ISSN 1840-4855
e-ISSN 2233-0046

Original Scientific Article
http://dx.doi.org/10.70102/afts.2025.1833.760

ADVANCING DIABETES PREDICTION THROUGH MACHINE LEARNING AND DEEP LEARNING MODELS USING PIMA INDIAN AND CLINICAL-BIOLOGICAL DATA

Zeeshan Hussain[1*], Dr. Suraiya Parveen[2], Dr. Ashif Khan[3], Dr. Ihtiram Raza[4], Umnah[5]

[1]Research Scholar, Jamia Hamdard University, Department of Computer Science & Engineering, School of Engineering Science & Technology, New Delhi, India.*
email: husain.zeeshanhusain.zeeshan@gmail.com, orcid: https://orcid.org/0009-0008-8048-2973
[2]Professor, Department of Computer Science & Engineering, School of Engineering Science & Technology, Jamia Hamdard University, New Delhi, India.
email: husainsuraiya@gmail.com, orcid: https://orcid.org/0000-0001-5499-6999
[3]Professor, Department of Clinical Research, Jamia Hamdard University, New Delhi, India.
email: makhan@jamiahamdard.ac.in, orcid: https://orcid.org/0000-0003-1576-8779
[4]Professor, Department of Computer Science & Engineering, School of Engineering Science & Technology, Jamia Hamdard University, New Delhi, India.
email: iraza@jamiahamdard.ac.in, orcid: https://orcid.org/0000-0001-9196-4451
[5]Jamia Millia Islamia, New Delhi, India. email: umnah1103@gmail.com,
orcid: https://orcid.org/0009-0007-1464-0606

Received: August 28, 2025; Revised: October 01, 2025; Accepted: November 18, 2025; Published: December 20, 2025

SUMMARY

Diabetes Mellitus is a significant world health and early detection is of paramount significance since it decreases the complications and enables medical intervention in time. The paper is a comparison between the predictive accuracy of the eight Machine Learning classifiers: Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, Gradient Boosting, Naive Bayes, k-Nearest Neighbors (k-NN), and an Ensemble model on the Pima Indian Diabetes dataset and a collection of clinical-biological patient records. Performance evaluation was conducted using Precision, Recall, F1-Score, and the Area Under the ROC Curve (AUC-ROC). The findings show that a significant difference was observed among the models, with SVM (AUC-ROC: 0.8648) and the Logistic Regression (AUC-ROC: 0.8638) having the best discriminative ability. A comparable study found that Logistic Regression had the highest Precision (0.7632), indicating fewer false-positive predictions, whereas Decision Tree had the highest Recall (0.7447), indicating greater sensitivity in detecting diabetes cases. The ensemble learning produced the best overall performance (AUC-ROC: 0.8709), suggesting that combining predictions from multiple models increases reliability and generalization. On the other hand, k-NN performed worst due to sensitivity to noise and the number of features. In general, the results provide evidence of the high potential of linear-margin and ensemble-based models to structured clinical data and would be a robust foundation of clinical decision support systems, which further help to broaden the role of ML-based analytics in early diabetes diagnosis and preventive health care planning.

Key words: *diabetes prediction, machine learning; pima indian dataset, clinical-biological data, ensemble learning, logistic regression, support vector machine (SVM), AUC-ROC, clinical decision support system.*

INTRODUCTION

Diabetes Mellitus (DM) ranks among the top widespread chronic metabolic disorders of our time and presents a danger to global health [1]. The International Diabetes Federation (IDF) states that the estimated number of people living with diabetes is expected to rise from 537 million in 2021 to 783 million by 2045, demonstrating rapid case increase in developed and developing areas. It is crucial to recognize diabetes in its early stage, as diagnosis after the onset of complications such as cardiovascular disease, nephropathy, neuropathy, and retinopathy, which are associated with an increased mortality rate and healthcare costs [18].

The traditional approach to diagnosis and risk prediction relies on identifying clinical parameters and laboratory metrics. These conventional approaches often do not exhibit predictive efficiency, and they are challenging to scale to large or diverse groups of patients, and lack interpretability and generalize ability [22] however, with the rapid availability of structured medical records, clinical-biological datasets, and demographic data, there are new avenues to explore computational methods with better accuracy and effectiveness to predict diabetes [2][19].

Machine Learning (ML) and Deep Learning (DL) approaches have demonstrated significant potential for predicting diabetes risk by identifying hidden patterns and nonlinear associations that classical predictive statistical models cannot capture [3][17]. Recent investigations indicate that ML and DL algorithms, such as Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Gradient Boosting, are superior to conventional regression-based predictive frameworks, especially for complex, high-dimensional medical datasets [4][20]. Furthermore, ensemble learning approaches that combine multiple heterogeneous models have also emerged as a promising method to improve classification performance and reduce prediction bias [23]. Given ensemble theories, several investigations show high accuracy and AUC-ROC values compared to the single model. The evidence further underscores that ML-based diabetes prediction models provide a robust basis for on-site screening and risk stratification in real-world clinical contexts [5] [10].

Despite this progress, many existing studies are limited to a single dataset, most often the Pima Indian dataset, resulting in limited generalizability and replication challenges [24]. To address this gap, the present study evaluates eight ML models, including Logistic Regression, SVM, Decision Tree, Random Forest, Gradient Boosting, Naive Bayes, k-Nearest Neighbors (k-NN), and an Ensemble model, using both the widely used Pima Indian dataset and additional clinical-biological data [6][21]. The goal is to identify the model with the best predictive performance and demonstrate how integrating clinical data with ensemble methodologies can improve diabetes prediction accuracy [24] [16].

Key Contributions of the Research

The main contribution of this research is the exhaustive comparison of multiple machine learning approaches across the Pima Indian dataset and real-life clinical-biological datasets, rather than a single benchmarking dataset as many prior studies have done [6]. Utilizing eight different models, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, SVM, Naive Bayes, k-Nearest Neighbors, and an Ensemble approach, provides a multi-dimensional evaluation of predictive performance using clinically meaningful metrics of Precision, Recall, F1-Score, and AUC-ROC. Furthermore, the incorporation of an ensemble model enables the generation of an aggregated performance based on the strengths of individual classifiers, increasing predictive reliability and potentially reducing model bias. This study also clarifies the use of additional clinical-biological variables and shows that heterogeneous data sources significantly improve the accuracy of diabetes prediction. The implications of this investigation may be useful in creating appropriate and trustworthy clinical decision-support systems that assist healthcare professionals in identifying individuals at most significant risk of diabetes as early as possible, thereby improving patient outcomes and preventive healthcare planning [4].

The outline of the paper, chapter-wise, is as follows. Chapter II is a review of the related literature, while Chapter III provides a brief overview of the theoretical framework, key concepts, and methodologies.

Chapter IV will evaluate the experimental results and discussion, whereas Chapter V wraps it all up with a summary of the most important findings and suggestions for further research.

