Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(4), 784-797
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Early Detection of Diabetes Using a Hybrid Approach Based on the Voting Classifier

Adlès Francis Kouassi*, Bi Irié Cyrille Dje, Kigninman Désiré Kone, Olivier Asseu*

Ecole Supérieure Africaine des TIC (ESATIC), LASTIC, Abidjan, Côte d'Ivoire
Email: *darelorthiniel@yahoo.fr, iriecyrille.djebi@esatic.edu.ci, desire.kone@esatic.edu.ci, *oasseu@yahoo.fr

How to cite this paper: Kouassi, A.F., Dje, B.I.C., Kone, K.D. and Asseu, O. (2025) Early Detection of Diabetes Using a Hybrid Approach Based on the Voting Classifier. *Open Journal of Applied Sciences*, **15**, 784-797.
https://doi.org/10.4236/ojapps.2025.154052

Received: February 28, 2025
Accepted: March 24, 2025
Published: March 27, 2025

Abstract

The early detection of type 2 diabetes is a major challenge for healthcare professionals, as a late diagnosis can lead to severe and difficult-to-manage complications. In this context, this paper proposes an innovative hybrid approach based on an ensemble method using Voting, designed to improve the accuracy of diabetes prediction. Our methodology is based on three main steps. First, we balanced the dataset classes using the SMOTEENN method to correct imbalances and ensure a fair representation of positive and negative classes. Next, we combined three complementary algorithms—Extra Trees Classifier (ETC), XGBoost (XGB), and K-Nearest Neighbors (KNN)—using the Voting strategy. This combination allows us to leverage the specific strengths of each model while reducing their individual limitations. Finally, we applied GridSearch to optimize hyperparameters, ensuring maximum model performance. The results obtained from experiments conducted on the Pima Indians Diabetes Dataset are remarkable. Our hybrid model achieves an overall accuracy of 95.50%, a precision of 93.22%, a recall of 98.21%, an F1-Score of 95.65%, and an AUC-ROC of 98.83%. These performances surpass those of individual models, demonstrating the potential of this approach for developing reliable and effective tools dedicated to the early diagnosis of type 2 diabetes.

Keywords

Diabetes Prediction, Extra Trees Classifier, XGBoost, K-Nearest Neighbors, SMOTEENN, Machine Learning, Hyperparameter Optimization

1. Introduction

Diabetes mellitus is a chronic disease characterized by persistent hyperglycemia resulting from insulin deficiency. If not managed in time, it can lead to serious complications such as cardiovascular diseases, kidney failure, and blindness. The

prevalence of diabetes continues to grow: in 2019, approximately 463 million people worldwide were affected by this disease, and projections estimate that this number could reach 700 million by 2045, particularly in middle-income countries [1] [2]. Given this alarming trend, improving diagnostic tools is crucial to enable early detection and reduce severe complications such as amputations and cardiovascular disorders.

Several key clinical parameters are used to diagnose diabetes, including age, body mass index (BMI), triceps skinfold thickness, serum insulin, plasma glucose level, and diastolic blood pressure. However, traditional diagnostic methods have several limitations: they are time-consuming and complex, sometimes requiring several weeks or even months to obtain reliable results [3] [4]. In response to these challenges, machine learning advancements have emerged as a promising solution. By leveraging large-scale medical datasets, these approaches accelerate and enhance diagnostic accuracy, offering an efficient alternative to conventional methods [5].

Among these advancements, ensemble learning has emerged as an effective analytical method, which mimics human learning by combining multiple machine learning models. One of the key advantages of this approach is its ability to reduce bias, optimize performance, and improve prediction accuracy by leveraging the complementary strengths of different models [6]. By integrating multiple algorithms, ensemble methods provide more robust and precise models, offering promising prospects for early diabetes diagnosis and management.

Several studies have explored ensemble learning techniques to enhance the accuracy of diabetes classification.

