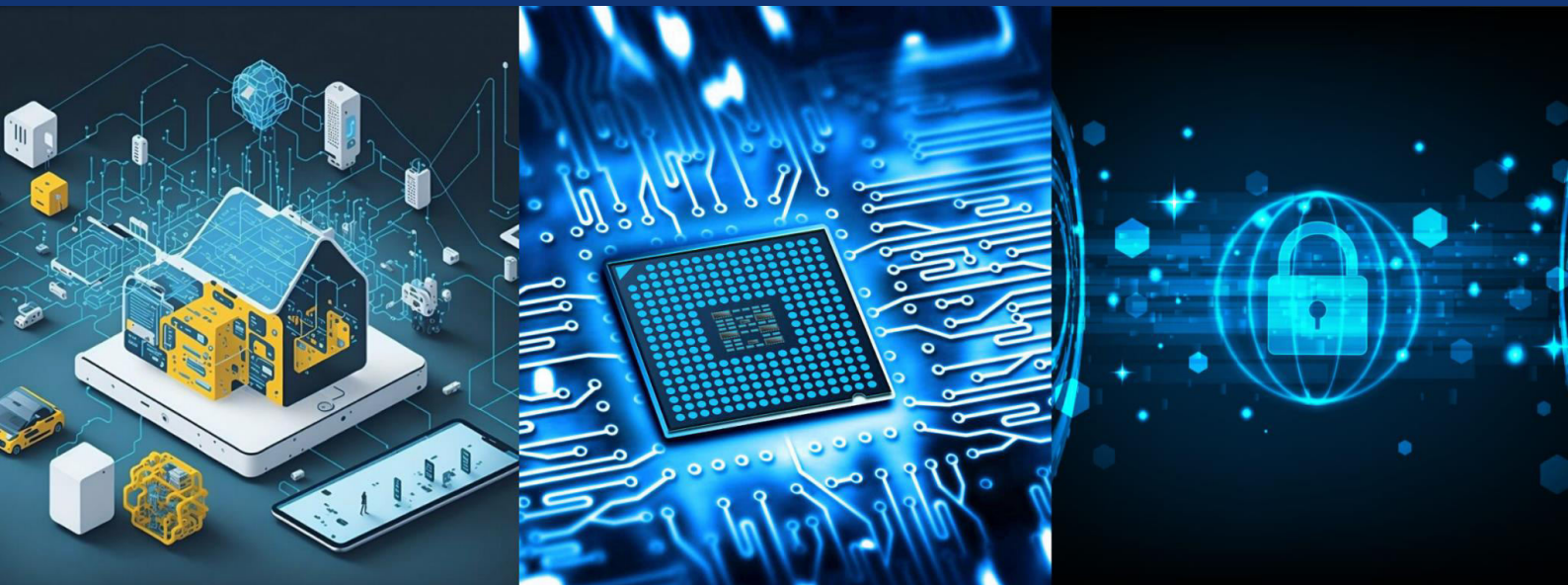
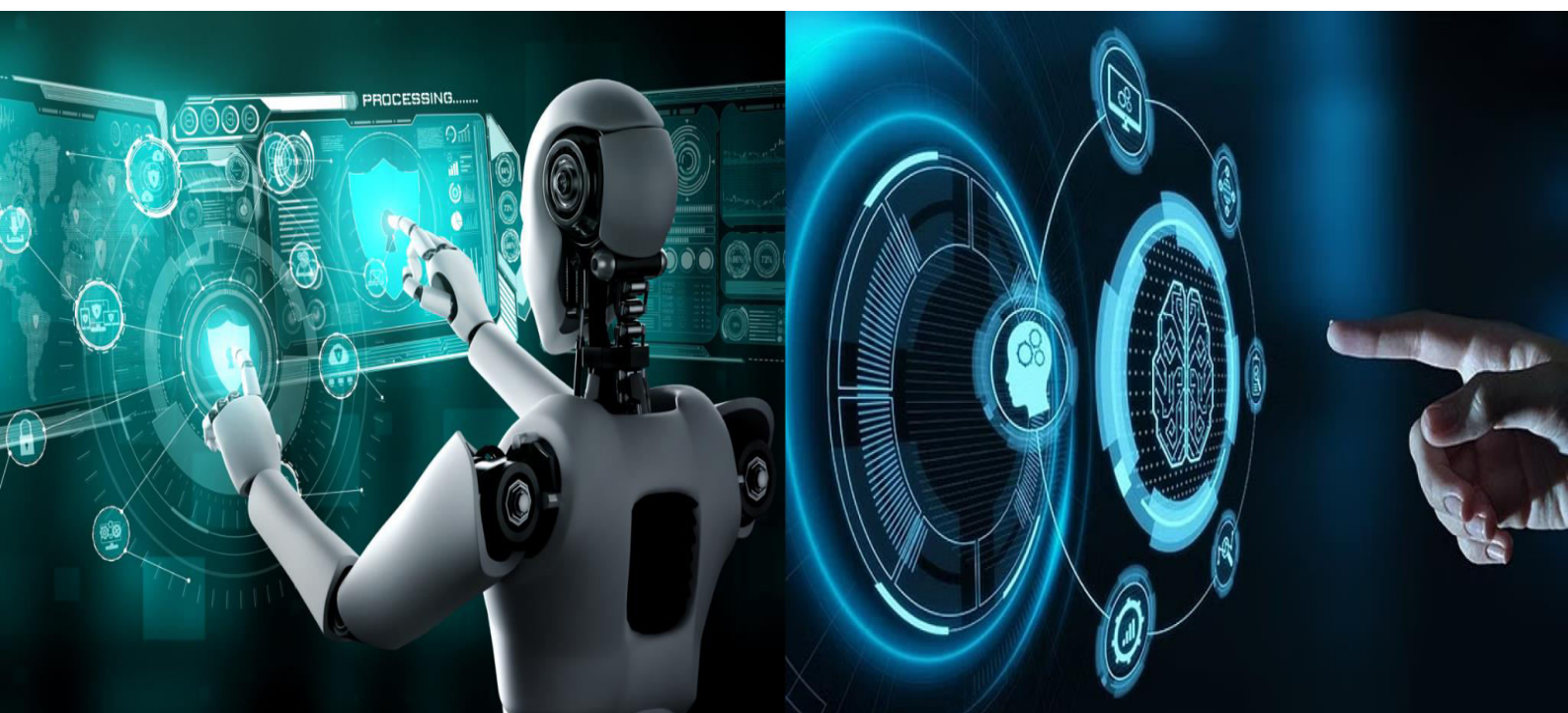