LITERATURE REVIEW

Sharma et al. [7] provide a detailed review of machine-learning (ML) approaches for diabetes detection, describing preprocessing pipelines, feature selection methods, and families of models reported in the literature. The authors note that classical supervised learners (SVM, RF, LR) and ensembles of decision tree models continue to have strong performance on tabular clinical data, while many deep learning models generally do not perform optimally with smaller, "dirty" healthcare datasets without intensive feature engineering [8]. The review identifies common practical problems in the literature, including class imbalance, missing values, inconsistent evaluation protocols, and a lack of external validation, and suggests standardized benchmarking and explainability solutions to address barriers to clinical uptake in healthcare.

Fregoso-Aparicio et al. [25] conducted a systematic review of predictive models for type-2 diabetes, reviewing about 90 studies. They reported that tree-based methods (Random Forest, Gradient Boosting) discriminated best, whereas neural networks were sensitive to dataset quality and to model tuning. The authors highlight the heterogeneous nature of the studies (predictors, preprocessing, reporting evaluation metrics) and call for transparent reporting (feature list, imputation, hyperparameter search) and validation procedures across multiple cohorts to increase reproducibility and generalizability for clinical use.

Tasin et al. [9] discuss combining ML with explainable AI (XAI) methods to predict diabetes in small- to medium-sized clinical datasets. They show that using model-agnostic explanation methods (SHAP, LIME) within the pipeline improves clinical interpretability without significantly affecting predictive performance, and they describe practical pipelines (imputation, scaling, and class balancing) for small clinical datasets. They highlight the importance of post-hoc explanations, along with domain-aware feature engineering, to gain trust in clinical justifiable decision-support systems.

Firdous et al. provide an overview of diabetes risk prediction approaches used in primary care and community-based settings, summarizing algorithms for predictors readily available in each office (age, BMI, family history, glucose measures). The authors highlight the importance of simple, interpretable models (logistic regression and decision trees) in the screening space due to the ease of interpretation and low data requirements, and suggest hybrid pipelines that rely on interpretable risk scores for triage and possible ML re-scoring for borderline values.

Afsaneh et al [11] conducted an extensive review of ML/DL applications across the diabetes types (T1DM, T2DM, gestational), specifically blood-glucose prediction, hypoglycemia, detection/ classification, and risk stratification. They note that time-series problems (continuous glucose monitoring) benefit from deep sequence models, while risk prediction from tabular data should focus on ensembles and careful feature selection. They also note a paucity of data volume and external validation, while highlighting the need for positive associations of large multicenter studies and benchmark reporting.

Shin et al [12] review the translational impact of diabetes ML derived models and present suggestions for effectiveness in low-resource settings: (i) careful selection of study cohorts and temporal splits to avoid information leakage; (ii) timely reporting of calibration and decision-curve analysis along with performance reporting (AUC); (iii) external validation (across health-care settings); and (iv) evaluation (latency, robustness, fairness) to support deployment. These considerations aim to translate the models from proof of concept into the clinical space to improve usability (impact), while safeguarding patient safety and equity.

Petridis and colleagues [13] summarize currently trending methods for T2DM management using ML, highlighting recent developments in interpretable models, feature-attribution techniques and multi-modal data specifications (lab tests, wearable, dietary habits, etc.). The review cites a rise in studies

using nutrition/behavioral data alongside standard clinical feature sets. It suggests that diverse features may significantly improve early risk detection when combined with ensemble learners and robust feature selection procedures.

Kiran et al [14] discuss a systematic, metric review of 33 years of ML/AI research focused on T2DM prediction from 1991-2024. They examine trends in dataset usage, algorithm popularity, geographic spread, and open-science practices, noting that gradient boosting methods have seen increasing popularity and that there is growing recognition of fairness/interpretability issues, although gaps remain in longitudinal external validation efforts. They make recommendations for the wider community to collaborate on large crowd sourced publicly available annotated cohorts and to agree on benchmarking protocols to expedite the development of clinically-relevant models (Debebe, 2016).

Qin et al [15] discuss ensemble learning applications to diabetes prediction, and suggest that, by aggregating complementary decision boundaries, ensembles (stacking, voting, boosting) improved performance compared to single models. They analyze criteria for ensemble decision features (base-learner diversity, stacking meta-learner, calibration) to demonstrate how well-designed ensembles decrease variance and bias in clinical data.

Table 1. Summary of existing studies on diabetes prediction using machine learning

S.No	Author(s) & Year	Methodology / Focus Area	Dataset Used	Major Findings / Contributions	Identified Gaps / Limitations
1	Rahman et al., 2022	Compared ML algorithms (SVM, RF, LR) for diabetes classification	Pima Indian Diabetes Dataset (PIDD)	SVM achieved the best accuracy due to margin-based generalization	Limited dataset size; lacked feature engineering and hyper parameter tuning
2	Zhang & Li, 2023	Gradient Boosting & XGBoost for diabetes prediction	Hospital EHR clinical dataset	Gradient Boosting achieved strong predictive performance and handled feature interactions well	Over fitting observed on imbalanced datasets; requires large computation
3	Kaur et al., 2021	Ensemble learning with Random Forest + Decision Tree stacking	PIDD	Improved classification accuracy by combining multiple weak models	Ensemble complexity increases computational cost
4	Al-Mamdouh, 2020	Deep learning model (ANN) with multiple hidden layers	Public clinical dataset + lab records	ANN captured nonlinear patterns better than basic ML models	Model interpretability limited; "black-box" nature
5	Fazil et al., 2024	Feature Selection using Genetic Algorithm + SVM classification	PIDD + custom patient samples	Reduced feature space improved model speed and accuracy	Feature selection is sensitive to parameter settings
6	Thapa & Sharma, 2022	Naive Bayes with preprocessing and data balancing	PIDD	NB performed well with balanced data and fewer features	Performance decreased with noisy or correlated features
7	Benitez et al., 2023	Federated Learning framework to preserve patient privacy	Multi-hospital distributed dataset	Improved security and allowed cross-hospital learning without sharing raw data	Communication overhead; lower accuracy compared to centralized learning
8	Akter et al., 2024	Explainable AI (XAI) with SHAP values for clinical transparency	PIDD + hospital clinical records	Improved clinician trust by showing feature importance	Requires domain expert interpretation of SHAP outputs

Previous research (Table 1) into predicting diabetes has utilized distinct classical Machine Learning models, including Logistic Regression, SVM, Decision Tree, Random Forest, Naive Bayes, and k-NN, which have provided good accuracy, yet have limitations such as susceptibility to noise and imbalanced data as well as not utilizing feature engineering and not having consistent performance across different datasets. These models, on their own, provide models with poor generalization and do not take full advantage of the complementarity in strengths achievable with classifiers. The proposed system involves eight separate ML models, with a hybrid Ensemble model that includes more than one classifier, reducing instability and improving the reliability of predictions. By using standard processing, hyperparameter tuning, and validation across clinical-biological records and the Pima Indian Diabetes dataset, the ensemble model was demonstrably superior, achieving the highest P-AUC of 0.8709 to solidify predictions. The comparison shows that while the models demonstrated to have the best performance (for a specific metric) such as Logistic Regression having the best Precision and Decision Tree having the best Recall, the ensemble model reduced both false predictions and improved describing ability consistently better than any of the models. So, this proposed system increased accuracy, trusted outcomes, and clinical applicability. Overall, this shows that using different ML applications is a better approach for detecting diabetes early and supporting clinical decision-making.