Patil *et al.* [7] proposed an ensemble learning approach that combines various machine learning techniques. Compared to conventional methods such as Boosting, Bagging, Random Forest, and Random Subspace, this approach improved accuracy and reduced diagnostic time, achieving 82% accuracy on the Pima Indians Diabetes Dataset.

Bhopte and Rai [8] explored a hybrid deep learning model (CNN-LSTM) for diabetes detection, reaching an accuracy of 89.30%. Their study compared the effectiveness of their approach with other classification models on the same dataset.

Lei Qin [9] developed an ensemble learning-based diabetes prediction model, integrating logistic regression (LR), k-nearest neighbors (KNN), decision trees (DT), Gaussian Naive Bayes (GNB), and support vector machines (SVM). In their approach, four of these algorithms were used as base learners, combined with an SVM meta-learner, achieving 81.6% accuracy.

Kumari *et al.* [10] proposed a weighted voting ensemble approach, combining Random Forest (RF), Logistic Regression (LR), and Naïve Bayes (NB). Their comparative evaluation against AdaBoost, SVM, XGBoost, and CatBoost on the PIMA dataset demonstrated that their ensemble achieved 79.04% accuracy and an F1-score of 80.6%, surpassing several individual models.

Abdulaziz *et al.* [11] developed a stacking-based ensemble model for diabetes prediction. Their methodology integrated Random Forest (RF) and Logistic Regression (LR) as base learners, with XGBoost as the meta-learner, achieving 83% accuracy on the PIMA dataset.

Rashid *et al.* [12] introduced a Voting Classifier ensemble that combines decision trees (DT), logistic regression (LR), k-nearest neighbors (KNN), random forest (RF), and XGBoost. They applied advanced data preprocessing techniques, including standardization, missing value imputation, and anomaly detection using the Local Outlier Factor (LOF). Their ensemble approach reached 81% accuracy, demonstrating improved performance in sensitivity and specificity metrics.

Bhuvaneswari *et al.* [13] developed an advanced ensemble learning approach, achieving 88.89% accuracy on the PIMA dataset.

Talari *et al.* (2024) [14] employs SMOTE to balance class distributions and applies an ensemble model based on bagging with decision trees. This approach achieves an accuracy of 99.07% with an optimized execution time of 0.1 ms, outperforming other techniques in terms of recall and F1-score.

Similarly, Nagassou *et al.* (2023) [15] explores an alternative ensemble approach by combining LightGBM and CatBoost. While LightGBM is highly efficient, it is prone to overfitting; however, CatBoost compensates for this limitation by incorporating an overfitting detector and balanced predictors. To further enhance robustness, Bayesian hyperparameter optimization is employed, leading to an F1-score and accuracy of 99.37%.

Building upon the strengths of hybrid models, Taha *et al.* (2022) [16] presents a methodology that integrates fuzzy clustering (Fuzzy C-Means) with logistic regression. Instead of relying solely on traditional classifiers, their approach first trains six machine learning models and then utilizes a hybrid meta-classifier to group predictions into fuzzy clusters, which are subsequently refined by logistic regression. The results demonstrate 99.00% accuracy on PIDD and 95.20% on SDD, outperforming conventional models.

Their model combined Random Forest (RF), Radial Support Vector Machine (R-SVM), and K-Nearest Neighbors (KNN) in a Voting Classifier, optimizing classification robustness and improving the reliability of diabetes prediction.

These studies highlight the potential of ensemble learning approaches in diabetes diagnosis. However, many of these models still face challenges such as hyperparameter tuning, class imbalance, and feature selection. Our research aims to address these limitations by optimizing hyperparameters via GridSearch, balancing classes using SMOTEENN, and leveraging an improved ensemble model (ETC, XGBoost, and KNN). This study demonstrates how ensemble hybrid approaches can lead to more reliable, precise, and robust predictive models for early diabetes detection.

2. Materials and Methods

2.1. Dataset and Preprocessing

The dataset used in this study comes from the renowned Kaggle platform, which

provides publicly accessible datasets. To ensure optimal performance of the machine learning models, several data preprocessing steps were implemented.

