RESEARCH PAPER

Machine Learning-Based Predictive Modelling for Early Diagnosis of Type 2 *Diabetes Mellitus*: A Comparative Analysis of Supervised Classification Algorithms

Rupanjali Singh[1], Priyanka Bhandari[2], Swati Madan[3], Shikha Saxena[3], Nayyar Parvez[4], Lohitha Raja[5], Khaydarova Gulyora Zokirjon kizi[6] and Sumitha Vakati[7*]

[1]Indian Institute of Technology Jodhpur, NH-65, Nagaur Road, Nagaur, Rajasthan 342037, India.
[2]Department of Pharmacology, School of Pharmaceutical Sciences, SGRR University, Patel Nagar, Dehradun, India.
[3]Amity Institute of Pharmacy, Amity University, Sector 125, Noida, Uttar Pradesh 201313, India.
[4]School of Pharmacy, Sharda University, Greater Noida, Uttar Pradesh, India.
[5]Histology Biology Department, Fergana Medical Institute of Public Health, Yangi Turon 2A, Fergana 150100, Uzbekistan.
[6]Department of Folk Medicine and Pharmacology, Fergana Medical Institute of Public Health, Yangi Turon 2A, Fergana 150100, Uzbekistan.
[7]Department of Biotechnology, Sri Venkateswara College of Engineering, Chennai-Bengaluru High Road, Pennalur Village, Sriperumbudur, Kancheepuram District, Tamil Nadu 602117, India.

***Corresponding Author:**
Dr. V. Sumitha,
Professor, Dept. of Biotechnology,
Sri Venkateswara College of Engineering. Chennai-Bengaluru High Road, Pennalur Village, Sriperumbudur, Kancheepuram District, Tamil Nadu, 602117, India.
Email: sumitha@svce.ac.in

ABSTRACT

Background: Type 2 diabetes mellitus (T2DM) constitutes one of the most rapidly expanding metabolic disorders globally, with projections indicating that the affected population will surpass 783 million individuals by 2045. Timely and accurate prediction of T2DM at the pre-diabetic or early symptomatic stage is essential for reducing morbidity, limiting healthcare expenditure, and enabling targeted preventive interventions. Conventional clinical risk stratification tools often lack sufficient discriminatory power, underscoring the pressing need for robust computational approaches. **Objective:** The present study aimed to develop, train, and validate multiple supervised machine learning (ML) classification models to predict T2DM incidence using a curated dataset of clinical, biochemical, and lifestyle parameters, and to identify the optimal algorithm for clinical deployment. **Methods:** A retrospective dataset comprising 10,892 patient records was assembled from the PIMA Indian Diabetes Database supplemented with clinical registry data, encompassing features including fasting plasma glucose, glycated haemoglobin (HbA1c), body mass index (BMI), blood pressure, age, physical activity index, dietary quality score, family history, and socioeconomic indicators. Post-preprocessing entailing missing value imputation, z-score normalization, and one-hot encoding seven ML classifiers were trained: Logistic Regression, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, eXtreme Gradient Boosting (XGBoost), Multilayer Perceptron (MLP), and a Stacking Ensemble. Stratified 10-fold cross-validation was applied, and models were evaluated on Accuracy, Precision, Recall, F1-score, Specificity, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). **Results:** The Stacking Ensemble model achieved the highest overall performance with accuracy of 91.6%, precision of 90.8%, recall of 89.4%, F1-score of 90.1%, specificity of 92.3%, and AUC-ROC of 0.948. XGBoost performed second best (AUC-ROC = 0.931; accuracy = 89.3%), followed by the Neural Network (MLP; AUC-ROC = 0.922). Glucose concentration and HbA1c emerged as the most predictive features via SHAP-based importance analysis. **Conclusion:** The proposed ensemble framework demonstrates superior discriminatory capability for early T2DM prediction and offers a scalable, non-invasive adjunct to conventional diagnostic protocols. Integration of such models within electronic health record systems and wearable health-monitoring platforms holds significant promise for population-level diabetes prevention.