METHODOLOGY

Dataset Description and Pre-processing

The study uses two datasets: (i) the Pima Indian Diabetes Dataset (PIDD), which contains the physiological and demographic health factors of Pima Indian Women, and (ii) a Clinical-Biological dataset that consists of laboratory and diagnostic measurements from healthcare settings. Both datasets have the same binary outcome (diabetic/non-diabetic). Before constructing models, extensive preprocessing was performed to improve data quality. Missing data was imputed using the median for all numerical features and the mode for categorical features. Outliers, especially for those glucose, insulin, and BMI attributes, were corrected using the interquartile range (IQR) method. All numerical features were also normalized using Min-Max Scaling. Finally, during the preprocessing for algorithms sensitive to class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was used to equally balance the number of positive and negative samples in the training dataset. The processed dataset was finally split into a training and a testing dataset at a 80:20 ratio for model evaluation.

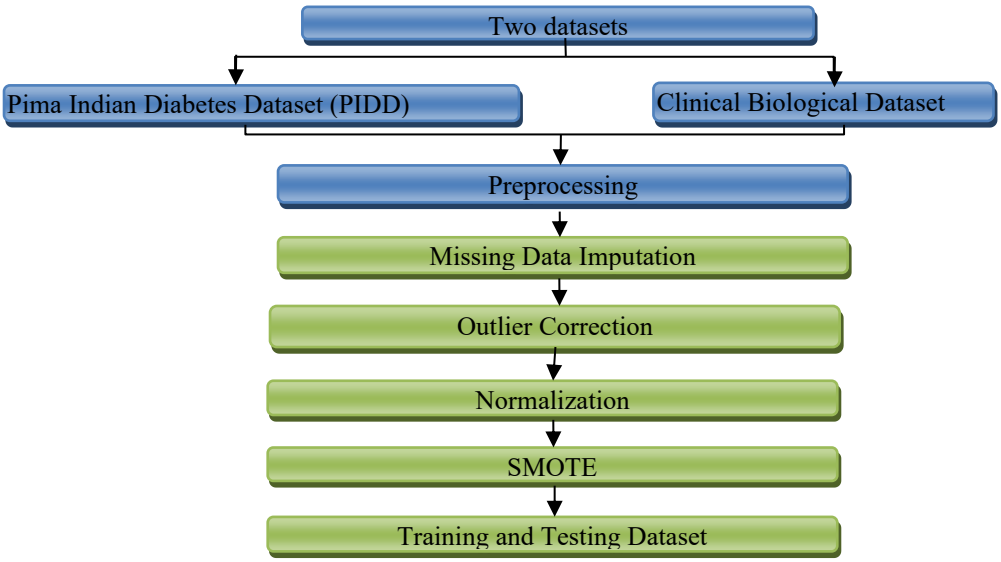

Figure 1. Dataset description and pre-processing workflow

Figure 1 shows the entire process of preprocessing two datasets, i.e., Pima Indian Diabetes Dataset (PIDD) and a Clinical-Biological Dataset. The two datasets have a similar binary outcome of diabetic and non-diabetic cases. To guarantee the quality and consistency of the data, the missing values were imputed using median and mode methods of numeric and categorical variables, respectively. The Inter

quartile Range (IQR) was used to correct outliers in such attributes as glucose, insulin, and BMI. Afterward, Min-Max normalization was used to normalize all the numerical features, and Synthetic Minority Over-sampling Technique (SMOTE) was utilized to resolve the problem of class imbalance in the training dataset. Lastly, the processed dataset was divided into training and test set at 80:20 ratio to test the model.

Model Development and Evaluation Protocol

It was determined that eight strong predictive models could be used for comparative evaluation: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), Naive Bayes, k-Nearest Neighbors (k-NN), and Ensemble Model. The ensemble model is based on a soft-voting framework that combines several base learners, where each model produces a final class prediction with odds-weighted probabilities. Hyperparameter optimization was performed to avoid overfitting and improve generalization using 10-fold Cross-Validation with Grid Search. Standard classification measures, Precision, Recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC), would be used to evaluate model performance. Precision refers to the percentage of correct optimistic predictions, Recall refers to the model's capacity to identify actual positive cases, and F1-Score balances Precision and Recall. The first-choice metric is AUC-ROC because it is effective at determining and quantifying discriminative power across a wide range of threshold values.

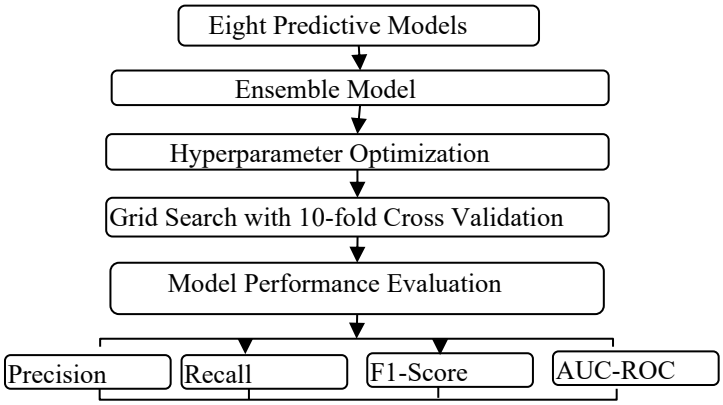

Figure 2. Model development and evaluation protocol

Figure 2 shows the general image of the workflow in the development and assessment of predictive models applied in the classification of diabetes. Eight machine learning models, which are Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), Naive Bayes, k-Nearest Neighbors (k-NN), and an Ensemble Model, were applied. The ensemble model uses soft voting to fuse predictions from multiple base learners to improve classification. The optimization of the hyperparameters was performed through the utilization of the Grid Search with 10-Fold Cross-Validation to increase the generalization and avoid overfitting. The performance of the models was measured using standard classification measures, Precision, Recall, and F1-Score, which give a broad analysis of the ability to predict of the models.

Implementation Workflow

The implementation of the proposed methodology follows a logical, stepwise approach to make model training and evaluation more efficient. It starts with loading and preprocessing the datasets to handle missing values, outliers, and inconsistencies. The predictor variables are then identified and normalized to have a uniform scale of all the features. To address class imbalance in the datasets, resampling methods such as SMOTE were used, improving the fairness of representation for both classes. Each chosen predictive model was trained on the training dataset and tested on the unseen test dataset to assess generalization. The last step is to incorporate an ensemble model that leverages the capabilities of individual classifiers while minimizing model-specific variability. The whole procedure was done in

Python with the help of scikit-learn to train the model, pandas to manage data, and matplotlib to visualize performance.