First, the dataset consists of 768 samples exclusively from female patients. Among them, 268 are diagnosed as diabetic, while 500 are non-diabetic. This distribution highlights an imbalance between the classes, with non-diabetic patients being almost twice as represented as diabetic ones. Such class imbalance can significantly impact the performance of machine learning models, particularly those that prioritize overall accuracy over recall for the minority class (diabetic patients).

In addition to the class imbalance, another major challenge in this dataset is the presence of multiple missing or anomalous values. For instance, several features contain zero values, which are biologically implausible and likely indicate missing data. Specifically:

- 227 individuals have a skinfold thickness of zero
- 35 individuals have a diastolic blood pressure of zero
- 27 patients have a body mass index (BMI) of zero.

To address these issues, several preprocessing techniques were applied:

1) Class Balancing: The SMOTEENN method was used to improve the representation of the 268 diabetic samples while reducing the risk of overfitting associated with artificially adding samples.

2) Data Normalization: The variables were scaled using Standard Scaler, ensuring their compatibility with algorithms sensitive to variations in scale.

3) Outlier Handling: The MICE (Multiple Imputation by Chained Equations) **method is** an advanced statistical technique used to handle missing values in datasets.

These steps enhance the quality of the dataset, ultimately improving the performance of machine learning models.

2.2. Methods Used for Diabetes Prediction

Numerous machine learning methods have been developed for diabetes prediction, with varying performance across models. In this study, the XGBoost (XGB), Extra Trees Classifier (ETC), and K-Nearest Neighbors (K-NN) algorithms were selected for their efficiency and are analyzed and briefly presented.

✓ **XGBoost (XGB):**

XGBoost (XGB) is a widely used boosting model in supervised regression due to its efficiency in optimizing objective functions and improving prediction accuracy. It is based on ensemble learning, combining multiple models to generate a single prediction, making it a robust ensemble method. As noted by [17] [18], it integrates several complementary algorithms into a coherent model, thereby enhancing overall performance. Its approach consists of analyzing the residual errors of an initial model and then adjusting an additional model to better predict these residuals [17] [18]. Finally, XGBoost stands out from traditional Gradient Boosting (GB) methods by finding an optimal balance between bias and variance,

ensuring more robust and precise predictions [17] [18].

✓ **K-Nearest Neighbors (K-NN):**

K-Nearest Neighbors (K-NN) is an algorithm that captures the local structures of data and enhances ensemble model performance by leveraging specific relationships between observations. It is commonly used for diabetes prediction, although its effectiveness depends on parameter selection and data preprocessing [19]. As highlighted by Karyono, G., selecting an optimal K-value is essential since excessively high values can reduce the model's efficiency. Additionally, Kandhasamy *et al.* compared KNN with other algorithms and emphasized the importance of handling noisy data to improve its overall performance [20]. Moreover, P. Sinha *et al.* demonstrated its effectiveness in other medical applications, particularly in comparison with models such as Support Vector Machines (SVM) [21].

✓ **Extra Trees Classifier (ETC):**

Extra Trees Classifier (ETC) enhances the diversity and stability of predictions through increased randomization while reducing the risk of overfitting. It is characterized by a highly random node-splitting process, where attributes and split points are randomly selected, which, in extreme cases, can generate trees that are entirely independent of the output values in the training sample [22] [23].

The synergistic integration of ETC, XGBoost, and KNN is the cornerstone of their collective efficiency within our hybrid model. Each algorithm contributes a distinct perspective to diabetes prediction: Extremely Randomized Trees (ETC) excel at capturing non-linear patterns through high randomness, XGBoost systematically improves accuracy by correcting misclassified instances, and KNN enhances the model by identifying local similarities between patients, which traditional hierarchical approaches may overlook.

This algorithmic diversity leads to complementary errors, enabling ensemble voting to mitigate individual weaknesses. In the context of diabetes detection, this synergy ensures better coverage of the problem by simultaneously considering complex risk factor interactions, ambiguous borderline cases, and atypical patient profiles. As a result, the model delivers greater robustness, improved accuracy, and reduced susceptibility to overfitting.