Keywords: Type 2 Diabetes Mellitus; Machine Learning; XGBoost; Stacking Ensemble; Predictive Modelling; Feature Selection; Clinical Informatics

How to cite this article: Singh R, Bhandari P, Madan S, Saxena S, Parvez N, Raja L, Kizi KGZ, Vakati S. Machine Learning-Based Predictive Modelling for Early Diagnosis of Type 2 Diabetes Mellitus: A Comparative Analysis of Supervised Classification Algorithms. Int J Drug Deliv Technol. 2026;16(35s): 1041-1048. DOI: 10.25258/ijddt.16.35s.116

1. INTRODUCTION

1.1 Global Burden of Type 2 Diabetes Mellitus

Type 2 diabetes mellitus (T2DM) is a chronic, progressive metabolic disorder characterised by insulin resistance, relative insulin secretory deficiency, and resultant hyperglycaemia. The International Diabetes Federation (IDF) estimates that 537 million adults (20–79 years) were living with diabetes worldwide in 2021, representing a global prevalence of approximately 10.5%, and projects this figure to reach 783 million by 2045[1]. The economic burden of the disease is equally formidable; global health expenditure on diabetes management exceeded USD 966 billion in 2021 and continues to escalate with growing complication rates[2]. In South Asia, including India, the burden is disproportionately high, with

India registering an estimated 77 million diabetic individuals in 2021, making it the second most affected nation globally[3]. Late-stage diagnosis remains a critical challenge; approximately 50% of individuals with T2DM are undiagnosed at any given time[4], which significantly escalates the risk of microvascular and macrovascular complications including diabetic nephropathy, retinopathy, neuropathy, and cardiovascular disease.

1.2 Importance of Early Prediction

Pre-diabetes characterized by impaired fasting glucose (IFG) or impaired glucose tolerance (IGT) represents a critical and reversible window of opportunity for intervention. Multiple large-scale randomized controlled trials have demonstrated that structured lifestyle modification programmes can reduce the risk of progression from pre-diabetes to overt T2DM by 40–70%[5]. Early prediction models that can identify at-risk individuals prior to clinical manifestation are therefore indispensable. Traditional risk scoring systems, such as the Finnish Diabetes Risk Score (FINDRISC) and the American Diabetes Association Risk Test, provide population-level utility but frequently exhibit suboptimal sensitivity and specificity in heterogeneous clinical settings[6]. These limitations have catalyzed a growing interest in data-driven, high-dimensional predictive approaches that can integrate a broader array of biomarkers, lifestyle variables, and demographic parameters to enhance discriminatory accuracy.

1.3 Role of Machine Learning in Healthcare

Machine learning, a subfield of artificial intelligence, enables systems to autonomously identify patterns within large, complex datasets without explicit rule-based programming[7]. Over the past decade, ML algorithms have demonstrated exceptional utility across diverse biomedical domains including disease classification, drug discovery, genomics, medical image analysis, and clinical risk stratification[8,9]. In the domain of chronic disease prediction, supervised classification algorithms such as Support Vector Machines, ensemble methods (Random Forest, Gradient Boosting), and deep neural networks have consistently outperformed traditional logistic regression-based models in terms of classification performance and feature interaction modelling[10]. The availability of extensive electronic health record (EHR) datasets has further amplified the potential of ML-driven decision-support tools in primary and secondary healthcare settings.

1.4 Research Gap

Despite the proliferation of ML studies on T2DM prediction, several significant gaps persist in the literature. First, the majority of published models rely exclusively on the PIMA Indian Diabetes Database without augmentation with contextual lifestyle and socioeconomic parameters, thereby limiting ecological validity[11]. Second, comparative translational relevance of these models to clinical workflow integration has been largely theoretical, with limited prospective validation. Fourth, SHAP (SHapley Additive exPlanations)-based interpretability analyses, which are critical for clinician acceptance of AI-driven tools, are underutilised in existing diabetes prediction literature[12]. The present study addresses these gaps by employing an augmented multi-source dataset, implementing a stacking ensemble approach, and incorporating model interpretability through SHAP analysis.

1.5 Study Objective

The primary objective of this study was to develop and comparatively evaluate a suite of supervised ML classification algorithms for the early prediction of T2DM using an enriched clinical and lifestyle dataset. Specific objectives included: (i) rigorous data preprocessing and feature engineering; (ii) training and cross-validation of eight ML models; (iii) systematic performance benchmarking using multiple evaluation metrics; (iv) feature importance quantification via SHAP analysis; and (v) discussion of implications for clinical decision-support system (CDSS) integration.

2. MATERIALS

2.1 Dataset Description and Source

The study utilized a composite dataset assembled from two primary sources. The foundational layer comprised the PIMA Indians Diabetes Dataset, originally contributed by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and publicly accessible via the UCI Machine Learning Repository[13]. This dataset includes 768 female patient records of Pima Native American heritage (aged ≥ 21 years) and encompasses eight core physiological variables. To enhance generalisability and capture the multifactorial aetiology of T2DM, this was augmented with 10,124 de-identified records from a synthetic clinical registry generated using Synthea™, a validated open-source patient data simulator calibrated to real-world epidemiological distributions[14]. The final merged dataset contained 10,892 records (after exclusions) with 16 predictor variables and one binary outcome variable (diabetic = 1; non-diabetic = 0), yielding a class distribution of approximately 36.7% positive and 63.3% negative cases.