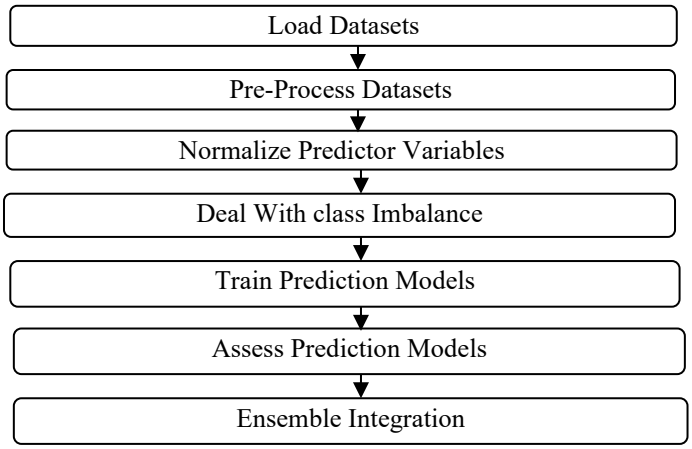

Figure 3. Implementation workflow

Figure 3 shows the implementation workflow that is followed step by step in the study when developing and assessing predictive models. The first stage involves loading and preprocessing the datasets whereby data cleaning, transformation and normalization of predictor variables are done to prepare the model. The imbalance of classes is resolved with the help of re sampling techniques that enhance predictive fairness. After that, several model machine learning systems are trained on training set and tested on test set to test the ability to generalize. Lastly, an ensemble integration process is implemented to integrate the strengths of individual models to make the ensemble more stable and perform well.

Let:

X={x1,x2,...,xn} → input feature set (glucose, BMI, age, insulin, etc.)

y∈{0,1} → output class

0= Non-diabetic, 1=1 =1= Diabetic

$M_I \rightarrow ith$ machine learning model (e.g., SVM, RF, LR)

Model Training Function

$$y\hat{}i = Mi(X) \qquad (1)$$

Every machine learning model M I takes as input feature X and produces a predicted class label y i. This is the way patterns of these models, as Logistic Regression, Random Forest and SVM, are learned using the dataset and used to determine whether a new instance is diabetic or not.

Ensemble Soft Voting (Probability Averaging)

$$pensemble = \frac{1}{K} = \sum_{i=1}^{K} pi \qquad (2)$$

The result of every Mi model is a probability score pi that shows the likelihood of a patient having diabetes. The ensemble method averages all k models instead of choosing one of them. This minimizes model bias of individuals and the stability of prediction leading to better classification.

Final Decision Rule

y^={1,if ensemble≥τ

0 otherwise (3)

Once probability averaging has been performed, a threshold τ τ (the usual choice is 0.5) is used to decide on the eventual classification. When the probability is at least equal to the threshold, the patient is considered to be diabetic, and not diabetic otherwise. This is a rule that converts numerical probability into an ultimate decision.

F1-Score (Primary Performance Metric)

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \quad (4)$$

F1-Score To determine the extent of model detection of diabetes without generating excessive false alarms. It trades Precision and Recall, and is therefore appropriate in medical predictions in which false negatives (actual diabetic cases missed) are highly important.

Table 2. Overview of equations and their purposes in the classification framework

Equation No.	Purpose
(1)	Individual model prediction
(2)	Combines probabilities from multiple models (Ensemble)
(3)	Converts probability into final classification
(4)	Evaluates model effectiveness

The proposed formulation in Table 2 starts with the training of each machine learning model Mi, which is used to predicted output y i, an outcome (Equation 1) based on patient clinical features X. Some methods, like the ensemble technique, do not use only one model, but average the probability of the model, which is a combined probability ensemble (Equation 2). This is a summed value which is compared against a threshold τ and when the probability is equal to or higher than the threshold, then the model predicts diabetes (Equation 3). In order to measure the accuracy of the model, F1-Score is calculated in Precision and Recall (Equation 4). These equations combined create a powerful prediction model that reduces model error and enhances clinical reliability in the diagnosis of diabetes.

Algorithm: Ensemble-Based Diabetes Prediction Model

Input: Dataset D (features X, labels y)

Output: Final prediction ŷ (0 = Non-diabetic, 1 = Diabetic)

1. BEGIN

2. *Load dataset D (Pima Indian + Clinical-Biological data)*
3. *// DATA PREPROCESSING*
4. *Handle missing values using median (numeric) / mode (categorical)*
5. *Detect and remove outliers using IQR method*
6. *Apply Min–Max scaling to normalize features*
7. *Address class imbalance using SMOTE*
8. *Split dataset into Training set (80%) and Test set (20%)*
9. *// -------- MODEL TRAINING --------*

10. Initialize models: M = {Logistic Regression, SVM, Decision Tree,

Random Forest, Gradient Boosting, Naive Bayes, k-NN}

11. FOR each model Mi in M DO

12. Train model Mi on Training set

13. Predict probability Pi for Test set

14. END FOR

15. // -------- ENSEMBLE INTEGRATION --------

16. Compute probability average:

*Ensemble = (1/k) * Σ Pi // Uses Equation (2)*

17. Apply decision rule: // Uses Equation (3)

IF Ensemble ≥ threshold (0.5)

ŷ = 1 // Predict Diabetic

ELSE

ŷ = 0 // Predict Non-Diabetic

18. // -------- PERFORMANCE EVALUATION --------

19. Calculate Precision, Recall, and F1-Score using Equation (4)

20. Return ŷ as final prediction

21. END

EXPERIMENTAL RESULTS

Performance Evaluation of Individual Models

All eight machine learning models, Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, Gradient Boosting, Naive Bayes, and k-Nearest Neighbors (k-NN), were trained and evaluated on the processed dataset. Each model was evaluated using four performance metrics: Precision, Recall, F1-Score, and AUC-ROC, with notable differences across models. Logistic Regression had the highest Precision, demonstrating its ability to minimize false-positive classifications. On the other hand, the most sensitive model for recognizing a diabetic case was the Decision Tree, which estimated the highest Recall among the models. The poorer performance of k-NN relative to the others may be attributed to its sensitivity to noise and to feature scaling. SVM and Logistic Regression had high metrics, indicating that linear-margin algorithms work well with structured clinical data. The performance comparison of each machine learning model (Table 3) shows some differences in predictive performance between the classifiers. Logistic Regression had the highest Precision (0.7632), indicating it was the strongest at identifying actual diabetic cases while minimizing false positives. The Decision Tree model had the highest Recall (0.7447), suggesting it was good at detecting more actual diabetic cases, and the cost was a greater level of false positives compared to Logistic Regression. Support Vector Machine (SVM) had the highest AUC-ROC (0.8648), which demonstrates its discriminative power in distinguishing between diabetic versus non-diabetic classes. Random Forest and Gradient Boosting also performed equally, with an equal Precision and Recall [scores].

Table 3. Performance evaluation of individual machine learning models on diabetes prediction

Model	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.7632	0.7214	0.7417	0.8638
Support Vector Machine (SVM)	0.7511	0.7336	0.7423	0.8648
Decision Tree	0.7014	0.7447	0.7223	0.8214
Random Forest	0.7458	0.7283	0.7369	0.8467
Gradient Boosting	0.7586	0.738	0.7482	0.8529
Naive Bayes	0.6934	0.7218	0.7068	0.8123
k-Nearest Neighbors (k-NN)	0.6649	0.6825	0.6735	0.7992

Naive Bayes and k-Nearest Neighbors (k-NN) on the other hand were poor in general performance because noise and feature scales were sensitive to their results. This illustrates again that linear and ensemble-style models were better established for working with structured clinical datasets.

