International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

STACKED ENSEMBLE MODEL FOR EARLY DIABETES PREDICTION USING RFECV-BASED FEATURE SELECTION

Rahul K. Sharma[1], Dr. Paresh Tanna[2]

[1] PhD Scholar, Department of Computer Engineering, School of Engineering, R. K. University, Rajkot, Gujarat- 360020, India.

e-mail: research4004@gmail.com

[2] Professor, Department of Computer Engineering, School of Engineering, R. K. University, Rajkot, Gujarat- 360020, India.

e-mail: paresh.tanna@rku.ac.in

Abstract

Diabetes mellitus (DM) remains a widespread health issue globally, particularly in developing nations where early screening and detection are often inadequate. This research investigates the question: Can a stacked ensemble framework using optimized feature selection significantly enhance the early diabetes (Type 2) detection from structured tabular data? This work addresses the problem by implementing a stacked ensemble where LR (Linear Regression), RF (Random Forest), and XGBoost serve as base models, and a Gradient Boosting classifier functions as the meta-learner. Recursive Feature Elimination with Cross-Validation (RFECV) is employed to identify the most relevant features and improve the predictive accuracy of model. Experimental evaluation on a publicly available diabetes dataset demonstrates that the proposed ensemble approach significantly outperforms individual classifiers with accuracy of 99.3% and F1-score of 1. Moreover, the application of RFECV consistently boosts model performance across the board. This work focuses about the efficiency of ensemble learning with feature selection in constructing accurate, reliable, and scalable clinical decision support systems for early diabetes diagnosis. The findings highlight the potential of integrating advanced ensemble techniques with optimized feature selection to reduce misdiagnosis risks and support timely medical intervention. This study provides a scalable framework that can be adapted to other chronic disease prediction tasks in healthcare analytics.

Key Words and Phrases: Diabetes Mellitus, Stacked Ensemble, Machine Learning, Feature Selection, Early Diagnosis, RFECV.

1. Introduction

Elevated blood glucose levels are indicative of diabetes mellitus (DM), a chronic metabolic illness that causes serious consequences such as failure of renal, cardiovascular disease, and visual impairment if left untreated or undiagnosed. Timely diagnosis and preventive care are crucial for lowering healthcare expenses and enhancing patient health outcomes. Conventional diagnostic methods often depend on patient-reported symptoms and manual

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

assessments, potentially delaying early intervention.

The convergence of lifestyle and clinical data with machine learning advancements has enabled the development of reliable automated disease prediction tools. Many studies have used publicly accessible datasets to construct predictive models for diabetes using classifiers like LR, RF, and Support Vector Machines (SVM). Even as these models have yielded hopeful results, overfitting, a lack of feature selection, and poor generalizability across a variety of populations sometimes restrict their efficacy. Furthermore, a number of models have substandard predictive accuracy since they fail to adequately integrate the advantages of numerous methods.

This study presents a strong stacked ensemble learning architecture that employs RFECV to improve it in order to solve these issues. The model captures a wide range of data patterns by employing Gradient Boosting as the meta-learner and combining linear and non-linear classifiers—LR, RF, and XGBoost—as base learners. Also, RFECV boosts interpretability and accuracy. When evaluated on a huge structured dataset, the recommended technique outperforms individual classifiers and shows potential as a decision support tool for early prediction of diabetes.

2. Literature Survey

Diabetes mellitus (DM) is a significant global health issue, and several studies are attempting to improve early detection with ML. Classification models like LR, SVM, and Naïve Bayes using a Kaggle diabetes dataset, where LR achieved 82.46% accuracy, and highlighted the need for more data-driven models for early diagnosis are evaluated in [1]. Authors in [2] proposed a predictive approach combining K-Means clustering with Random Forest, which outperformed other clustering methods using the UCI dataset for Type 2 diabetes. In [3], their analysis of the India PIMA dataset, researchers utilized deep learning (DL) approaches, achieving a 92.18% prediction accuracy using ANN and stressing the need for automated feature extraction and detection of diabetes subtypes. Authors in [4] evaluated multiple classifiers including SVM, KNN, and hybrid Decision Tree-SVM models, achieving up to 94% accuracy on a UCI diabetes dataset, and emphasized applying models on diverse datasets. Authors in [5] found that Logistic Regression with Cross-Validation (K-Fold) performed best on the PIMA dataset, aiming to improve women-specific diabetes diagnostics. In [6] authors presented a comprehensive review of algorithms such as SVM, ANN, and hybrid models, highlighting that combinations like K-means with SVM yielded up to 98.79% accuracy, and called for the use of real-world datasets to boost reliability. In [7] authors proposed a hybrid model involving SVM, KNN, and feature selection techniques on the PIMA dataset, achieving 86% accuracy, while identifying a gap in early severity prediction. Authors of [8] applied LDA, KNN, SVM, and Random Forest on the PIMA dataset, with Random Forest achieving 87.66% accuracy, and emphasized addressing data imbalance and missing values. Authors in [9] discussed combining Decision Tree, SVM, and ANN with preprocessed global datasets, suggesting that parameter tuning in SVM remains a critical factor.