International Journal of Innovative Research in Computer and Communication Engineering

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





Machine Learning-Based Early Prediction of Type 2 Diabetes Mellitus using Clinical and Demographic Features: A Comparative Analysis of Classification Models

Nancy Singhal¹, ²Dr. Saurabh Charaya, Dr. Sandeep Kumar³

Research Scholar, Department of Computer Science and Engineering, Om Sterling Global University, Hisar, Haryana, India¹

Professor, Department of Computer Science, School of Engineering and Technology, Om Sterling Global University, Hisar, Haryana, India²

Professor and Associate Dean - Data Science, School of Computer Science & Artificial Intelligence, SR University, Warangal, Telangana, India³

ABSTRACT: Type 2 diabetes mellitus (T2DM) is a major chronic metabolic disease in which delayed identification can increase the risk of vascular, renal, neurological and cardiovascular complications. This study evaluates the usefulness of supervised machine learning for classifying diabetes status from demographic, behavioural and routinely available clinical variables. The analysis used a publicly available Diabetes Prediction Dataset containing 100,000 observations before cleaning. After the removal of 3,854 duplicate records, 96,146 observations remained. Predictors included gender, age, hypertension, heart disease, smoking history, body mass index (BMI), HbA1c level and blood-glucose level, with diabetes status as the binary target. Categorical variables were numerically encoded, numerical predictors were scaled, and an 80:20 hold-out strategy with random state 42 was used for model development and testing. Logistic Regression, Decision Tree and Random Forest classifiers were compared using confusion-matrix-derived accuracy, precision, sensitivity, specificity and F1-score. On the reported test set of 19,230 observations, the Decision Tree achieved the highest overall accuracy (97.06%) and F1-score (80.32%) with 100% specificity, whereas Random Forest achieved the highest sensitivity for diabetes-positive cases (69.09%) and the fewest false negatives (532). Logistic Regression provided an interpretable baseline but had the lowest sensitivity (62.52%). The findings show that a model with the highest overall accuracy is not necessarily the most suitable model for screening. In early case finding, sensitivity and the false-negative burden require particular attention. The study therefore supports clinically supervised use of machine learning as a decision-support aid, subject to external validation, calibration, fairness assessment and careful deployment.

KEYWORDS: Type 2 diabetes mellitus, machine learning, artificial intelligence, diabetes prediction, Logistic Regression, Decision Tree, Random Forest, clinical decision support.