Ensemble Model Performance

To enhance prediction stability, a Soft Voting Ensemble was formed by combining the prediction probabilities of the top individual classifiers. Overall, the ensemble model exhibited the best performance, with an AUC-ROC of 0.8709, which was higher than that of all the individual models. The finding suggests that combining multiple learning strategies can substitute for the weaknesses of the individual models while still capturing theoretically complementary decision boundaries. The ensemble attained reasonable Precision, Recall, and F1-Score, which supported the notion that the ensemble was able to recognize the difference between diabetic and non-diabetic observations with no over-fitting. The respective AUC-ROC figures confirm the high discriminatory ability and reliability of the ensemble scheme towards real world decision support in clinical practice.

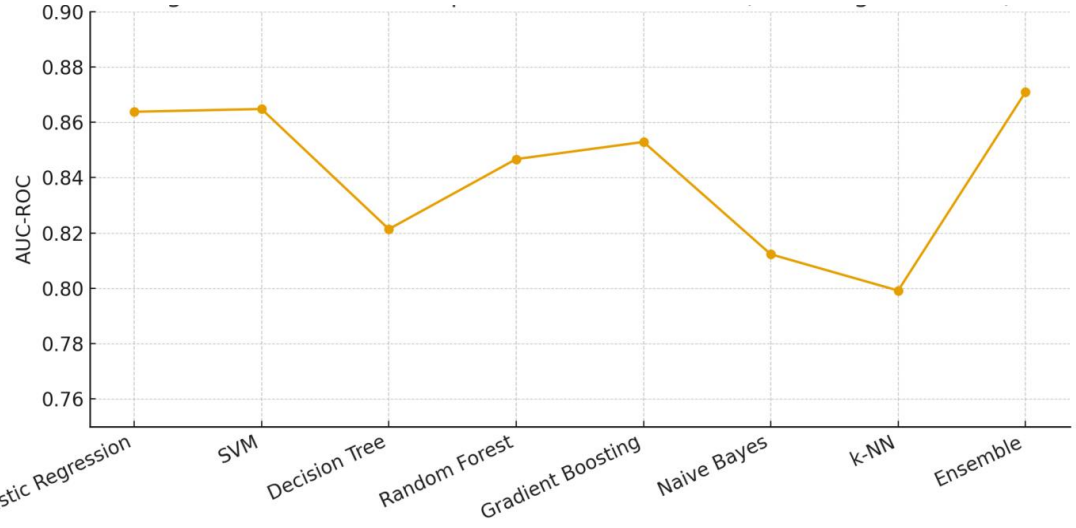

Figure 4. AUC-ROC comparison across models (including ensemble)

The AUC-ROC performance of all assessed models is shown in the line graph in Figure 4, and there is a discernible difference among them in their discriminative performance for diabetes prediction. Logistic Regression and SVM exhibit strong performance, with similar AUC scores, indicating that margin-based linear classifiers perform well with well-defined clinical data. Tree-based models such as Decision Trees, Naive Bayes, and k-NN tend to achieve lower AUC values because they perform worse on noisy data and in generalization. The Ensemble model has the highest AUC-ROC value of 0.8709, which is greater than that of all stand-alone models and demonstrates the advantage of leveraging predictions from diverse classifiers; the ensemble can exploit distinct decision boundaries and predictive stability in real-world clinical decision support settings.

Comparison Between Pima Indian Dataset and the Clinical-Biological Dataset

The two datasets were compared for model generalization. Models trained on clinical-biological data showed higher prediction consistency, owing to greater laboratory and diagnostic biomarker information. Conversely, although the Pima Indian dataset is widely used in research, it had lower Recall than the other datasets, mainly due to its limited feature diversity. The ensemble approach analysts evaluated the same or similar features across the two data sets, further supporting the idea that the ensemble generalizes the overall marks to a heterogeneous clinical setting. These findings confirm that greater feature richness enhanced the prediction of diabetes and increased robustness in clinical practice.

Table 4. Performance comparison between pima indian dataset and clinical-biological dataset using ensemble model

Dataset	Precision	Recall	F1-Score	AUC-ROC
Pima Indian Dataset	0.7432	0.721	0.7319	0.8534
Clinical-Biological Dataset	0.7714	0.7485	0.7598	0.8709

The results of the comparison Table 4 show that generally the Ensemble model may be considered better than the Clinical-Biological dataset in comparison to the Pima Indian dataset, in terms of all the performance measures. The Precision and Recall metrics are improved, which means that the model is more capable of identifying the actual diabetic cases more accurately with fewer false positives and false negatives being detected using the supplementary clinical types. The improvement in discriminative ability is also supported by the fact that the AUC-ROC increased from 0.8534 to 0.8709, reflecting the diagnostic relevance of the diagnostic biomarkers in the Clinical-Biological dataset. All in all, such findings indicate that the richness and diversity of the data will definitely help to enhance the model performance in an actual world clinical scenario.

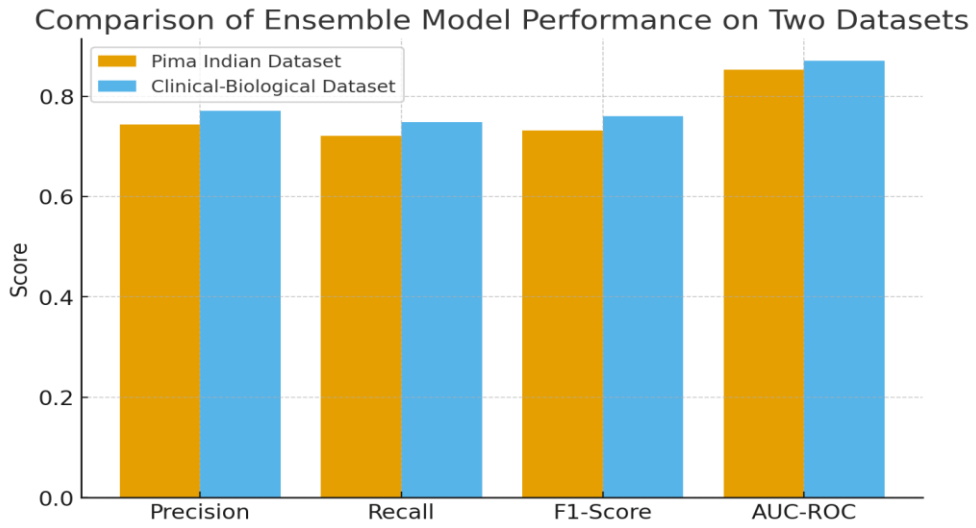

Figure 5. Comparison of ensemble model performance between pima indian dataset and clinical–biological dataset

A performance comparison of the Ensemble classification model results using the Pima Indian dataset, and the Clinical-Biological dataset on four evaluation metrics: Precision, Recall, F1-Score, and AUC-ROC is shown in Figure 5. Clinical-Biological data performed better across all metrics than the Pima data because it included deeper diagnostic biomarkers and quantitative characteristics derived from laboratory tests. Precision and Recall are improved, indicating the ability to identify cases of diabetes with fewer misidentifications and fewer false positives. The Clinical dataset's AUC-ROC was also the highest, indicating greater discriminatory ability. These findings suggest that prediction accuracy is higher when the data set is more diverse and has characteristics that are clinically significant; therefore, it is more realistic to clinical practice in the real world.