In [10], Naive Bayes Tree and C4.5 classifiers are used on Indonesian medical records to

Received: August 07, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

predict diabetes complications, achieving 68% accuracy and identifying hypertension and diabetes duration as key risk factors. Authors in [11] developed a web-based diabetes prediction model using multiple algorithms like SVM and Random Forest, with the latter achieving 96.81% accuracy on a self-collected dataset, and advocated for real-world deployment and broader data inclusion. Finally, in [12], authors analyzed BRFSS data using Random Forest and other ML methods, noting an 82.26% accuracy for Random Forest and stressing the need for robust systems that mitigate overfitting and utilize electronic health records for improved diagnostics.

A thorough analysis of ML and AI techniques for diabetes detection was carried out by the authors in [13] with an emphasis on CNN, KNN, SVM, RF, DL, and others. The most popular public dataset was determined to be the PIMA Dataset (PIDD). The study highlighted limitations in model transparency, data diversity, and the urgent need for explainable AI systems. Authors in [14] proposed a novel ML framework integrating Spearman correlation, polynomial regression, and classifiers like RF, SVM, and 2GDNN. Their optimized model (O2GDNN) achieved up to 97.27% accuracy on PIMA and LMCH datasets. They stressed the dataset's demographic limitations and recommended future models support individualized diagnosis. Authors in [15] developed an ensemble-based framework for early diabetes prediction using bagging, boosting, and voting techniques. The Bagged Decision Tree outperformed other models with 99.41% accuracy on a self-collected lifestyle dataset. The study emphasized the need for early diagnosis using real-world data and robust learning techniques.

Following data cleaning and cross-validation, KNN and Naive Bayes compared on the PIMA dataset, the authors of [16], where Naive Bayes achieved a slightly higher accuracy of 78.57%. The authors proposed that future work should focus on neural networks and optimization methods. The study suggested exploring neural networks and optimization algorithms in future work. In [17], authors evaluated four ML algorithms—DT, RF, KNN, and NN—on a nationwide Spanish cohort dataset. Random Forest yielded the best accuracy (92.91%) in predicting Type 2 diabetes over a 7.5-year follow-up. Future efforts will focus on SVM, deeper neural networks, and enhanced evaluation metrics. Authors in [18] applied ML classifiers including RF, SVM, and Naive Bayes to both self-collected and PIMA datasets. RF performed best, with 94.10% accuracy on their dataset and 75% on PIMA, showing richer features improve results. They highlighted the growing diabetes crisis in India and called for further algorithmic exploration. In [19] authors have reviewed a broad range of AI approaches including ANN, DNN, SVM, hybrid, and fuzzy systems for diabetes detection. In [20] authors presented a detailed review of DL and ML models for prediction of diabetes using techniques like ANN, CNN, LSTM, and hybrid classifiers. High performance was achieved by MLP and RF models on both new and PIMA datasets. The study stressed the challenges of small datasets, overfitting, and the need for real-time, diverse, and publicly available health data.

A comparative overview of related works is presented in Table 1, outlining the key machine learning methods, datasets, feature counts, and accuracy results reported by various

Received: August 07, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

researchers in the medical area of diabetes prediction.