I. INTRODUCTION

Type 2 diabetes mellitus is characterized by insulin resistance together with progressive impairment of insulin secretion. Because its development may be gradual, a substantial period can pass before affected individuals receive a diagnosis. During this interval, metabolic abnormalities may progress and the risk of long-term complications can increase. Early recognition is therefore an important component of preventive healthcare and clinical risk management. Risk is not determined by a single characteristic. Age, excess body weight, hypertension, inactivity, family history, smoking behaviour and glycaemic abnormalities can occur in different combinations. In clinical practice, HbA1c and blood-glucose measurements are central indicators of glycaemic status, while demographic and comorbidity information helps place these measurements in a broader risk context. The combination of multiple heterogeneous predictors creates a setting in which data-driven classification can be useful.



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

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

Artificial intelligence (AI), particularly supervised machine learning (ML), offers a mechanism for learning patterns between labelled outcomes and multiple predictor variables. In diabetes research, methods such as Logistic Regression, Decision Trees, Random Forests, support vector machines and neural networks have been used for risk stratification, diagnostic support and complication prediction. These models differ in interpretability, flexibility, sensitivity to nonlinear relationships and susceptibility to overfitting. Consequently, model comparison should not be based on accuracy alone.

A clinically useful predictive model should be evaluated in relation to the intended use. A screening-oriented system must reduce missed positive cases and therefore requires close attention to sensitivity and false negatives. A confirmatory or triage-oriented tool may place comparatively greater emphasis on precision and specificity. In addition, the usefulness of any medical AI model depends on transparency, calibration, population representativeness, fairness, privacy protection and integration into clinician-directed workflows.

The present paper compares three commonly used supervised classifiers - Logistic Regression, Decision Tree and Random Forest - using a large public diabetes dataset. The paper follows the comparative and clinically oriented structure of the supplied journal sample by presenting a detailed literature context, transparent methodology, model-wise results, interpretation of trade-offs, practical implications, limitations and responsible-use considerations.

II. LITERATURE REVIEW

Machine Learning in Diabetes Prediction

Prior work has established diabetes prediction as a prominent application area for machine learning and data mining. Kavakiotis et al. (2017) reviewed a wide range of diabetes-related prediction tasks and emphasized the diversity of data sources, algorithms and clinical objectives. This broader literature indicates that predictive performance is strongly influenced by data quality, variable selection, preprocessing and the definition of the outcome being predicted.

Population-based modelling has also demonstrated the feasibility of using routinely collected demographic and health information for diabetes-related risk prediction. Dinh et al. (2019) examined data-driven prediction of diabetes and cardiovascular disease, while Lai et al. (2019) evaluated predictive models for diabetes mellitus using machine learning techniques. These studies contribute to the rationale for combining demographic variables, comorbidities and glycaemic indicators in classification frameworks.

Comparative Classification Approaches

Comparative studies are important because different algorithms optimize different aspects of performance. Logistic Regression is frequently used as an interpretable statistical baseline. Its coefficients provide a relatively transparent account of direction and relative contribution, and predicted probabilities can be communicated in a clinically meaningful manner. However, a linear decision structure may not capture complex nonlinear relationships unless interaction terms or transformations are explicitly incorporated.

Decision Trees offer easily traceable decision rules and can model nonlinear relationships and interactions without requiring a predefined functional form. Their interpretability makes them attractive for clinical explanation, but a single tree can be unstable and prone to overfitting if depth, pruning and minimum-sample settings are not carefully controlled. Random Forest addresses some of this instability by combining many decision trees generated from resampled data and subsets of features. The ensemble can improve robustness and sensitivity, though it is less directly interpretable than a single tree.

Hasan et al. (2020) examined ensemble-based diabetes prediction, while Tigga and Garg (2020), Joshi and Dhakal (2021), and Khanam and Foo (2021) compared multiple classification approaches for Type 2 diabetes prediction. Their inclusion in the literature base supports a comparative rather than single-model approach. Fregoso-Aparicio et al. (2021) further synthesized machine learning and deep learning prediction approaches, reinforcing the need to interpret reported performance in relation to dataset composition and validation design.

Preprocessing, Explainability and Clinical Translation

Preprocessing is a central methodological issue in diabetes prediction. Olisah et al. (2022) focused on diabetes prediction and diagnosis from a preprocessing and machine-learning perspective, demonstrating why encoding, scaling,



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

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

class imbalance and leakage control can materially affect the apparent quality of a model. In the present study, categorical variables are encoded and continuous variables are scaled before modelling. Because the positive diabetes class represents a relatively small proportion of the cleaned dataset, metrics that focus on the minority class are essential.