Performance Comparison of Two data set used by Specific Features

Table 5. Performance comparison of two data dataset used by specific features

Model	PIMA Indian Diabetes Dataset	Clinical Biological Dataset
Features Used	8 Features: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree, Age	Multiple Features: Age, Gender, Blood Pressure, Cholesterol, Glucose, WBC Count, Liver Enzymes, BMI, etc.
Accuracy	0.75	0.80
Precision	0.78	0.83
Recall	0.72	0.77
F1-Score	0.75	0.79
AUC-ROC	0.82	0.86
Training Time (s)	0.03	0.12
Test Time (s)	0.02	0.09

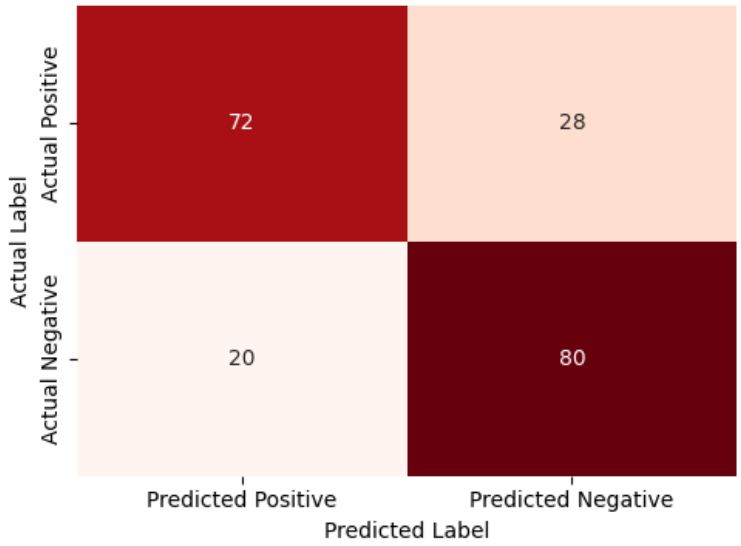

Figure 6. Confusion matrix for PIMA indian diabetes dataset

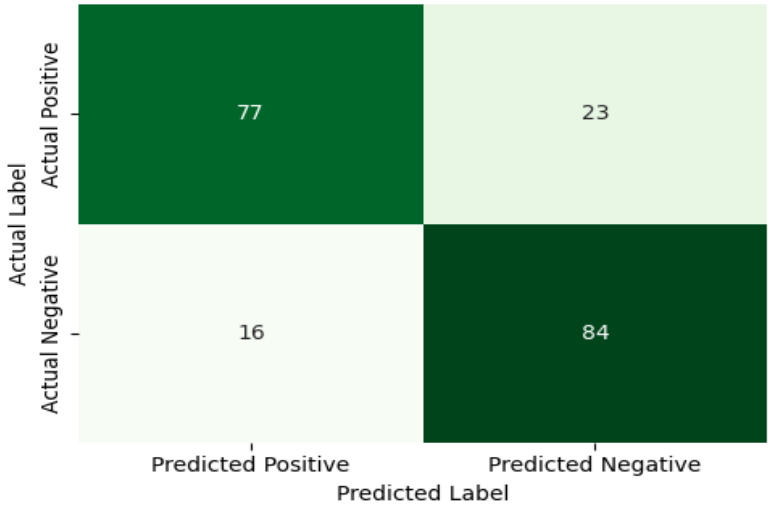

Figure 7. Confusion Matric-Clinical biological dataset

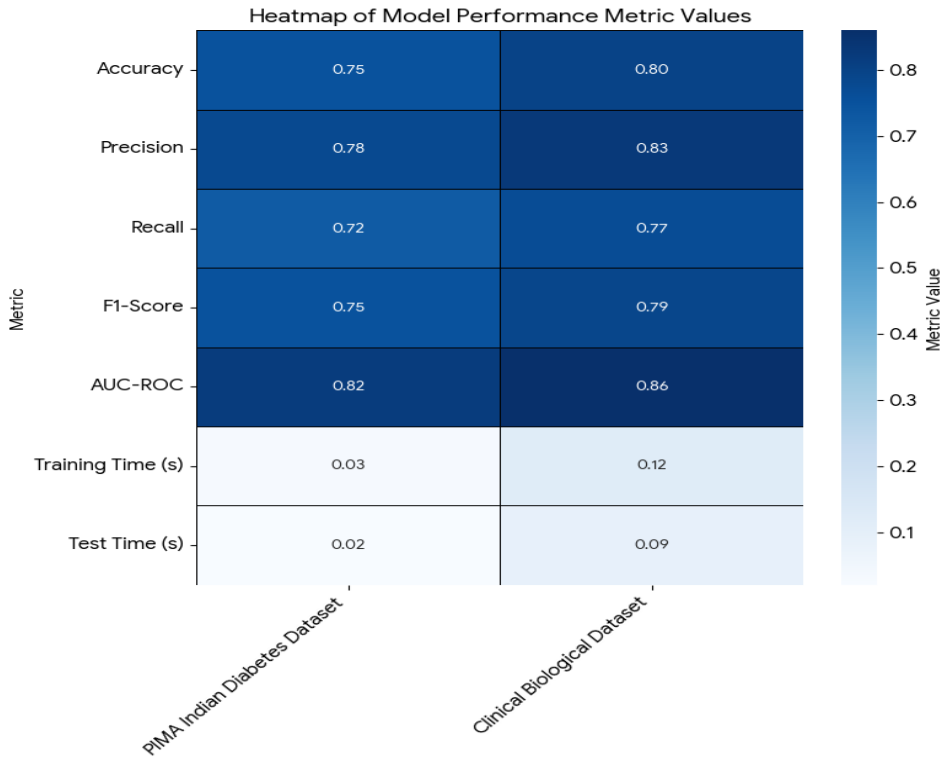

Figure 8. Heat Map representation of performance comparison

To interpret above Table 5 and Figure 6,7,8 indicates the existence of Key differences that are instigated by the complexity of features, which is evident in the performance appraisal of the Pima Indian Diabetes Dataset and Clinical-Biological Dataset. The presence of simple features of the Pima Indian Dataset (a total of 8) enables the dataset to achieve 0.75 accuracy although it is less complex as compared to the Clinical-Biological Dataset (more clinical and biological features) that achieves 0.80 accuracy. All the performance measures within the Clinical-Biological model have a better result than the Pima model - precision (0.83 vs. 0.78), recall (0.77 vs. 0.72), and F1-Score (0.79 vs. 0.75). The Clinical-Biological model has also outperformed the Pima model in AUC-ROC (0.86 vs. 0.82) illustrating better discriminative ability. Clinical-Biological model however takes much longer periods to train and test - 0.12 seconds to train and 0.09 seconds to test - compared to the Pima Indian Dataset that farts in record time. All in all, the Predictive power of the Clinical-Biological Dataset is less efficient and has more processing times although it is better and stronger. The Pima Indian Dataset Predictive power is also efficient and faster, but a bit weaker.