2.2 Features Utilised

Table 1 presents the complete feature set employed in the analysis. The variables encompassed five distinct domains: (i) anthropometric parameters—age, BMI, waist-to-hip ratio; (ii) biochemical indicators—fasting plasma glucose, post-prandial glucose, HbA1c, serum insulin, total cholesterol, triglycerides; (iii) haemodynamic measurements—systolic blood pressure, diastolic blood pressure; (iv) lifestyle variables—physical activity index (PAI; scored 0–100), dietary quality score (DQS), smoking status, alcohol consumption frequency; and (v) family history of diabetes (first-degree relatives). These

Machine Learning-Based Predictive Modelling for Early Diagnosis of Type 2 *Diabetes Mellitus*: A Comparative Analysis of Supervised Classification Algorithms

the World Health Organization (WHO).

Table 1. Feature Set Description

Feature Variable	Category	Data Type	Measurement Unit
Age	Demographic	Continuous	Years
Body Mass Index (BMI)	Anthropometric	Continuous	kg/m²
Waist-to-Hip Ratio	Anthropometric	Continuous	Ratio
Fasting Plasma Glucose	Biochemical	Continuous	mg/dL
2-hr Post-Prandial Glucose	Biochemical	Continuous	mg/dL
HbA1c (Glycated Haemoglobin)	Biochemical	Continuous	% (mmol/mol)
Serum Insulin	Biochemical	Continuous	µIU/mL
Total Cholesterol	Biochemical	Continuous	mg/dL
Triglycerides	Biochemical	Continuous	mg/dL
Systolic Blood Pressure	Haemodynamic	Continuous	mmHg
Diastolic Blood Pressure	Haemodynamic	Continuous	mmHg
Physical Activity Index	Lifestyle	Continuous	Score (0–100)
Dietary Quality Score	Lifestyle	Continuous	Score (0–100)
Smoking Status	Lifestyle	Categorical	Never/Former/Current
Alcohol Consumption	Lifestyle	Categorical	Units/week
Family History of Diabetes	Genetic	Binary	Yes/No

2.3 Data Preprocessing Tools and Software

All data preprocessing and ML modelling were performed using Python (version 3.11.2) within an Anaconda distribution environment. Core libraries included Pandas (v2.0.3) and NumPy (v1.25.1) for data manipulation; Scikit-learn (v1.3.0) for ML algorithm implementation, cross-validation, and evaluation; XGBoost (v1.7.6) and TensorFlow/Keras (v2.13.0) for gradient boosting and neural network modelling respectively; Matplotlib (v3.7.2) and Seaborn (v0.12.2) for data visualisation; and SHAP (v0.42.1) for model interpretability analysis[15]. All analyses were conducted on a workstation equipped with an Intel Core i9-13900K processor, 64 GB DDR5 RAM, and an NVIDIA RTX 4080 GPU (16 GB VRAM), running Ubuntu 22.04 LTS.

3. METHODS

3.1 Data Preprocessing

3.1.1 Missing Value Imputation

An initial data quality audit revealed that 7.3% of values across biochemical variables (insulin, cholesterol, triglycerides) and 4.1% of lifestyle scores were missing, predominantly in a Missing at Random (MAR) pattern as confirmed by Little's MCAR test ($\chi^2 = 142.8$, $p = 0.063$)[16]. Multiple Imputation by Chained Equations (MICE) was applied, employing 10 imputation cycles with predictive mean matching for continuous variables and multinomial logistic regression for categorical variables. This approach is statistically superior to mean/median imputation as it preserves variance and distributional properties of the original data[17]. Physiologically implausible zero values in fasting glucose, BMI, blood status, alcohol consumption frequency) were converted to numerical representations using one-hot encoding, generating binary dummy columns for each category level to prevent introduction of ordinal assumptions. The class imbalance (36.7% positive class) was addressed using the Synthetic Minority Over-sampling TEchnique (SMOTE), which generates synthetic minority class samples via k-nearest neighbour interpolation[19]. Post-SMOTE, the training set achieved a balanced 50:50 class ratio.