Recent research has increasingly emphasized explainable AI. Tasin et al. (2023) examined diabetes prediction using machine learning and explainable AI techniques. Explainability is particularly relevant in health contexts because clinicians need to understand why a system produces a given risk classification and whether the influential features are clinically plausible. Methods such as feature importance, local explanation and SHAP analysis can support this requirement, although such analyses were not reported in the current dataset results.

More recent studies, including Mansoori et al. (2023), Modak and Jha (2024), Abdollahi and Aref (2024), and Rönn et al. (2024), illustrate the continuing development of feature-based and higher-dimensional predictive approaches. Collectively, the reviewed literature does not support the idea of a universally optimal diabetes classifier. Instead, model choice should reflect the target population, available variables, class distribution, error costs, validation evidence and intended clinical role.

III. OBJECTIVES OF THE STUDY

- To develop an AI/ML-based framework for predicting Type 2 diabetes from demographic, behavioural and clinical variables.
- To compare Logistic Regression, Decision Tree and Random Forest models on an independent hold-out test split.
- To interpret model performance using accuracy, precision, sensitivity, specificity and F1-score rather than accuracy alone.
- To examine the false-negative burden of the models because missed diabetes-positive cases are important in screening-oriented use.
- To identify practical implications, limitations and safeguards for responsible clinical use of diabetes prediction systems.

IV. MATERIALS AND METHODS

Study Design and Data Source

The study used a retrospective secondary-data design based on the publicly available Diabetes Prediction Dataset hosted on Kaggle and attributed in the source paper to Mustafa et al. (2021). The original file contained 100,000 observations. After removal of 3,854 duplicate records, 96,146 unique records remained for analysis. The dependent variable was binary diabetes status, coded as non-diabetic (0) or diabetic (1). Because this is a secondary dataset, the analysis is suitable for methodological comparison but does not constitute a prospective clinical cohort or a population-specific diagnostic validation study.

Predictor Variables

The predictor set combined demographic, behavioural, anthropometric, comorbidity and glycaemic information. This combination allows the models to evaluate both background risk indicators and direct clinical markers. The variables used in the classification framework are summarized in Table 1.



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

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

Table.1. Variables used for Type 2 diabetes classification

Variable	Type	Role in prediction
gender	Categorical	Demographic predictor
age	Continuous	Demographic predictor
hypertension	Binary	Comorbidity predictor
heart disease	Binary	Comorbidity predictor
smoking history	Categorical	Behavioural predictor
BMI	Continuous	Anthropometric predictor
HbA1c level	Continuous	Long-term glycaemic marker
blood glucose level	Continuous	Clinical glycaemic marker
diabetes	Binary	Outcome/target

Data Preprocessing

Gender and smoking-history categories were converted to numerical representation using label encoding in the reported workflow. Continuous predictors were normalized using min-max-based scaling. The data were then prepared for supervised classification. The cleaned dataset was divided into training and testing partitions using an 80:20 hold-out strategy with random state = 42, producing a reported test set of 19,230 observations.

The class distribution was imbalanced: approximately 8.82% of observations were diabetes-positive. This imbalance is methodologically important because a classifier can achieve a high overall accuracy by correctly predicting the dominant non-diabetic class while still missing a clinically important share of positive cases. For this reason, sensitivity, precision and F1-score were interpreted alongside accuracy and specificity.

Model Development

Three supervised classifiers were evaluated. Logistic Regression served as the interpretable baseline model. The Decision Tree represented a transparent nonlinear classifier capable of capturing interaction-like decision rules. Random Forest represented an ensemble approach designed to reduce the instability associated with a single decision tree by aggregating predictions across multiple trees. The models were compared on the same reported hold-out test set so that confusion-matrix counts and derived performance measures could be interpreted consistently.

Evaluation Measures

Performance was evaluated from true negatives (TN), false positives (FP), false negatives (FN) and true positives (TP). Accuracy represents the proportion of all observations correctly classified. Precision represents the proportion of predicted diabetes-positive cases that were truly positive. Sensitivity represents the proportion of actual diabetes-positive cases correctly detected. Specificity represents the proportion of actual non-diabetic cases correctly classified. F1-score provides a harmonic balance between precision and sensitivity.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$$

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Sensitivity}) / (\text{Precision} + \text{Sensitivity})$$

For a screening-oriented application, sensitivity is especially important because a false negative represents a person with diabetes who is classified as non-diabetic. The clinical cost of such an error can be greater than the cost of a false positive that leads to additional confirmatory assessment. Therefore, the interpretation gives explicit attention to the number of false negatives produced by each model.