2.3. Proposed Diabetes Detection Approach

The dataset used in this study comes from the well-known Kaggle platform, which hosts publicly accessible databases. To enhance the efficiency of machine learning models, several preprocessing steps were implemented, including class balancing using the SMOTEENN method, data normalization, and outlier handling. Once preprocessing was completed, the dataset was split into 80% for training and 20% for testing.

The proposed methodology for diabetes detection is based on an ensemble learning approach, combining three algorithms: K-Nearest Neighbors (K-NN), Extra Trees Classifier (ETC), and XGBoost (XGB).

The objective of this approach is to improve robustness and accuracy by aggre-

gating the predictions of the three models. Each algorithm has its own strengths and limitations, and their combination often leads to better performance, as illustrated in Figure 1.

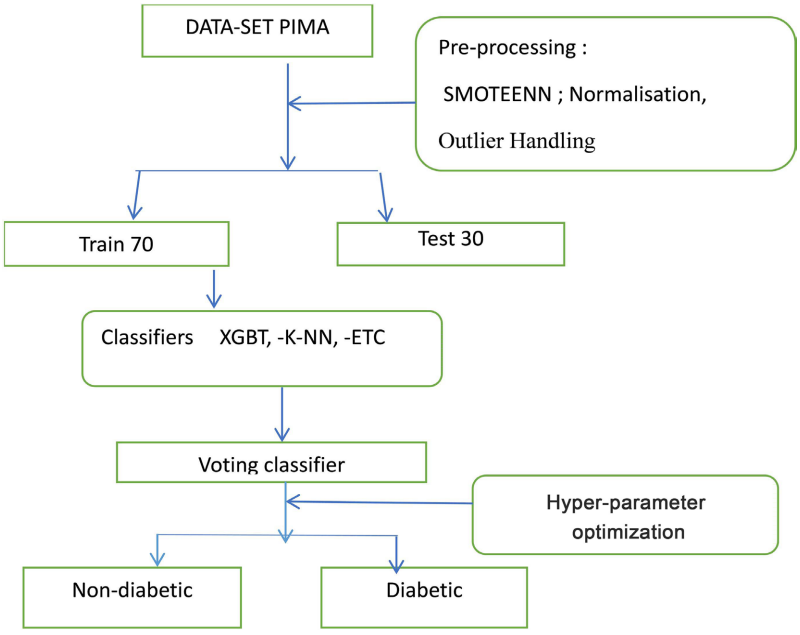

Figure 1. Proposed methodology for diabetes detection.

In this approach, the three models are trained independently on the same dataset, with each model generating prediction probabilities for each class of the target variable. These probabilities are then combined to produce a final prediction.

This fusion is performed by averaging the predicted probabilities, while assigning weight to each model based on its performance on a validation set.

Thus, the most effective models have a greater influence on the final decision, improving the model's robustness and reducing the risk of overfitting.

The class with the highest vote count is then selected by averaging the probability scores from the combined predictions of all classifiers in the ensemble model (Figure 2).

After calculating the average predicted probabilities from the K-NN, ETC, and XGB models, the class with the highest probability score is chosen as the final prediction. This approach ensures that the ensemble model makes an informed decision by leveraging the strengths of each classifier effectively.

3. Results and Analysis

3.1. Results and Discussion

✓ **Handling Missing (Multiple Imputation by Chained Equations)**

The comparative analysis of imputation methods highlights MICE as the most effective approach (Table 1), achieving the highest accuracy (95.50%), precision