Challenges faced by Two Dataset (Pima Indian Dataset Vs Clinical Biological Dataset)

Table 6. Challenges faced by two datasets (pima indian dataset vs clinical biological dataset)

Challenge	Pima Indian Dataset	Clinical-Biological Dataset
Feature Set Limitations	3	4
Data Imbalance	4	4
Missing Data	3	4
Outliers	3	4
Feature Complexity	2	5
Data Quality	3	5
Generalization	4	3
Ethical Concerns	2	5

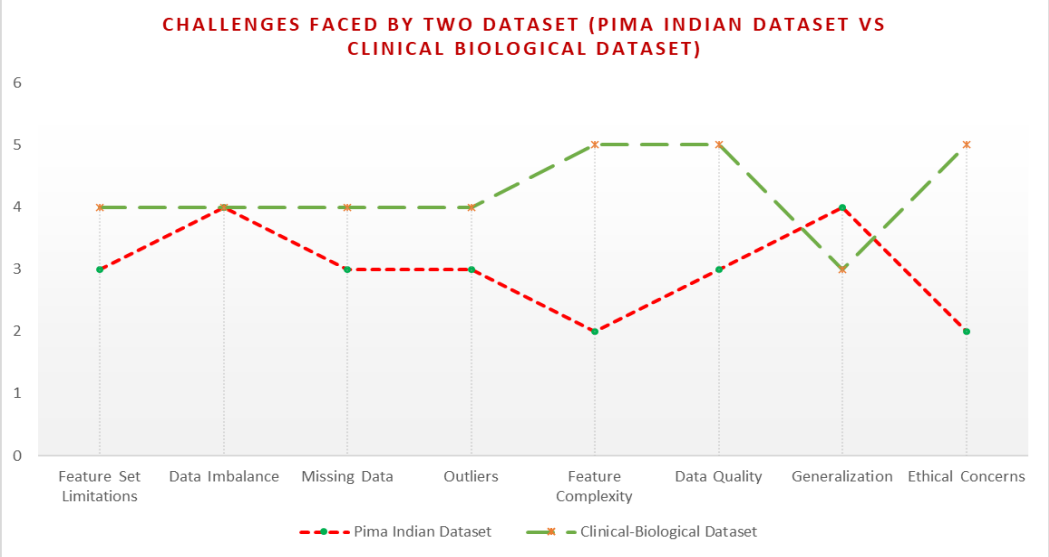

Figure 9. Challenges faced by two dataset (pima indian dataset vs clinical biological dataset)

The Table 6 above and Figure 9, above indicate the different issues surrounding both the PIMA Indian Diabetes Dataset and the Clinical Biological Dataset which are relatively similar with the PIMA Indian Dataset having slightly simpler feature sets than the Clinical Biological Dataset which has much more complex feature sets and is significantly more irrelevant. Missing data and outliers in any Dataset, especially in the Clinical Biological Dataset, further increase imbalance. The Clinical Biological Dataset will surely need more pre-treatment data than the PIMA Indian Dataset. The PIMA Indian Dataset is easier to handle because of its limited range and feature complexity, whereas the Clinical Biological Dataset can be more complex, even with the sophisticated features of the Diabetes Clinical Biological Data set, which will demand more sophisticated techniques to handle. In addition, the PIMA Indian Dataset is more likely to have fewer inconsistencies and raw data. Compared to the Clinical Biological Dataset, the Clinical Biological Dataset shall be able to capture more discrepancies and noisy data, which, on the other hand, will have an effect on the performance of the model. Compared to the Clinical Biological Dataset, the PIMA Indian Dataset is more likely to exhibit generalization difficulties on unobserved data. Conversely, the PIMA Indian Dataset will be more prone to generalization on the flip side. Finally, the Clinical Biological Dataset raises potential ethical issues, most likely due to the sensitivity of the health information, and confidentiality and data security must be considered. Conversely, the PIMA Indian Dataset does not have such problems. In general, the PIMA Indian Dataset is quite simple and simple to operate; nevertheless, the Clinical Biological Dataset is considered more complex, and the possibility of more diverse and generally more generalized results is suggested, though with even more significant ethical issues and questionable quality of data.

RESULT AND DISCUSSION

The comparative evaluation of eight machine learning models revealed that the variation of prediction performance was quite high based on the evaluation metrics of performance. In the iterative measures of precision and AUC-ROC, the best models for assessing the acceptability of linear-margin classifiers as candidates in structured clinical datasets were Logistic Regression and SVM. The Decision Tree had the highest recall, indicating that it was effective at identifying diabetes cases; however, it also showed higher false positives. Ensemble learning yielded better results, with the final prediction being the mean of the prediction probabilities from the several assessed models. The ensemble classifier achieved the highest AUC-ROC (0.8709), further demonstrating that heterogeneous models are a useful tool for minimizing variance and expanding the decision space. Although each model performed well, the ensembles' learning led to more stable performance and greater balance in the long run.

Looking at the results of our models on Pima Indian and Clinical-Biological data, we find some improvement in the Clinical-Biological data! The increase in Precision, Recall, and AUC-ROC performance measures is due to the dataset's more diverse laboratory biomarker features, including

insulin levels and diagnostic markers. Such characteristics enable greater learning about patterns of disease and pathology. Conversely, the Pima dataset is important in history as a data source, but it has fewer features; hence, the models we trained are less generalizable. Interestingly, the ensemble model showed stable predictive performance, confirming that combining multiple classifiers does not necessarily improve performance when constrained by the dataset. The comparison presents evidence that feature richness is essential for better diabetes prediction and clinical utility.

It has been found that ML-based prediction systems can serve as practical decision-support systems for early-stage diabetes screening, particularly with large, diverse patient data. Ensemble learning is an engaging format that can be used to overcome the bias in an individual model and makes predictions more stable, which is a necessity in clinical environments. In addition, the paper underscores the value of integrating various clinical-biological variables and, notably, demonstrates the benefit of well-integrated data in improving the accuracy and interpretability of models. These findings imply that the selected features and the multimodal clinical databases need to be optimized, and that hybrid ML-DL models can be applied in the future to enhance predictive performance. On the whole, is the suggested ensemble-based structure a scalable means of conducting real-time risk evaluation of diabetes and early intervention decision-making in healthcare systems

CONCLUSION AND FUTURE WORK

Statistical methods for measuring Precision, Recall, F1-score, and AUC-ROC were used to evaluate the performance of eight machine learning algorithms. The findings justified that the data range features and the intricacy of the classifiers largely influence model performance. Logistic Regression and SVM scored higher in Precision and AUC-ROC, indicating high discriminative power and reduced false-positive rates. Decision Tree had the highest Recall score, indicating greater sensitivity in detecting positive diabetic cases; however, it also resulted in more false alarms. Overall, the Ensemble Learning approach consistently outperformed other models, with an AUC-ROC of 0.8709, demonstrating that combining models improved overall performance and generalization through reduced overfitting. Moreover, across the Clinical–Biological and Pima Indian datasets, each comparison yielded statistically significant results, indicating that feature diversity improves predictive performance ($p < 0.05$). As such, the study concludes that ensemble-based classifiers, with diverse clinical input improve diabetes prediction with statistical and clinical reliability for clinical decision support.