V. RESULTS

Descriptive Characteristics of the Analytical Dataset

The final analytical dataset contained 96,146 observations after duplicate removal. It included a broad age range and clinically relevant information on adiposity, hypertension, heart disease, smoking history, HbA1c and blood glucose.



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

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

The descriptive summary reported a diabetes-positive prevalence of 8.82%, hypertension prevalence of 7.8%, heart-disease prevalence of 4.1% and mean BMI of 27.32 kg/m². The imbalance of the diabetes target means that minority-class performance measures are necessary for meaningful interpretation.

Table.2. Key descriptive findings reported in the source analysis.

Indicator	Reported value	Interpretation
Final cleaned observations	96,146	Large secondary dataset after duplicate removal
Train/test allocation	80% / 20%	Reported test-set size: 19,230
Diabetes-positive prevalence	8.82%	Imbalanced outcome; minority-class metrics are necessary
Hypertension prevalence	7.8%	Clinically meaningful comorbidity
Heart-disease prevalence	4.1%	Low frequency but relevant predictor
Mean BMI	27.32 kg/m ²	Average in overweight range

Confusion-Matrix Results

Table 3 presents the reported test-set confusion matrices. These counts provide a direct view of each model's error pattern. Logistic Regression produced 645 false negatives, Decision Tree produced 566, and Random Forest produced 532. Thus, Random Forest missed the fewest diabetes-positive cases. The Decision Tree produced no false positives in the reported hold-out results, while Random Forest produced 65 and Logistic Regression produced 182.

Table.3. Reported test-set confusion matrices (n = 19,230 for each model).

Model	TN	FP	FN	TP
Logistic Regression	17,327	182	645	1,076
Decision Tree	17,509	0	566	1,155
Random Forest	17,444	65	532	1,189

The Decision Tree's absence of false positives resulted in 100% specificity. However, it still failed to identify 566 positive cases. Random Forest correctly identified the largest number of positive observations (1,189) and generated the smallest false-negative count (532). Logistic Regression remained a transparent baseline but identified the fewest positive cases and produced the greatest false-negative burden among the three models.

Comparative Performance of the Classifiers

Table.4. Performance metrics derived from the reported confusion-matrix counts; positive-class metrics refer to diabetes-positive observations.

Model	Accuracy	Precision	Sensitivity	Specificity	F1-score
Logistic Regression	95.70%	85.53%	62.52%	98.96%	72.22%
Decision Tree	97.06%	100.00%	67.12%	100.00%	80.32%
Random Forest	96.90%	94.81%	69.09%	99.63%	78.90%



