making users and healthcare professionals gain a better understanding of the logic behind predictions.

Application-wise, the system can be scaled into a mobile platform to enhance accessibility, particularly in the remote areas or resource limited areas. Combined with electronic health records or wearables, automated data gathering and constant health establishment could be provided.

Further applications can also consider the use of longitudinal analysis, which will be predicted using patient information over time instead of using a one-time input. This would help track the progression of the disease and detect the changes between pre-diabetes and diabetes more objectively.

Lastly, the system would be applicable in the real-world scenario with high scalability by deploying it to a secure cloud infrastructure that would be properly authenticated, encrypted, and aligned with the healthcare data protection standards. These improvements have the potential of making the existing system an all-inclusive and scalable healthcare decision-support tool.

References

1. M.A.B. Khan, M.J. Hashim, J.K. King, R.D. Govender, H. Mustafa, and J. Al Kaabi, "Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends," *Journal of Epidemiology and Global Health*, vol. 10, pp. 107–111, 2020. https://doi.org/10.2991/JEGH.K.191028.001
2. Templer, S. Abdo, and T. Wong, "Preventing Diabetes Complications," *International Medical Journal*, vol. 54, pp. 1264–1274, 2024. https://doi.org/10.1111/imj.16455
3. B.P. Kovatchev, "Metrics for Glycaemic Control: From HbA1c to Continuous Glucose Monitoring," *Nature Reviews Endocrinology*, vol. 13, pp. 425–436, 2017. https://doi.org/10.1038/nrendo.2017.3
4. G.E. Umpierrez and B. Kovatchev, "Glycaemic Variability: How to Measure and Its Clinical Implication for Type 2 Diabetes," *American Journal of Medicine Science*, vol. 356, pp. 518–527, 2018. https://doi.org/10.1016/j.amjms.2018.09.010
5. W. Barclay et al., "Glycemic Index, Glycemic Load, and Chronic Disease Risk: A Meta-Analysis," *American Journal of Clinical Nutrition*, vol. 87, pp. 627–637, 2008.
6. E. Gerontiti et al., "Low Glycemic Index Diets in Type 2 Diabetes: An Updated Review," *Hormones (Basel)*, 2024. https://doi.org/10.1007/s42000-024-00566-7
7. B.J. Erickson and F. Kitamura, "Performance Metrics for Machine Learning Models," *Radiology: Artificial Intelligence*, vol. 3, 2021. https://doi.org/10.1148/ryai.2021200126
8. P. Kavakiotis et al., "Machine Learning and Data Mining Methods in Diabetes Research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017.
9. Horestani F.J., Mehdi O., and Owrang M., "Predicting Diabetes Using Machine Learning and Health Factors," n.d.
10. Choudhury et al., "Machine Learning Approaches for Diabetes Prediction," *PLOS ONE*, 2024. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0300785
11. RIT Thesis Repository, "Comparative Study of ML Algorithms for Diabetes Detection," *Rochester Institute of Technology*. https://repository.rit.edu/cgi/viewcontent.cgi?article=13493
12. *Nature Scientific Reports*, "Explainable Machine Learning Models in Healthcare," 2025. https://www.nature.com/articles/s41598-025-03718-w
13. Chawla et al., "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, 2002. https://pmc.ncbi.nlm.nih.gov/articles/PMC5524285/
14. *MDPI Healthcare*, "Interpretable AI in Medical Decision Support Systems," 2025. https://www.mdpi.com/2227-9032/13/20/2588
15. Horestani, F.J., and Mehdi Owram, M., "Predicting Diabetes with Machine Learning: Analysis of Income and Health Factors," 2024.
16. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I., "Machine Learning and Data Mining Methods in Diabetes Research," *Computational and Structural Biotechnology Journal*, 2017.
17. Zhang, Y., et al., "Multi-Class Diabetes Prediction with Class Imbalance Consideration," *PLOS ONE*, 2024.
18. Rothermel, J., "Comparative Evaluation of Machine Learning Models for Diabetes Prediction," *Rochester Institute of Technology Thesis*, 2023.
19. Erickson, B.J., and Kitamura, F., "Performance Metrics and Interpretability of Machine Learning Models in Healthcare," *Scientific Reports*, 2025.
20. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P., "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, 2002.
21. Dimitriadis, S., et al., "Explainable Machine Learning Models for Real-Time Clinical Decision Support," *Healthcare Journal (MDPI)*, 2024.