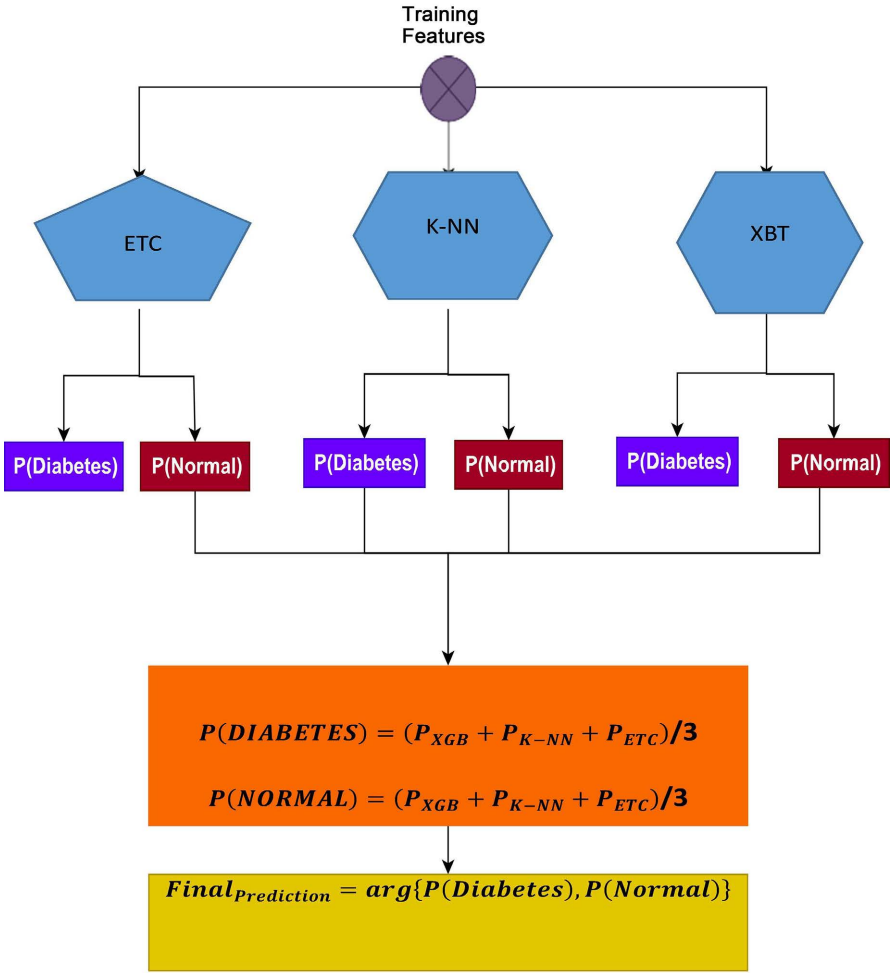

Figure 2. Architecture of the prosed voting classifier.

(93.22%), and recall (98.21%). k-NN performs moderately well, with an accuracy of 94.82%, but its recall is lower than that of MICE. Mean imputation, while simple to implement, shows the weakest performance, with an accuracy of 93.69% and a precision of only 90.16%. MICE stands out for its optimal balance between precision and recall, making it the best choice for handling missing values in diabetes prediction. This confirms that MICE is the most reliable method, offering superior performance across all key evaluation metrics.

Table 1. Performance comparison of imputation methods for diabetes prediction.

IMPUTATION METHOD	ACCURACY	PRECISION	RECALL	F1-SCOORE	AUC-ROC
Average	93.69	90.16	98.21	95.65	98.83
k-NN	94.82	92.58	94.51	94.83	95.80
MICE	**95.50**	**93.22**	**98.21**	**95.65**	**98.83**

✓ **Ablation study to demonstrate the synergy of algorithms**

The analysis of the performance table highlights the superiority of the ETC +

XGBoost + K-NN combination, which achieves the best accuracy (95.50%), precision (93.22%), and recall (98.21%). This ensemble approach clearly outperforms all other tested methods, reinforcing the effectiveness of combining multiple classifiers.

A comparison between combined methods and individual classifiers further confirms this trend. The three ensemble models (ETC + XGBoost + K-NN, ETC + XGBoost, and ETC + K-NN) consistently surpass individual classifiers, demonstrating the benefits of ensemble learning. Notably, the full combination of three algorithms results in a +3.7% accuracy gain over the best two-algorithm combination, illustrating the advantage of leveraging diverse model capabilities.