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

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

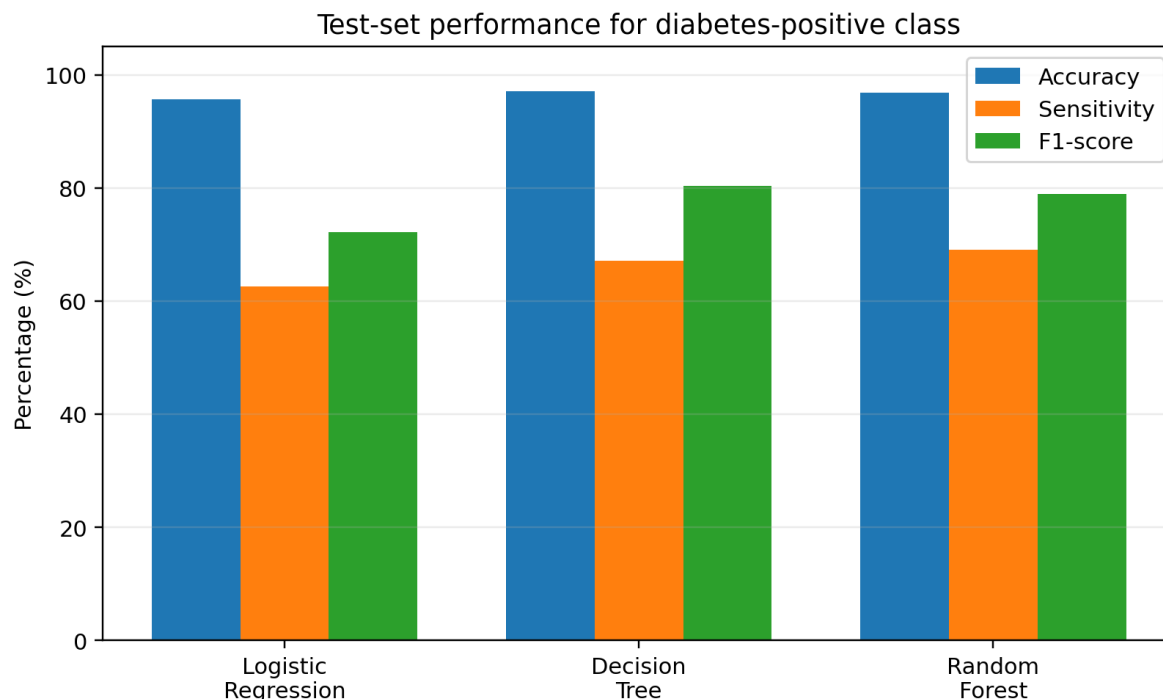


Figure.1.Comparative test-set performance Random Forest achieved the highest sensitivity, whereas Decision Tree achieved the highest accuracy and F1-score.

The Decision Tree achieved the highest overall accuracy at 97.06%, followed closely by Random Forest at 96.90% and Logistic Regression at 95.70%. The Decision Tree also achieved the highest F1-score (80.32%) and reported perfect precision and specificity on the hold-out split. Random Forest achieved the highest sensitivity (69.09%), indicating the strongest detection rate for diabetes-positive observations. Logistic Regression had the lowest sensitivity (62.52%) and F1-score (72.22%).

Model-Wise Interpretation

Decision Tree: The Decision Tree showed the strongest overall accuracy and F1-score in the reported comparison. Its 100% specificity means that no non-diabetic observation in the reported test set was classified as diabetic. However, the model still missed 566 positive cases. Because a single tree can be sensitive to the exact train-test split and to its structural settings, this result should not be interpreted as proof that the model will retain perfect specificity in new populations. Pruning, repeated validation and external testing are required before clinical use.

Random Forest: Random Forest achieved 96.90% accuracy and the highest sensitivity (69.09%). It correctly detected 1,189 diabetes-positive observations and produced only 532 false negatives, the lowest among the three models. This profile makes Random Forest comparatively attractive when the intended use places greater weight on finding positive cases. Its precision (94.81%) and specificity (99.63%) also remained high, indicating that the gain in sensitivity did not require a large increase in false positives.

Logistic Regression: Logistic Regression achieved 95.70% accuracy, 85.53% precision, 62.52% sensitivity, 98.96% specificity and a 72.22% F1-score. Its principal advantage is interpretability and the potential to express risk through model coefficients and predicted probabilities. Its lower sensitivity suggests that, if early case finding is the priority, threshold adjustment, class weighting or carefully controlled resampling methods should be considered in future work.



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

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

VI. COMPARATIVE STUDY

The study conducted by us substantially extends the earlier diabetes-prediction study by Noronha et al. (2025). The reference study used 768 records and mainly compared SVM and Random Forest, with SVM achieving 75.97% accuracy and an overall F1-score of 0.76. In comparison, proposed's study used 96,146 cleaned observations and systematically evaluated Logistic Regression, Decision Tree, and Random Forest.

Performance Indicator	Reference Study	Proposed Best Result	Improvement
Accuracy	75.97%	97.06%	+21.09 percentage points
F1-score	76.00%	80.32%	+4.32 percentage points
Diabetes-positive Recall/Sensitivity	65.00%	69.09%	+4.09 percentage points
Dataset Size	768	96,146	~125.2× larger

The comparison indicates that the major improvement in proposed study is not limited to higher accuracy. The study uses a substantially larger dataset, provides complete confusion-matrix results, distinguishes between accuracy and screening sensitivity, and explicitly considers the clinical consequences of false-negative predictions. The Decision Tree improved the reported best accuracy from 75.97% to 97.06%, representing an increase of 21.09 percentage points, while Random Forest provided the strongest screening-oriented performance with 69.09% sensitivity and only 532 false negatives.

VII. DISCUSSION

The comparison demonstrates that model selection changes when performance is viewed through a clinical rather than purely statistical lens. If overall accuracy were the only criterion, the Decision Tree would be preferred because it reached 97.06%. However, diabetes screening places particular importance on the identification of positive cases. Under that criterion, Random Forest performed better because it achieved the highest sensitivity and generated the fewest false negatives.