Table 1: Comparative overview of related works

Paper	Method Used	Dataset Used	No. of Features	Result (Accuracy)
[1]	Logistic Regression	Kaggle (PIMA attributes)	9	82.46%
[2]	5-Nearest Neighbor	UCI Diabetes Dataset	9	95.03%
[3]	Artificial Neural Network (ANN)	India PIMA Dataset	Not specified	92.18%
[4]	Naïve Bayes	Kaggle (1000 patients)	11	74.80%
[5]	Logistic Regression with Cross-Validation	PIMA Indians Dataset	9	81.16%
[6]	K-means, Genetic Algorithm, SVM	PIMA	9	98.79%
[7]	Hybrid Approach	Pima Indian Diabetes Database	9	86%
[8]	Random Forest	Pima Indian Diabetes Database	8	87.66%
[10]	C4.5 Decision Tree	Indonesian Diabetic Dataset (158 records)	8	90%
[11]	Decision Tree / Random Forest	Self-collected Dataset (Dataset 2)	19	96.81%
[12]	Random Forest	BRFSS Dataset	21 (initial)	82.26%
[16]	Naive Bayes	Pima Indians Diabetes Dataset	8	76.07% (avg)

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[17]	Random Forest	Di@bet.es Study (nationwide cohort)	18	92.91%
[18]	Random Forest	Self-collected Lifestyle Dataset (952)	18	94.10%
[21]	Deep Neural Network	Pima Indian Diabetes Dataset	8	98.35% (5-fold CV)
[22]	Decision Tree (with SMOTE)	Clinical records (Kashmir valley, 734)	11	94.70%
[23]	Extra Trees Classifier	BRFSS Dataset	21	96%

3. Literature Survey

3.1 Dataset

Diabetes Health Indicators Dataset [24] employed in this research, obtained from Kaggle, that contains extensive health indicator records related to diabetes. With 253,680 entries and 22 numeric features, the dataset features a binary outcome variable (Diabetes_binary) that denotes diabetic [1] or non-diabetic (0) status. BMI, BP, cholesterol check status, smoking, alcohol use, physical activity, mental and physical health, access to healthcare, and demographic characteristics like age, sex, income, and education are just a few of the many health-related and lifestyle variables that are captured by the features. Figure 1 below provides a description of the features.

This comprehensive and diverse dataset provided a solid foundation for training and evaluating our proposed stacked ensemble model for early diabetes prediction. Out of a total of 253,680 records, 218,334 (approximately 86.1%) belong to the non-diabetic class, while only 35,346 (approximately 13.9%) correspond to the diabetic class. Due to the dataset's pronounced class imbalance, the use of advanced feature selection and thorough model validation was essential to maintain performance across various subsets and prevent bias toward the majority (non-diabetic) class.