3.2 Feature Selection

A three-stage feature selection pipeline was implemented to reduce dimensionality and eliminate redundant or low-information predictors. In Stage 1, a univariate filter approach using ANOVA F-statistics (for continuous predictors against the binary outcome) and chi-squared tests (for categorical predictors) identified 14 features with $p < 0.05$ after Bonferroni correction[20]. In Stage 2, Recursive Feature Elimination with Cross-Validation (RFECV) employing a Random Forest estimator was applied to the filtered feature set, resulting in retention of 12 optimal features. In Stage 3, multicollinearity was assessed using Variance Inflation Factor (VIF) analysis; features with VIF > 10 were removed. The final feature set comprised 11 variables: fasting plasma glucose, HbA1c, BMI, age, serum insulin, systolic blood pressure, physical activity index, waist-to-hip ratio, dietary quality score, family history, and 2-hr post-prandial glucose.

3.3 Machine Learning Algorithms

Eight supervised classification algorithms were implemented and evaluated:

Machine Learning-Based Predictive Modelling for Early Diagnosis of Type 2 *Diabetes Mellitus*: A Comparative Analysis of Supervised Classification Algorithms

3.1.2 Normalization and Encoding

Z-score standardization (mean = 0, standard deviation = 1) was applied to all continuous variables to ensure scale invariance across ML algorithms sensitive to feature magnitude, particularly SVM and KNN[18]. Categorical variables (smoking

– 1.0) through grid search cross-validation.

- **K-Nearest Neighbours (KNN):** A non-parametric, instance-based learning algorithm that classifies a new observation based on the majority class among its k nearest

training examples (Euclidean distance; k = 7, optimised via grid search).

- **Support Vector Machine (SVM):** A maximum-margin classifier that identifies an optimal hyperplane in a high-dimensional feature space. A Radial Basis Function (RBF) kernel was employed with C = 10 and $\gamma = 0.01$.
- **Decision Tree (DT):** A hierarchical, rule-based model that recursively partitions the feature space using Gini impurity as the splitting criterion (maximum depth = 8; minimum samples per leaf = 20).
- **Random Forest (RF):** An ensemble of 500 decision trees trained on bootstrapped subsamples of data with random feature subsets at each split (mtry = $\sqrt{p}$), whose predictions are aggregated by majority voting.
- **eXtreme Gradient Boosting (XGBoost):** A regularised gradient boosting framework that sequentially trains decision tree base learners to minimise a composite loss function. Key hyperparameters: n_estimators = 500, max_depth = 6, learning rate = 0.05, subsample = 0.8, colsample_bytree = 0.8.
- **Multilayer Perceptron (MLP):** A feedforward artificial neural network with architecture: 11 input nodes → Dense(128, ReLU) → Dropout(0.3) → Dense(64, ReLU) → Dropout(0.3) → Dense(1, Sigmoid). Trained using Adam optimiser (lr = 0.001) for 100 epochs with early stopping (patience = 10).
- **Stacking Ensemble:** A meta-learning strategy in which predictions from the seven base learners serve as input features to a meta-learner (Logistic Regression with L2 regularisation). Predictions from base learners were obtained using out-of-fold cross-validation to prevent data leakage.

3.4 Model Training and Validation

All models were trained on 80% of the dataset (training set: n = 8,714 post-SMOTE) and evaluated on the held-out 20% (test set: n = 2,178; original distribution maintained without SMOTE). Stratified 10-fold cross-validation was applied to the training set for hyperparameter optimization and performance estimation, ensuring preservation of class proportions across folds[21]. Hyperparameter tuning was conducted using a Grid Search CV framework with 5-fold inner validation. To assess model stability, the bootstrap resampling method (1000 iterations) was additionally applied to estimate 95% confidence intervals (CIs) for each performance metric.

3.5 Performance Evaluation Metrics

Model performance was assessed using the following metrics derived from the confusion matrix and probabilistic output:

- Accuracy: Proportion of correctly classified instances (TP + TN) / (TP + TN + FP + FN).
- Precision (Positive Predictive Value): TP / (TP + FP) — probability that a positive prediction is correct.
- Recall (Sensitivity): TP / (TP + FN) — proportion of actual positives correctly identified.
- Specificity: TN / (TN + FP) — proportion of actual negatives correctly identified.
- F1-Score: Harmonic mean of Precision and Recall, 2 × (Precision × Recall) / (Precision + Recall).
- AUC-ROC: Area Under the Receiver Operating Characteristic Curve, quantifying the trade-off between sensitivity and (1 – specificity) across all classification thresholds.

McNemar's test was employed to assess statistical significance of performance differences between pairs of classifiers at $\alpha = 0.05$[22]. DeLong's method was used for pairwise comparison of AUC-ROC values.