The class imbalance explains why this distinction matters. With only about 8.82% diabetes-positive observations, a model can perform extremely well on the majority non-diabetic class and still fail to identify a meaningful proportion of people with diabetes. All three models showed high specificity, but their sensitivities were substantially lower. Therefore, the principal modelling challenge in this dataset is not the recognition of non-diabetic observations; it is the reduction of missed positive cases.

The Decision Tree produced a particularly strong reported hold-out result, including perfect specificity and precision. Such performance is attractive but should be interpreted cautiously. A single tree can fit idiosyncratic partitions of the development data, and unusually strong hold-out values may not reproduce in a different sample. The source analysis itself recommends pruning, cross-validation and external validation. These steps are essential before concluding that the Decision Tree is consistently superior.

Random Forest offered the most balanced screening-oriented profile. Its sensitivity of 69.09% was higher than both the Decision Tree and Logistic Regression, while its specificity remained 99.63%. The ensemble therefore reduced the false-negative burden without producing a major loss of specificity. This is consistent with the theoretical advantage of aggregating multiple trees to reduce the instability of a single tree. Nevertheless, a sensitivity below 70% means that a substantial proportion of positive cases would still be missed if the model were used as a stand-alone screening mechanism.

Logistic Regression remains important despite lower predictive sensitivity because clinical adoption requires interpretability. A transparent baseline can reveal whether the performance gains of more flexible models justify their additional complexity. In practice, the selection between Logistic Regression and tree-based methods should consider not only discrimination but also calibration, explanation, deployment requirements, computational burden and ease of governance.



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

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

Clinical Interpretation of Error Costs

A false negative in diabetes prediction can delay confirmatory testing, counselling or intervention. For this reason, the lowest false-negative count in the reported comparison is a meaningful advantage for Random Forest. By contrast, a false positive can lead to additional testing or short-term anxiety, but it does not carry the same risk of leaving a true case unidentified. This asymmetry supports threshold selection strategies that explicitly prioritize sensitivity when a model is intended for preliminary screening.

At the same time, the variables HbA1c and blood-glucose are themselves central diagnostic indicators. Their availability may make classification easier than a true pre-diagnostic risk model that relies only on history, demographics and non-laboratory variables. Thus, high classification performance should not be interpreted as evidence that the models can replace established diagnostic testing. Instead, the models are better understood as analytical decision-support tools operating alongside clinical assessment.

Implications for Responsible AI Deployment

Before deployment, the models should be evaluated on geographically and clinically distinct cohorts. External validation is necessary to determine whether the reported rankings remain stable outside the public dataset. Calibration should also be assessed because a clinically useful risk score must not only rank people correctly but should provide probabilities that correspond reasonably to observed risk. Subgroup analysis by age, sex and other relevant characteristics is needed to detect uneven performance and potential bias.

Explainability should form part of the deployment framework. For tree-based models, feature importance and local explanation techniques can help clinicians identify which variables contributed to a classification. Methods such as SHAP can make individual predictions more transparent, but explanation should not be mistaken for proof of causality. A model may rely on statistically predictive relationships that do not represent causal mechanisms.

Data governance is equally important. Even when a model performs well, patient information must be protected through data minimization, controlled access, de-identification and appropriate regulatory or institutional review. Automated predictions should not directly determine treatment. The final decision should remain with qualified healthcare professionals who can integrate the model output with symptoms, examination findings, laboratory confirmation and the broader patient context.

VIII. PRACTICAL IMPLICATIONS AND RECOMMENDATIONS

The reported model comparison supports the use of ML as a clinician-directed decision-support component rather than an autonomous diagnostic system. A practical workflow could use a predictive model to flag individuals who may benefit from confirmatory testing, preventive counselling or earlier follow-up. For such a workflow, the following improvements are recommended:

- Optimize classification thresholds according to clinical risk tolerance, with explicit emphasis on reducing false negatives in screening settings.
- Apply stratified k-fold cross-validation and repeated validation so that performance is not dependent on one hold-out split.
- Conduct external and temporal validation on cohorts from different institutions, regions and time periods.
- Use one-hot or clinically defensible encoding for nominal variables and compare class-weighting or SMOTE-based approaches with strict leakage control.
- Report ROC-AUC, precision-recall AUC, calibration curves and decision-threshold analyses in addition to accuracy and F1-score.
- Assess subgroup performance and fairness by age, sex and other available demographic characteristics.
- Use explainability tools such as SHAP or local feature contributions to support clinician review of predictions.
- Monitor model drift, calibration and error patterns after deployment, with a defined process for model review or withdrawal if performance deteriorates.
- Protect sensitive data through access control, de-identification, data minimization and institutional governance.
- Maintain human oversight: the model should prompt assessment and confirmatory testing, not independently diagnose diabetes or prescribe treatment.