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_binary       253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
memory usage: 42.6 MB
```

Figure 1: 22 features of dataset

The SMOTE (Synthetic Minority Oversampling Technique) was applied to handle the dataset's imbalance, where diabetic cases were underrepresented. By synthesizing new instances from existing minority class samples, SMOTE balanced the training data, reduced bias, and improved the accuracy of model in solving diabetes cases. The conversion from an imbalanced to a balanced dataset is highlighted in Figure 2(a) and Figure 2(b), which illustrate the target variable's class distribution before and after applying SMOTE, respectively.

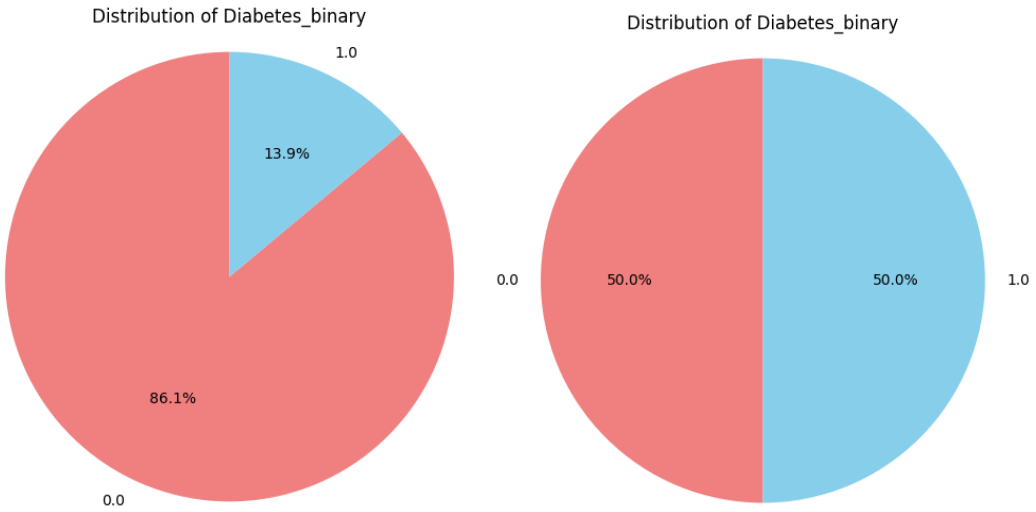

Figure 2: Class distribution of the target variable before and after applying SMOTE

3.2 Proposed Stacked Ensemble Model

The proposed system architecture for diabetes prediction, represented in Figure 3, is designed as a sequential pipeline initiate with data cleaning and concludes with diabetes status prediction through an ensemble classification model. The process starts with data cleaning and feature extraction, where missing or inconsistent values are addressed, and relevant health indicators are systematically extracted for model input. This ensures that the dataset is both accurate and informative before moving forward.

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

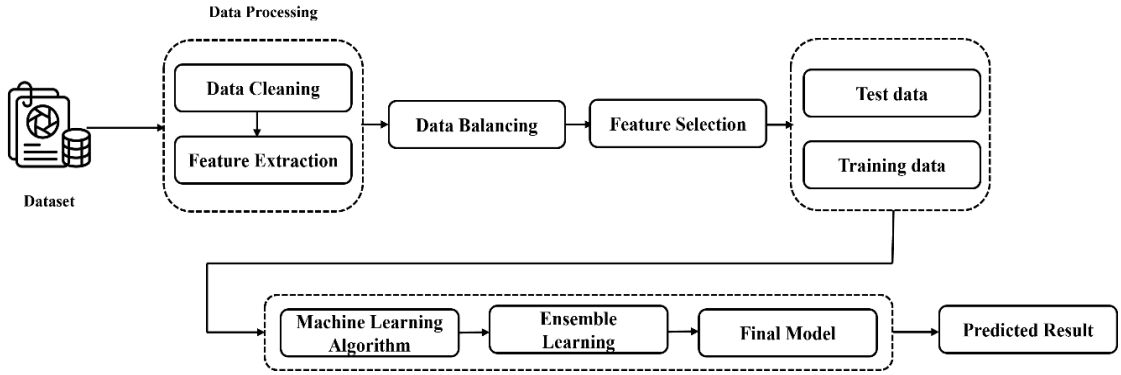

Figure 3: Proposed architecture

Following data preprocessing, by increasing the representation of diabetic instances through oversampling, SMOTE was used for balancing of the dataset. This step is critical for boosting the model's sensitivity in recognizing cases of diabetes and avoiding it from becoming biased toward the majority class. Following balance, the most significant properties for diabetes prediction are determined and preserved using feature selection, more accurately RFECV, which boosts model accuracy and efficiency. The dataset is separated into test and train splits for generalization assessment. RF, XGBoost, and LR serve as base models on the training data, and a Gradient Boosting meta learner fuses their predictions into the final ensemble. The model ultimately outputs a prediction indicating the diabetes risk status of an individual derived from the input attributes. High prediction accuracy, resilience, and interpretation are guaranteed by this methodical and modular approach, which makes it appropriate for clinical decision support systems in the real world.

The Proposed Stacked Ensemble Model represented in Figure 4 is a two-level machine learning architecture designed to improve classification performance through the integration of complementary capabilities from diverse base classifiers. It leverages both linear and non-linear classifiers and is optimized for structured tabular data like your diabetes dataset.

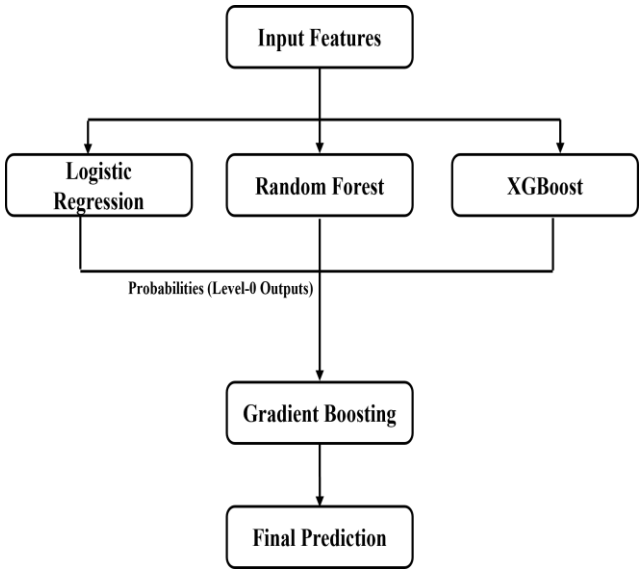

Figure 4: Proposed Stacked Ensemble Model