When examining individual classifiers, ETC (Extra Trees Classifier) emerges as the best standalone model, achieving 89% accuracy with high precision (92.00%). XGBoost also performs well with 82% accuracy and a good precision recall balance of 84.50%/88.00%, while K-NN, despite having the lowest accuracy (80.34%), maintains a relatively high recall (88.40%).

In terms of precision-recall balance, ETC + XGBoost + KNN not only achieves the highest accuracy but also maintains the best tradeoff between precision and recall. ETC alone, while precise (92.00%), has a lower recall (87%), suggesting a slight imbalance. See Figure 3 for visual comparison.

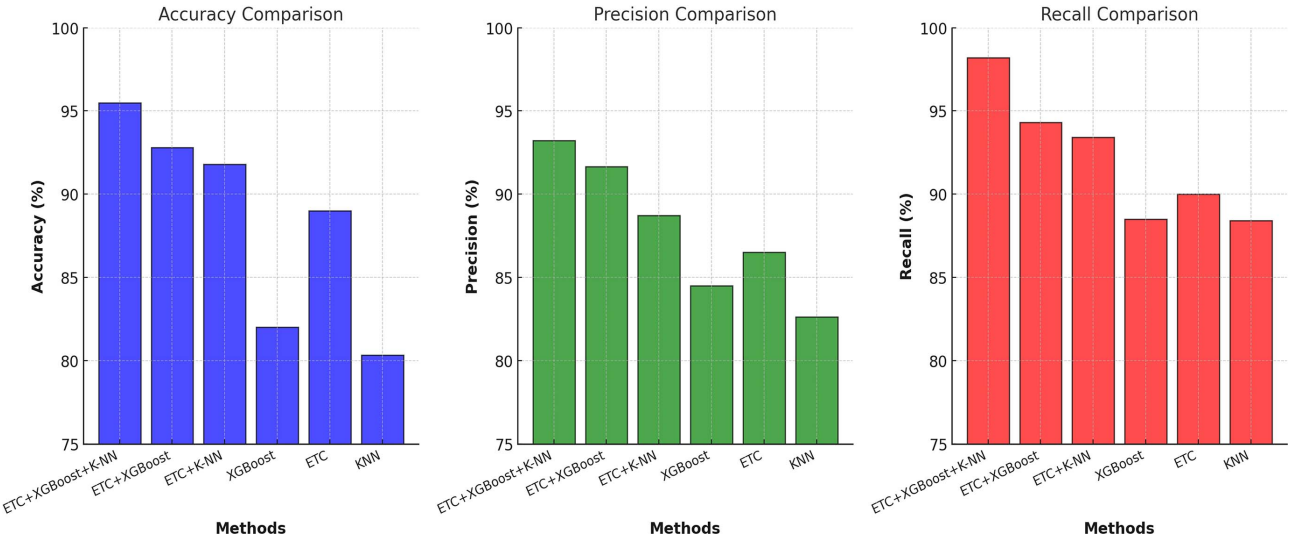

Figure 3. Comparative analysis of machine learning models for diabetes prediction.

✓ **Hyperparameter Analysis and Model Performance**

The hyperparameters chosen for the ensemble model combining Extra Trees (ET), k-Nearest Neighbors (k-NN), and XGBoost (XGB) are well-tuned, ensuring an optimal balance between robustness, performance, and generalization capability.

Analysis of Optimal Parameters

The hyperparameters chosen for the ensemble model combining Extra Trees (ET), k-Nearest Neighbors (k-NN), and XGBoost (XGB) are well-tuned, ensuring

an optimal balance between robustness, performance, and generalization capability.

KNN (K-Nearest Neighbors)

n_neighbors = 5: Provides an optimal balance between bias and variance.

weights = "distance": Gives more influence to closer neighbors compared to distant ones.

p = 2: Uses Euclidean distance (L2 norm), well-suited for continuous feature spaces.