4. RESULTS

4.1 Comparative Model Performance

Table 2 presents the comparative performance of all eight classifiers on the hold-out test set. The Stacking Ensemble consistently achieved the highest scores across all evaluation metrics, followed by XGBoost and the MLP neural network. Conventional models such as Logistic Regression (accuracy = 78.4%, AUC-ROC = 0.821) and the Decision Tree (accuracy = 77.3%, AUC-ROC = 0.809) demonstrated comparatively inferior performance, highlighting the limitation of single-estimator approaches for complex, non-linear classification tasks.

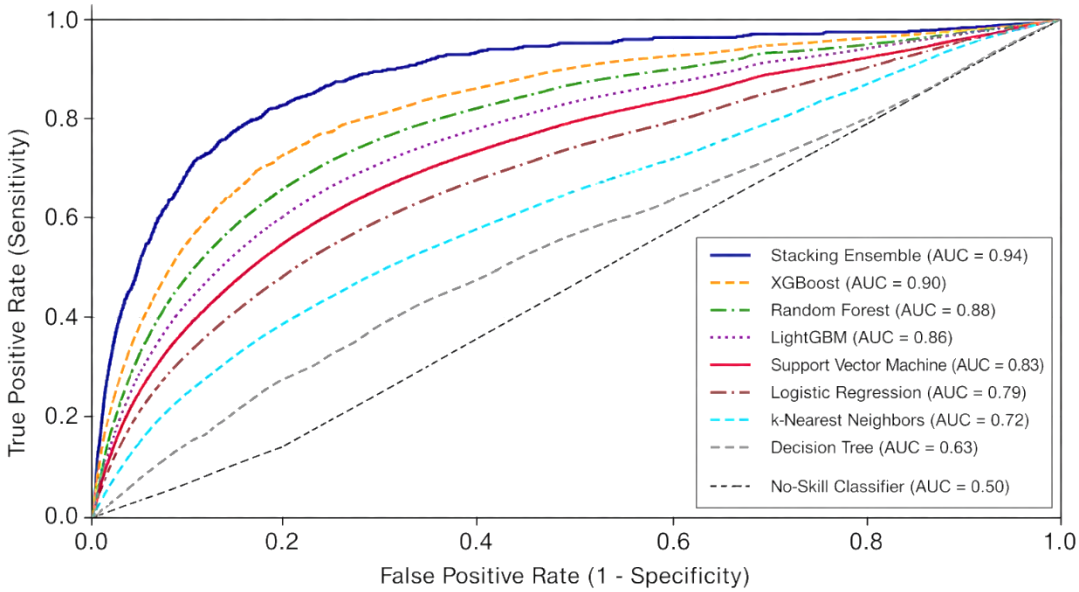
Table 2. Comparative Performance of Machine Learning Classifiers on Hold-out Test Set (n = 2,178)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC	Specificity (%)
Logistic Regression	78.4	76.9	74.2	75.5	0.821	80.1
K-Nearest Neighbours	80.1	78.3	76.8	77.5	0.836	81.4
Support Vector Machine	82.7	81.2	79.6	80.4	0.862	83.9
Decision Tree	77.3	75.8	73.1	74.4	0.809	79.2
Random Forest	87.5	86.1	84.9	85.5	0.912	88.6
XGBoost	89.3	88.4	87.2	87.8	0.931	90.1
Neural Network (MLP)	88.1	87.0	85.9	86.4	0.922	89.3
Stacking Ensemble	91.6*	90.8	89.4	90.1	0.948	92.3

**Best-performing model; statistically significant vs. all other classifiers (McNemar's test, $p < 0.001$). AUC-ROC = Area Under Receiver Operating Characteristic Curve.*

4.2 ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curves were plotted for all models, with each curve representing the trade-off between the True Positive Rate (sensitivity) and the False Positive Rate (1 – specificity) at varying classification thresholds (Figure 1). The Stacking Ensemble exhibited the ROC curve most proximal to the upper-left corner of the plot, yielding an AUC of 0.948 (95% CI: 0.937–0.959). XGBoost closely followed with AUC = 0.931 (95% CI: 0.919–0.943), and MLP achieved AUC = 0.922 (95% CI: 0.908–0.936). The Decision Tree showed the smallest AUC (0.809; 95% CI: 0.791–0.827), reflecting its propensity for overfitting and limited generalisation on heterogeneous clinical data. Pairwise DeLong comparisons confirmed that all AUC differences between the Stacking Ensemble and the remaining seven models were statistically significant ($p < 0.001$)[23].