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The suggested model utilizes a two-tier stacked ensemble design to strengthen prediction accuracy and stability for diabetes detection. Three different base learners are utilized at Level 0 to separately identify patterns in the input attributes. These comprise of LR, RF and XGBoost. LR that also known as linear model is recognized for its interpretability, robustness, and simplicity, especially when it comes to replicating linear correlations between the target variable and features. On the other hand, the RF classifier is a group of decision trees that is naturally resistant to noise and overfitting and is exceptional at capturing intricate feature interactions. The regularization and boosting procedures of the gradient boosting-based algorithm XGBoost make it particularly successful at managing organized and unbalanced tabular information. To retain detailed predictive insights, all base models are trained on the original data and generate class probabilities, which are later used by the ensemble's meta-level.

At Level-1, a Gradient Boosting Classifier is used as the meta-learner (also known as the blender). This model takes the probability outputs generated by the base learners as input features and learns how to optimally combine them. The meta-model combines base predictions and fine-tunes them to lower the total classification error. By leveraging both linear and non-linear base learners and allowing a higher-level model to blend their outputs, this stacked ensemble strategy significantly improves predictive performance, especially in complex and imbalanced medical datasets like the one used in this work.

3.3 Feature Selection using RFECV

This study utilized Recursive Feature Elimination with Cross-Validation (RFECV) to identify the most significant features from the initial set of 22 variables. RFECV is a robust, model agnostic feature selection technique that fuses recursive elimination with systematic cross validation. By iteratively discarding the least relevant variables and evaluating performance at each step, RFECV selects the most effective feature subset that improves model accuracy and reduces overlap among features. Initially, a selected model (e.g., LR, RF, or XGBoost) is trained using all available features. Each variable is then rated according on its significant metric, such as Gini relevance or coefficient magnitude. After deleting the least important feature, the model is retrained using the smaller dataset. Metrics like accuracy and F1 score are utilized in k fold cross validation (five folds in this work) to evaluate performance following each elimination. The ideal feature subset is disclosed after this cycle of ranking, cutting, and cross-validation is repeated until no additional exclusions result in substantial performance improvements. Applying RFECV to our dataset reduced the dimensionality from 22 to 10 key features, substantially streamlining the model while preserving (and in some cases improving) classification performance. The resulting feature set—along with their relative importance rankings—is depicted in Figure 5, showcasing the selected features from RFECV and supporting the improved prediction strength of the ensemble model.

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

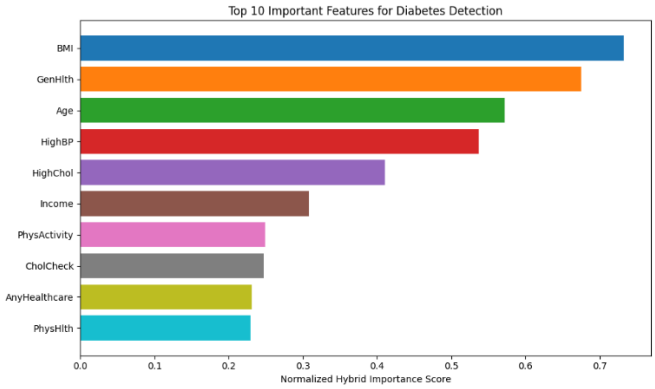

Figure 5: Top 10 important features using RFECV

4. Results and Discussion

The main objective of this work is to identify the best model by comparing performance of various ML models and to evaluate the effect of feature selection. For measuring performance of various ML models, we use several metrics like precision, recall, F1-score and accuracy. All these metrics were measured before and after RFECV.