XGBoost

n_estimators = 300: A high number of trees, ensuring a robust model that mitigates overfitting.

max_depth = 6: A moderate depth, balancing complexity and generalization.

learning_rate = 0.1332: A moderately low learning rate that ensures stable convergence.

subsample = 0.8057: Samples approximately 81% of the data for each tree, reducing overfitting.

colsample_bytree = 0.7846: Uses about 78% of features per tree, promoting diversity in decision boundaries.

ETC (Extra Trees Classifier)

n_estimators = 50: A moderate number of trees, sufficient for this model.

max_depth = 12: A relatively deep structure to capture complex relationships.

min_samples_split = 4: A reasonable threshold before splitting a node.

min_samples_leaf = 1: Allows leaves to contain a single sample, ensuring high precision.

max_features = "log2": Considers log2 (n_features) features per split, increasing randomness and generalization.

The superiority of the ETC + XGBoost + KNN combination is attributed to the complementarity of these three approaches:

- KNN excels in regions where classes are clearly separable and captures local structures within the data.
- XGBoost is highly effective at modeling complex, non-linear relationships while efficiently handling outliers.
- Extra Trees introduces additional randomness, promoting generalization and reducing variance.

This synergy results in a well-balanced model, leveraging KNN's local adaptability, XGBoost's structured learning, and Extra Trees' randomness-driven robustness, ultimately leading to superior predictive performance.

✓ **Model Performance Evaluation**

Superiority of the Ensemble Model

Compared to other models, the proposed ensemble model stands out significantly, achieving an impressive accuracy of 95.50%, a precision of 93.22%, and a recall of 98.21%. This high recall ensures excellent detection of positive cases while maintaining a good balance with precision, minimizing both false positives and

false negatives.

✓ **Precision-Recall Curve Analysis**

The Precision-Recall curve shows that all three models exhibit outstanding performance, with curves close to 1, demonstrating their effectiveness in classification. ETC + XGBoost + KNN stands out by maintaining high precision even at high recall levels, slightly surpassing ETC + XGBoost, while ETC + KNN shows a slight drop in precision at higher recall values. Thus, ETC + XGBoost + KNN emerges as the most robust and generalizable solution, ensuring reliable and optimized classification for diabetes detection. See Figure 4.

✓ **AUC-ROC Score and Model Robustness**

With an AUC-ROC score of 98.83%, the model demonstrates excellent class separation, further reinforcing its robustness.

✓ **Interpretation of the Confusion Matrix Results**

The ensemble model (ETC + XGBoost + KNN) exhibits a very low false negative rate, missing only one diabetes case out of 56, which is a major advantage for medical applications. However, the presence of 4 false positives out of 55 negatives indicates that some healthy individuals might be misclassified as diabetic, potentially leading to unnecessary medical tests. Despite this, the optimal balance between precision and recall ensures reliable classification, minimizing errors while effectively detecting diabetic patients. This trade-off between safety and accuracy makes this model a robust and effective solution for diabetes detection. See Figure 5.

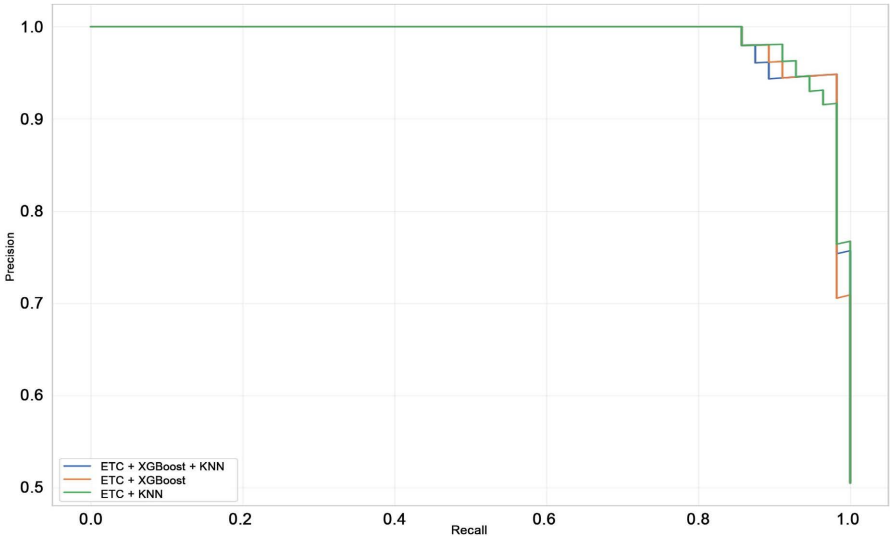

Figure 4. Precision-Recall curve.