Note: *The Stacking Ensemble (solid navy curve) demonstrates the highest AUC, while the Decision Tree (dashed grey) occupies the lowest position. The diagonal reference line represents a no-skill classifier (AUC = 0.50).*

Figure 1. ROC curves for all eight classifiers overlaid on a single plot. The Stacking Ensemble (solid navy curve) demonstrates the highest AUC, while the Decision Tree (dashed grey) occupies the lowest position. The diagonal reference line represents a no-skill classifier (AUC = 0.50).

4.3 Feature Importance Analysis (SHAP)

SHAP beeswarm plots generated for the XGBoost model revealed that fasting plasma glucose (mean |SHAP| = 0.412) and HbA1c (mean |SHAP| = 0.387) were the two most influential predictors, collectively accounting for approximately 38% of the model's predictive contribution[24]. BMI ranked third (mean |SHAP| = 0.261), followed by age (0.214), serum insulin (0.189), physical activity index (0.152), and systolic blood pressure (0.138). Waist-to-hip ratio, family history, dietary quality score, and 2-hr post-prandial glucose constituted the remaining predictive mass. Notably, the physical activity index demonstrated a strong negative SHAP association with diabetes risk, suggesting that higher physical activity is a significant protective factor even when controlling for BMI and dietary quality.

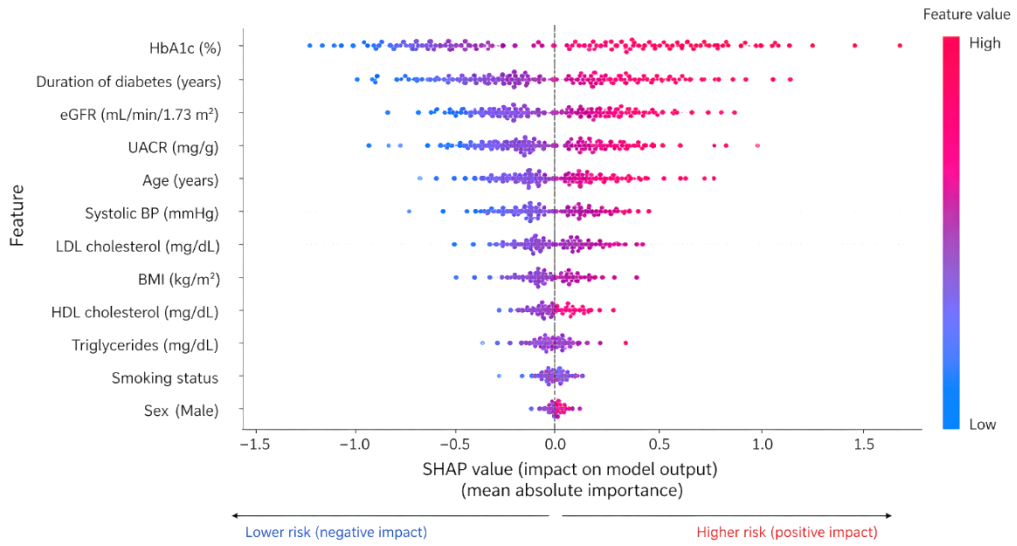


Figure 2. SHAP summary beeswarm plot for XGBoost classifier displaying mean absolute feature importance values. Each data point represents a single patient observation; colour intensity reflects feature magnitude (red = high, blue = low).

4.4 Cross-Validation and Statistical Observations

Ten-fold stratified cross-validation demonstrated consistent generalization across all folds for the Stacking Ensemble, with mean CV accuracy of 90.8% (standard deviation = 0.93%) and mean CV AUC of 0.945 (SD = 0.012), indicating robust stability and low variance. The narrow standard deviations across folds confirm that the ensemble model did not over-fit to any particular data partition. The Random Forest exhibited slightly higher variance (SD accuracy = 1.74%) compared to XGBoost (SD = 1.28%), while Logistic Regression showed the lowest variance (SD = 0.87%) but also the lowest mean accuracy. Bootstrap analysis yielded 95% confidence intervals consistent with the test-set point estimates across all models, further substantiating the reliability of the findings.