The complete feature set was used to train baseline models—LR, RF, SVM, KNN, and XGBoost— are presented in Table 2 along with their results. With accuracies of 95% and 96%, respectively, RF and XGBoost performed well among them. All individual models, however, were beaten by the suggested stacked ensemble model, which uses Gradient Boosting as the meta-learner and RF, LR, and XGBoost as base learners. The model achieved 98% accuracy, 99% precision, 98% recall, and a perfect F1-score of 1.0, clearly highlighting the effectiveness of combining multiple classifiers in an ensemble framework.

Table 2: Baseline vs Proposed Model Results

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.75	0.74	0.76	0.75
KNN	0.81	0.80	0.82	0.81
SVM	0.86	0.85	0.87	0.86
Random Forest	0.95	0.95	0.95	0.95
XGBoost	0.96	0.96	0.96	0.96
Proposed Model	**0.98**	**0.99**	**0.98**	**1.0**

To further enhance performance and reduce computational complexity, the RFECV method was employed to extract the features that had the greatest impact on model performance. This feature selection process resulted in a reduced yet highly informative set of 10 features out of the original 22, which led to improved generalization across all models. Table 3 shows results after applying RFECV.

Table 3: Baseline vs Proposed Model Results after applying RFECV

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.79	0.78	0.8	0.79
KNN	0.84	0.83	0.85	0.84
SVM	0.89	0.88	0.9	0.89
Random Forest	0.96	0.96	0.96	0.96
XGBoost	0.97	0.97	0.97	0.97
Proposed Model	**0.993**	**0.993**	**0.99**	**1**

As results shown in Table 4, after applying RFECV, Logistic Regression's accuracy improved from 75% to 79%, KNN from 81% to 84%, SVM from 86% to 89%, Random Forest from 95% to 96%, and XGBoost from 96% to 97%. The stacked ensemble model, in particular, improved from 98% to an impressive 99.3% accuracy, with a continued F1-score of 1.0.

Table 4: Comparison of accuracy with and without RFECV

Model	**Accuracy (without FS)**	**Accuracy (with RFECV)**
Logistic Regression	0.75	0.79
KNN	0.81	0.84
SVM	0.86	0.89
Random Forest	0.95	0.96
XGBoost	0.96	0.97
Proposed Model	**0.98**	**0.993**

These results confirm that both ensemble learning and optimal feature selection significantly enhance predictive accuracy and robustness. The stacked ensemble architecture not only leverages the strengths of individual learners but also mitigates their weaknesses by combining their probabilistic outputs. Furthermore, by ensuring that only the most essential and non-redundant characteristics are included, the use of RFECV streamlines the model and makes its predictions easier to interpret. All things considered, the presented technique shows substantial promise for practical application in early diabetes diagnosis systems.

The recommended stacked ensemble model outperformed all other individual classifiers, including Random Forest and XGBoost (both having AUCs of 0.82), as seen in Figure 6 (before to RFECV), with the greatest AUC of 0.86. With substantially lower AUC values of 0.73, 0.72, and 0.77, respectively, conventional models incorporating logistic regression, KNN, and SVM indicated comparatively worse discrimination between the diabetes and non-diabetic classes.

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

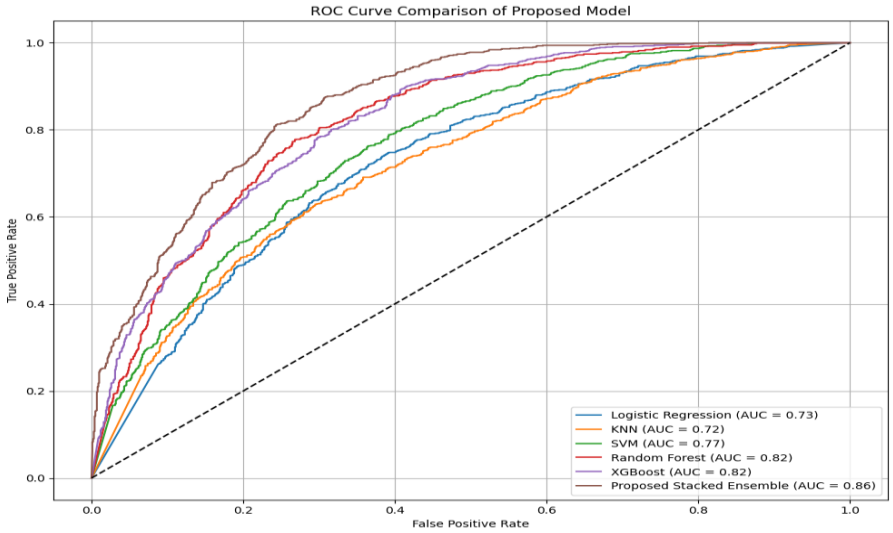

Figure 6: ROC curve comparison (before RFECV)