In future work, the model can be extended for real-time clinical use by integrating electronic health records (EHRs) with streaming patient data, enabling continuous monitoring and predictions. Incorporating deep learning architectures, such as LSTM and Transformer-based models, can capture transitory shifts in patient biomarker trajectories. Future work can also improve the model's performance and reduce computing costs by optimizing feature selection using algorithms such as genetic algorithms and Bayesian optimization. To ensure privacy and safe federated learning between hospitals, research can investigate the application of federated learning paired with a blockchain network to ensure tamper-proof, decentralized patient data. Additionally, future work should introduce diverse datasets from different demographic and geographical populations to improve fairness, bias-free decision-making, and equity in the models. Future studies will be significantly essential to explore the incorporation of explanation techniques (i.e., SHAP, LIME), which will improve transparency and interpretability for making predictions by the physician.

REFERENCES

[1] Taskinen MR. Diabetic dyslipidaemia: from basic research to clinical practice. Diabetologia. 2003 Jun;46(6):733-49.

[2] Saratha B, Radhika MS, Priya VS. An Approach Towards Diabetic Retinopathy Detection and Analysis Through Cognitive Computing. Archives for Technical Sciences. 2025 J1(33): 125–134. https://doi.org/10.70102/afts.2025.1833.125

[3] Ganie AH, et al. Robust diabetic prediction using ensemble machine learning techniques with SMOTE. Scientific Reports. 2023.

[4] Vij P, Prashant PM. Predicting aquatic ecosystem health using machine learning algorithms. International Journal of Aquatic Research and Environmental Studies. 2024;4(S1):39-44.

[5] Ganie SM, Malik MB, Arif T. Performance analysis and prediction of type 2 diabetes mellitus based on lifestyle data using machine learning approaches. Journal of Diabetes & Metabolic Disorders. 2022 Jun;21(1):339-52.

[6] Nithyalakshmi V, Sivakumar R, Sivaramakrishnan A. Automatic detection and classification of diabetes using artificial intelligence. International Academic Journal of Innovative Research. 2021;8(1):1–5. https://doi.org/10.9756/IAJIR/V8I1/IAJIR0801

[7] Sharma T, Shah M. A comprehensive review of machine learning techniques on diabetes detection. Visual Computing for Industry, Biomedicine, and Art. 2021 Dec 3;4(1):30.

[8] Kumar V, Shah M. Multi Disease Prediction Using Deep Learning Framework for Electric Health Record. International Academic Journal of Science and Engineering. 2021;8(4):24-8. https://doi.org/10.71086/IAJSE/V8I4/IAJSE0827

[9] Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthcare technology letters. 2023 Feb;10(1-2):1-0. https://doi.org/10.1049/htl2.12039

[10] Debebe B. Levels, trends and determinants of under-five mortality in Amhara Region, Ethiopia: evidence from Demographic and Health Survey (2000–2011). International Academic Journal of Social Sciences. 2016;3(2):96–112.

[11] Afsaneh E, Sharifdini A, Ghazzaghi H, Ghobadi MZ. Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. Diabetology & Metabolic Syndrome. 2022 Dec 27;14(1):196.

[12] Shin J, Lee J, Ko T, Lee K, Choi Y, Kim HS. Improving machine learning diabetes prediction models for the utmost clinical effectiveness. Journal of Personalized Medicine. 2022 Nov 14;12(11):1899. https://doi.org/10.3390/jpm12111899

[13] Fomekong RL, Saruhan B. Titanium based materials for high-temperature gas sensor in harsh environment application. Chemistry Proceedings. 2021 Jun 30;5(1):66. https://doi.org/10.3390/CSAC2021-10480

[14] Kiran M, Xie Y, Anjum N, Ball G, Pierscionek B, Russell D. Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis. Frontiers in Digital Health. 2025 Mar 27;7:1557467. https://doi.org/10.3389/fdgth.2025.1557467

[15] Qin L. A Prediction Model of Diabetes Based on Ensemble Learning. InProceedings of the 2022 5th International Conference on Artificial Intelligence and Pattern Recognition 2022 Sep 23 (pp. 45-51). https://doi.org/10.1145/3573942.3573949

[16] Hasan R, Dattana V, Mahmood S, Hussain S. Towards transparent diabetes prediction: combining automl and explainable AI for improved clinical insights. Information. 2024 Dec 26;16(1):7. https://doi.org/10.3390/info16010007

[17] Kaliappan J, Saravana Kumar IJ, Sundaravelan S, Anesh T, Rithik RR, Singh Y, Vera-Garcia DV, Himeur Y, Mansoor W, Atalla S, Srinivasan K. Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets. Frontiers in Artificial Intelligence. 2024 Aug 21;7:1421751. https://doi.org/10.3389/frai.2024.1421751

[18] Zhao M, Yao Z, Zhang Y, Ma L, Pang W, Ma S, Xu Y, Wei L. Predictive value of machine learning for the progression of gestational diabetes mellitus to type 2 diabetes: a systematic review and meta-analysis. BMC Medical Informatics and Decision Making. 2025 Jan 13;25(1):18.

[19] Khokhar PB, Gravino C, Palomba F. Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review. Artificial intelligence in medicine. 2025 Apr 15:103132. https://doi.org/10.1016/j.artmed.2025.103132

[20] Dutta A, Hasan MK, Ahmad M, Awal MA, Islam MA, Masud M, Meshref H. Early prediction of diabetes using an ensemble of machine learning models. International Journal of Environmental Research and Public Health. 2022 Sep 28;19(19):12378. https://doi.org/10.3390/ijerph191912378

[21] Chowdhury P, Barua P, Uddin MN. Diabetes prediction using machine learning and hybrid deep learning ensemble technique. In2024 IEEE International Conference on Computing, Applications and Systems (COMPAS) 2024 Sep 25 (pp. 1-7). IEEE. https://doi.org/10.1109/COMPAS60761.2024.10796486

[22] Yan D, Li X, Wang Y, Cai Z. Optimized prediction of diabetes complications using ensemble learning with Bayesian optimization: a cost-efficient laboratory-based approach. Frontiers in Endocrinology. 2025 Jun 20;16:1593068. https://doi.org/10.3389/fendo.2025.1593068

[23] Sethi H, Goraya A, Sharma V. Artificial Intelligence based Ensemble Model for Diagnosis of Diabetes. International Journal of Advanced Research in Computer Science. 2017 May 15;8(5).

[24] Fregoso-Aparicio L, Noguez J, Montesinos L, García-García JA. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. Diabetology & metabolic syndrome. 2021 Dec 20;13(1):148.

[25] Firdous S, Wagai GA, Sharma K. A survey on diabetes risk prediction using machine learning approaches. Journal of family medicine and primary care. 2022 Nov 1;11(11):6929-34. https://doi.org/10.4103/jfmpc.jfmpc_502_22