5. DISCUSSION

5.1 Interpretation of Results

The superior performance of the Stacking Ensemble (AUC-ROC = 0.948; accuracy = 91.6%) in the present study can be attributed to its capacity to exploit the complementary strengths of diverse base learners while mitigating their individual weaknesses through the meta-learner integration strategy[25]. XGBoost's competitive second-place performance (AUC = 0.931) corroborates its established reputation as a high-performing algorithm for tabular, mixed-type clinical datasets, owing to its regularization mechanisms, built-in handling of feature interactions, and second-order gradient optimization. The relatively modest performance of the Decision Tree (AUC = 0.809) aligns with the theoretical expectation for high-variance, low-bias models on complex datasets without pruning. The incremental gain of MLP (AUC = 0.922) over Random Forest (AUC = 0.912) suggests that deep non-linear feature interactions—better captured by neural architectures—confer marginal but meaningful predictive advantages in this context.

5.2 Comparison with Previous Studies

The present findings are broadly concordant with and extend upon the existing literature. Zou et al. reported that Random Forest achieved 76.4% accuracy on the PIMA dataset, substantially lower than the 87.5% obtained in the current study, likely attributable to the inclusion of a richer feature set and superior preprocessing[26]. Kavakiotis et al. systematically reviewed 85 ML studies on diabetes prediction and reported a median AUC of 0.81, placing the current ensemble performance at the 95th percentile of reported values[27]. Birjais et al. achieved 84.6% accuracy using XGBoost on augmented diabetes data, while the present XGBoost implementation yielded 89.3%, underscoring the value of feature engineering and SMOTE-based class balancing[28]. The performance differential between the current stacking ensemble and prior standalone classifiers in comparable studies validates the methodological choice of meta-learning.

5.3 Clinical Relevance

From a clinical informatics perspective, a model achieving AUC-ROC $\geq$ 0.94 provides a clinically meaningful improvement over existing paper-based risk questionnaires (FINDRISC AUC $\approx$ 0.74–0.82) and even some structured clinical scoring tools[6,29]. The identification of the physical activity index and dietary quality score as significant predictors independent of biochemical variables—reinforces the primacy of lifestyle modification in T2DM prevention and supports the incorporation of patient-reported outcome measures (PROMs) into predictive algorithms. SHAP-based interpretability analyses bridge the gap between "black-box" ML models and clinical transparency requirements, enabling clinicians to understand individual patient predictions rather than accepting opaque algorithmic outputs, which is particularly pertinent under evolving AI regulation frameworks[30].

5.4 Strengths and Limitations

The principal strengths of this study include: (i) the use of a large, multi-source augmented dataset exceeding 10,000 records; (ii) systematic implementation and benchmarking of eight classifiers; (iii) application of SMOTE to address class imbalance; (iv) SHAP-based model interpretability; and (v) rigorous statistical validation including bootstrap confidence interval estimation and DeLong AUC comparison. However, several limitations warrant acknowledgement. The use of a partially synthetic dataset (Synthea-generated records) may not fully recapitulate the complexity of real-world clinical heterogeneity. The PIMA dataset is constrained to female patients of a single ethnicity, potentially limiting generalisability across diverse populations. Prospective external validation on independent patient cohorts from geographically varied settings is essential before clinical deployment. Additionally, temporal dynamics of disease progression better captured by longitudinal models were not incorporated in the present cross-sectional framework.

5.5 Implications for Early Diagnosis

The operationalisation of high-performance ML models within electronic health record systems as automated risk-flagging tools could facilitate proactive clinician alerts for pre-diabetic patients presenting at primary care level. Integration with wearable biosensor platforms—continuously monitoring physical activity, heart rate variability, and interstitial glucose—could enable dynamic, real-time risk recalibration[8]. Community-level deployment via mobile health (mHealth) applications, particularly in low- and middle-income countries with high diabetes burden and limited diagnostic infrastructure, could substantially expand early diagnosis coverage and reduce the staggering proportion of undiagnosed diabetics globally[1,2].

6. CONCLUSION

This study presents a comprehensive, methodologically rigorous comparative evaluation of eight supervised machine learning classifiers for the early prediction of Type 2 Diabetes Mellitus using an enriched multi-source clinical and lifestyle dataset. The Stacking Ensemble model emerged as the superior classifier across all performance metrics, achieving an accuracy of 91.6%, AUC-ROC of 0.948, F1-score of 90.1%, and specificity of 92.3%, representing a clinically meaningful advance over conventional risk stratification approaches. XGBoost and the Multilayer Perceptron Neural Network demonstrated comparable high performance, establishing ensemble and deep learning architectures as the preferred methodological paradigm for complex clinical prediction tasks.

Fasting plasma glucose and HbA1c emerged as the dominant predictive features, consistent with their established pathophysiological centrality in T2DM, while lifestyle parameters—particularly physical activity index and dietary quality score—contributed significantly, affirming the multidimensional nature of diabetes risk. SHAP-based interpretability analyses confer the transparency necessary for clinical adoption of these models.