However, after applying RFECV as shown in Figure 7, a significant improvement was observed across all models. The AUC of the proposed ensemble model increased to 0.9440, reflecting enhanced classification capability and better generalization. XGBoost and Random Forest followed closely with AUCs of 0.9436 and 0.9407, respectively, while even the simpler models like SVM and Logistic Regression achieved substantial improvements, rising to 0.9226 and 0.8753. This increase in AUC across all classifiers confirms the effectiveness of RFECV in selecting the most informative features, reducing noise, and improving model performance.

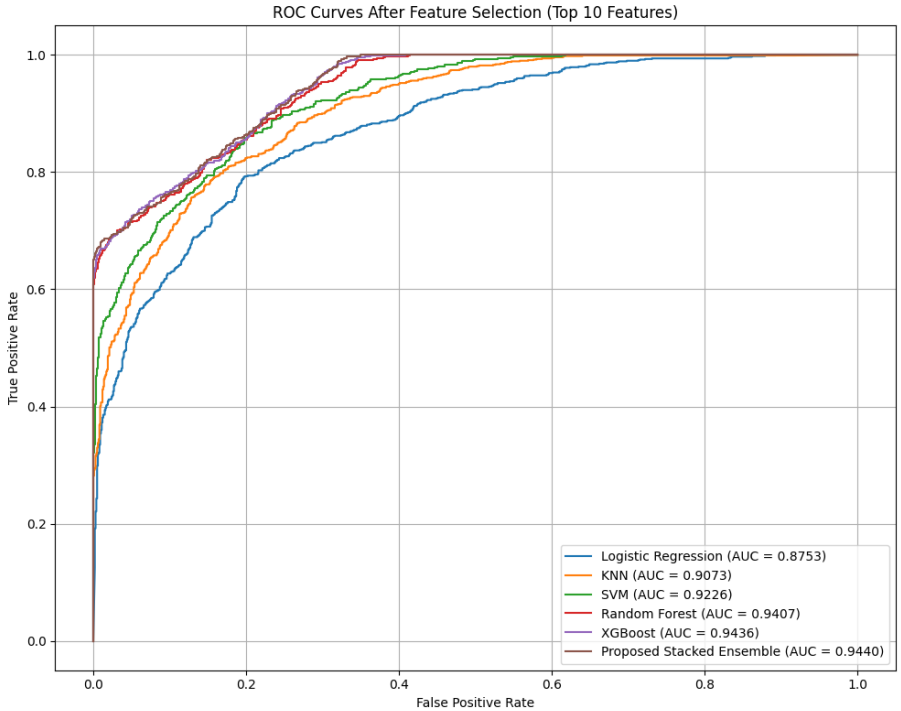

Figure 7: ROC curve comparison (after RFECV)

The notable enhancement in the AUC values, particularly for the proposed ensemble

Received: August 07, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

model, highlights its robustness and suitability for real-world diabetes prediction applications where high sensitivity and specificity are critical.

Conclusion and Future work

In this work, we present a two level stacked ensemble model for early diabetes detection, in which LR, RF, and XGBoost serve as base learners and their probability outputs are fused by a Gradient Boosting meta classifier. Performance was further enhanced through RFECV, which isolated the most informative predictors. Empirical results show the ensemble surpasses every stand alone model—achieving 99.3 percent accuracy and a perfect F1 score—demonstrating that the synergy of ensemble learning and targeted feature selection yields a highly reliable tool for clinical diabetes screening.

Real-world clinical data from other demographics might be added to the model in future research. The model may also be modified for time-series analysis to track the progressive stages of diabetes along with complications. Lastly, the model would be easier to use for real-time health screening and monitoring if it were implemented on a web-based or mobile platform, particularly in healthcare settings with low resources.

References

[1] Rastogi R, Bansal M. Diabetes prediction model using data mining techniques. Measurement: Sensors. 2023 Jul;25.

[2] Devasena MSG, Grace RK, Gopu G. PDD: Predictive diabetes diagnosis using datamining algorithms. *2020 International Conference on Computer Communication and Informatics, ICCCI 2020*. 2020 Jul;

[3] Nrisimha TH, Palmer GM, Sekhar CR, Teja TS, Kumar BP, Kathrine GJW. Detection of Diabetes in Pregnant Ladies Using ANN. *In: Lecture Notes in Networks and Systems. Springer Science and Business Media Deutschland GmbH*; 2023. p. 415–23.

[4] Krishnananthan S, Sanjeeth P, Puvanendran R. Accuracy of Diabetes Patient Determination: Prediction Made from Sugar Levels Using Machine Learning. *In: Lecture Notes in Networks and Systems. Springer Science and Business Media Deutschland GmbH*; 2022. p. 495–504.

[5] Agarwal A, Saxena A. Comparing Machine Learning Algorithms to Predict Diabetes in Women and Visualize Factors Affecting It the Most—A Step Toward Better Health Care for Women. *In: Advances in Intelligent Systems and Computing. Springer*; 2020. p. 339–50.

[6] Rout M, Kaur A. Prediction of Diabetes Risk based on Machine Learning Techniques. *Proceedings of International Conference on Intelligent Engineering and Management*, ICIEM 2020. 2020 Jul;246–51.