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

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

IX. LIMITATIONS

The study used a public secondary dataset and therefore cannot establish causal relationships or guarantee performance in any specific hospital, community or demographic group. The reported results are based on one 80:20 hold-out split. Without repeated cross-validation or an external test cohort, the stability of the model ranking remains uncertain. The Decision Tree's perfect specificity, in particular, should be treated as a hold-out result rather than a generalizable clinical property.

A second limitation concerns the predictor set. HbA1c and blood-glucose levels are powerful clinical glycaemic markers and are themselves central to diabetes assessment. Their inclusion can produce stronger classification performance than would be expected from a purely pre-diagnostic risk model using demographics, lifestyle and comorbidities alone. Future work should therefore compare full clinical models with restricted early-risk models that exclude direct glycaemic markers.

The available results do not report calibration, ROC-AUC, precision-recall AUC, formal feature importance, explainability analysis, subgroup fairness testing or external validation. The present paper does not invent these unavailable statistics. These analyses should be added in future studies before translating the model into clinical practice.

X. CONCLUSION

This study demonstrates the potential of supervised machine learning for classifying Type 2 diabetes status from demographic, behavioural, comorbidity, anthropometric and glycaemic variables. Using the reported hold-out test results, Decision Tree achieved the highest overall accuracy (97.06%) and F1-score (80.32%), while Random Forest achieved the highest sensitivity (69.09%) and the lowest false-negative count (532). Logistic Regression provided the most interpretable baseline but had lower sensitivity for positive cases.

The comparison shows why medical AI should not be judged by accuracy alone. In a screening context, missed positive cases are a central concern, making sensitivity and false-negative burden important criteria for model selection. Random Forest therefore offers an attractive screening-oriented balance in the reported dataset, even though the Decision Tree achieved slightly higher overall accuracy. No model should be treated as a replacement for clinical assessment or laboratory confirmation.

For responsible translation, future work should emphasize cross-validation, external validation, calibration, subgroup fairness, threshold optimization and explainability. When embedded within clinician-led pathways and governed by appropriate privacy and safety controls, machine learning can support earlier identification, targeted follow-up and more efficient use of healthcare resources.

XI. ETHICAL AND DATA GOVERNANCE STATEMENT

This paper reports secondary analysis of a publicly available, de-identified dataset and does not describe direct recruitment, intervention or collection of identifiable information from human participants by the authors. The source material provided for this revision does not report a new institutional ethics approval number for the present secondary analysis. Any future prospective or institution-based validation should obtain the approvals, consent procedures and data-governance permissions required by the participating organizations and applicable regulations. Model outputs should remain advisory and should be reviewed by qualified clinicians.

REFERENCES

1. American Diabetes Association. (2024). Standards of care in diabetes-2024. Diabetes Care, 47(Supplement 1).
2. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116.
3. Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Medical Informatics and Decision Making, 19, 211.



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

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

4. Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocrine Disorders*, 19, 101.
5. Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, 8, 76516-76531.
6. Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, 167, 706-716.
7. Joshi, R. D., & Dhakal, C. K. (2021). Predicting type 2 diabetes using logistic regression and machine learning approaches. *International Journal of Environmental Research and Public Health*, 18(14), 7346.
8. Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, 7(4), 432-439.
9. Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. *Diabetology & Metabolic Syndrome*, 13, 148.
10. Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220, 106773.
11. Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, 10(1-2), 1-10.
12. Mansoori, A., Sahranavard, T., Hosseini, Z. S., et al. (2023). Prediction of type 2 diabetes mellitus using hematological factors based on machine learning approaches: A cohort study analysis. *Scientific Reports*, 13, 663.
13. Modak, S. K. S., & Jha, V. K. (2024). Diabetes prediction model using machine learning techniques. *Multimedia Tools and Applications*, 83, 38523-38549.
14. Abdollahi, J., & Aref, S. (2024). Early prediction of diabetes using feature selection and machine learning algorithms. *SN Computer Science*, 5, 217.
15. Rönn, T., Perfilyev, A., Oskolkov, N., & Ling, C. (2024). Predicting type 2 diabetes via machine learning integration of multiple omics from human pancreatic islets. *Scientific Reports*, 14, 14637.
16. Noronha, F., Singh, S., & Poojari, A. (2025). Diabetes prediction and risk analysis using machine learning techniques. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 12(3), e175-e182.



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