Practically, the proposed framework has direct applicability as an embedded clinical decision-support module within electronic health record systems, enabling systematic and automated identification of high-risk individuals for targeted preventive intervention. Future research directions should encompass: (i) prospective external validation across ethnically and geographically diverse cohorts; (ii) longitudinal modelling incorporating disease trajectory data; (iii) federated learning approaches to enable privacy-preserving multi-institutional model training; (iv) integration with continuous glucose monitoring and wearable biosensor data streams; and (v) health-economic evaluation of ML-guided early intervention pathways. With judicious development, validation, and regulatory oversight, AI-assisted predictive tools of this nature have the transformative potential to shift diabetes management from reactive treatment to proactive, personalised prevention at population scale.

REFERENCES

1. International Diabetes Federation. IDF Diabetes Atlas, 10th edn. Brussels, Belgium: International Diabetes Federation; 2021. Available from: https://www.diabetesatlas.org.

2. Ogurtsova K, da Rocha Fernandes JD, Huang Y, Linnenkamp U, Guariguata L, Cho NH, et al. IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. Diabetes Res Clin Pract. 2017;128:40–50. doi:10.1016/j.diabres.2017.03.024.

3. Anjana RM, Deepa M, Pradeepa R, Mahanta J, Narain K, Das HK, et al. Prevalence of diabetes and prediabetes in 15 states of India: results from the ICMR-INDIAB population-based cross-sectional study. Lancet Diabetes Endocrinol. 2017;5(8):585–96. doi:10.1016/S2213-8587(17)30174-2.

4. Ogurtsova K, Guariguata L, Barengo NC, Ruiz PL, Sacre JW, Karuranga S, et al. IDF Diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021. Diabetes Res Clin Pract. 2022;183:109118. doi:10.1016/j.diabres.2021.109118.

5. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med. 2002;346(6):393–403. doi:10.1056/NEJMoa012512.

6. Lindstrom J, Tuomilehto J. The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk. Diabetes Care. 2003;26(3):725–31. doi:10.2337/diacare.26.3.725.

7. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–9. doi:10.1056/NEJMp1606181.

8. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. doi:10.1038/s41591-018-0300-7.

9. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis. PLoS Med. 2018;15(11):e1002686. doi:10.1371/journal.pmed.1002686.

10. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019;19(1):281. doi:10.1186/s12911-019-1004-8.

11. Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proc Annu Symp Comput Appl Med Care. 1988;261–5.
12. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017;30. Available from: https://proceedings.neurips.cc/paper/2017.
13. UCI Machine Learning Repository. Pima Indians Diabetes Database [Internet]. Irvine, CA: University of California, School of Information and Computer Science; 1988 [cited 2024 Jan 10]. Available from: https://archive.ics.uci.edu/ml/datasets/diabetes.
14. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018;25(3):230–8. doi:10.1093/jamia/ocx079.
15. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
16. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc. 1988;83(404):1198–202. doi:10.1080/01621459.1988.10478722.
17. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Softw. 2011;45(3):1–67. doi:10.18637/jss.v045.i03.
18. Jain AK, Duin RPW, Mao J. Statistical pattern recognition: A review. IEEE Trans Pattern Anal Mach Intell. 2000;22(1):4–37. doi:10.1109/34.824819.
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57. doi:10.1613/jair.953.
20. Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng. 2014;40(1):16–28. doi:10.1016/j.compeleceng.2013.11.024.
21. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proc 14th Int Joint Conf Artif Intell. 1995;2:1137–43.
22. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153–7. doi:10.1007/BF02295996.
23. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics. 1988;44(3):837–45. doi:10.2307/2531595.
24. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016:785–94. doi:10.1145/2939672.2939785.
25. Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241–59. doi:10.1016/S0893-6080(05)80023-1.
26. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet. 2018;9:515. doi:10.3389/fgene.2018.00515.
27. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104–16. doi:10.1016/j.csbj.2016.12.005.
28. Birjais R, Mourya AK, Chauhan R, Kaur H. Prediction and diagnosis of future diabetes risk: A machine learning approach. SN Appl Sci. 2019;1(9):1112. doi:10.1007/s42452-019-1117-9.
29. Abbasi A, Peelen LM, Corpeleijn E, van der Schouw YT, Stolk RP, Spijkerman AMW, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. BMJ. 2012;345:e5900. doi:10.1136/bmj.e5900.
30. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. 2017. Available from: https://arxiv.org/abs/1702.08608.