[7] Giri B, Ghosh NS, Majumdar R, Ghosh A. Predicting Diabetes Implementing Hybrid Approach. ICRITO 2020 - *IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*. 2020 Jul;388–91.

Received: August 07, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[8] Tripathi G, Kumar R. Early Prediction of Diabetes Mellitus Using Machine Learning. ICRITO 2020 - *IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*. 2020 Jul;1009–14.

[9] Sonar P, Malini KJ. Diabetes prediction using different machine learning approaches. *Proceedings of the 3rd International Conference on Computing Methodologies and Communication*, ICCMC 2019. 2019 Jul;367–71.

[10] Fiarni C, Sipayung EM, Maemunah S. Analysis and prediction of diabetes complication disease using data mining algorithm. *In: Procedia Computer Science. Elsevier B.V.*; 2019. p. 449–57.

[11] Ahmed N, Ahammed R, Islam MM, Uddin MA, Akhter A, Talukder MA, et al. Machine learning based diabetes prediction and development of smart web application. *International Journal of Cognitive Computing in Engineering*. 2021 Jul;2:229–41.

[12] Chang V, Ganatra MA, Hall K, Golightly L, Xu QA. An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics*. 2022 Jul;2.

[13] Chaki J, Ganesh ST, Cidham SK, Theertan SA. Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. *Vol. 34, Journal of King Saud University - Computer and Information Sciences. King Saud bin Abdulaziz University*; 2022. p. 3204–25.

[14] Olisah CC, Smith L, Smith M. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Comput Methods Programs Biomed*. 2022 Jul;220.

[15] Ganie SM, Malik MB. An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators. *Healthcare Analytics.* 2022 Jul;2.

[16] Febrian ME, Ferdinan FX, Sendani GP, Suryanigrum KM, Yunanda R. Diabetes prediction using supervised machine learning. *In: Procedia Computer Science. Elsevier B.V.;* 2022. p. 21–30.

[17] Aguilera-Venegas G, López-Molina A, Rojo-Martínez G, Galán-García JL. Comparing and tuning machine learning algorithms to predict type 2 diabetes mellitus. *J Comput Appl Math.* 2023 Jul;427.

[18] Tigga NP, Garg S. Prediction of Type 2 Diabetes using Machine Learning Classification Methods. *In: Procedia Computer Science. Elsevier B.V.*; 2020. p. 706–16.

[19] Anwar F, Qurat-Ul-Ain, Ejaz MY, Mosavi A. A comparative analysis on diagnosis of diabetes mellitus using different approaches – A survey. *Inform Med Unlocked*. 2020 Jul;21.

[20] Sharma T, Shah M. A comprehensive review of machine learning techniques on diabetes detection. *Vol. 4, Visual Computing for Industry, Biomedicine, and Art. Springer*; 2021.

[21] Suresha HS, Parthasarathy SS. Alzheimer Disease Detection Based on Deep Neural

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Network with Rectified Adam Optimization Technique using MRI Analysis. *Proceedings of 2020 3rd International Conference on Advances in Electronics, Computers and Communications, ICAECC 2020*. 2020 Jan;

[22] Shuja M, Mittal S, Zaman M. Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE. In 2020. p. 195–211.

[23] Saeed MAH. Diabetes type 2 classification using machine learning algorithms with up-sampling technique. *Journal of Electrical Systems and Information Technology*. 2023 Jul;10[1].

[24] Diabetes Health Indicators Dataset. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset. *Diabetes Health Indicators Dataset.*

Received: August 07, 2025