3.2. Analysis and Comparison of the Performance of the Four Models

Furthermore, Lei Qin [10] explored an ensemble method integrating multiple algorithms, including Logistic Regression (LR), KNN, Decision Trees (DT), Gaussian Naïve Bayes, and SVM, achieving an accuracy of 81.6%. Despite these results,

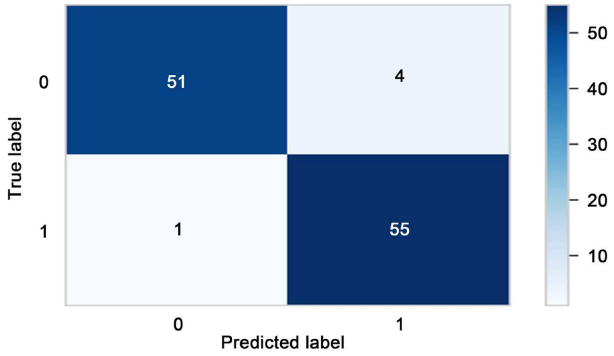

Figure 5. Confusion matrix.

the absence of optimal hyperparameter tuning and the limited dataset size prevented the achievement of optimal performance.

Additionally, Kumari *et al.* [11] proposed a weighted voting approach combining RF, LR, and NB, with an accuracy of 79.04%. However, the omission of cross-validation, a key element in assessing the robustness of a model, limits its reliability and potential for improvement.

On the other hand, Abdulaziz *et al.* [12] designed an ensemble approach combining RF and LR as base learners and XGBoost as a meta-learner, achieving 83% accuracy on the Pima dataset. While this method proves effective for diabetes prediction, further improvements remain possible.

Similarly, Rashid *et al.* [13] developed a voting ensemble approach, combining five algorithms (DT, LR, KNN, RF, and XGBoost) and incorporating an advanced preprocessing step (standardization, data imputation, and anomaly removal via the Local Outlier Factor (LOF)). With an accuracy of 81%, this approach stands out by evaluating metrics such as sensitivity and specificity, surpassing some previous methods.

As shown in **Table 2** and **Figure 6**, and in comparison, with these studies [7]-[12], our ensemble model clearly outperforms them in diabetes detection. Indeed, our methodology relies on hyperparameter optimization via GridSearchCV, while leveraging a balanced dataset, ensuring better model generalization.

This advancement contributes to improving diabetes diagnostic tools, reinforcing the importance of hybrid ensemble approaches in the medical field for more accurate and reliable predictions.

Table 2. Performance comparison with state-art-the-art studies.

Ref.	Year	Technique	Dataset	Accuracy
[7]	2023	Ensemble stacking approach (DT, NB, multilayer perceptron, SVM, and KNN)	Pima	81.9%
[8]	2022	Multilayer perception, GridSearchCV	Pima	89.30%
[9]	2022	Ensemble stacking approach (LR, KNN, DT, Gaussian Naive Bayes, and SVM)	Pima	82%

Continued

[10]	2021	Ensemble soft voting approach (RF, LR, and NB)	Pima	79.04%
[11]	2023	Ensemble stacking approach (LR, RF, XGboost, GridSearchCV, Cross-validation)	Pima	83%
[12]	2024	Ensemble soft voting approach (DT, LR, KNN, RF, XGBoost)	Pima	81%
Our proposed model	2025	Ensemble soft voting approach (ETC, XGBoost, KNN)	Pima	95.50%

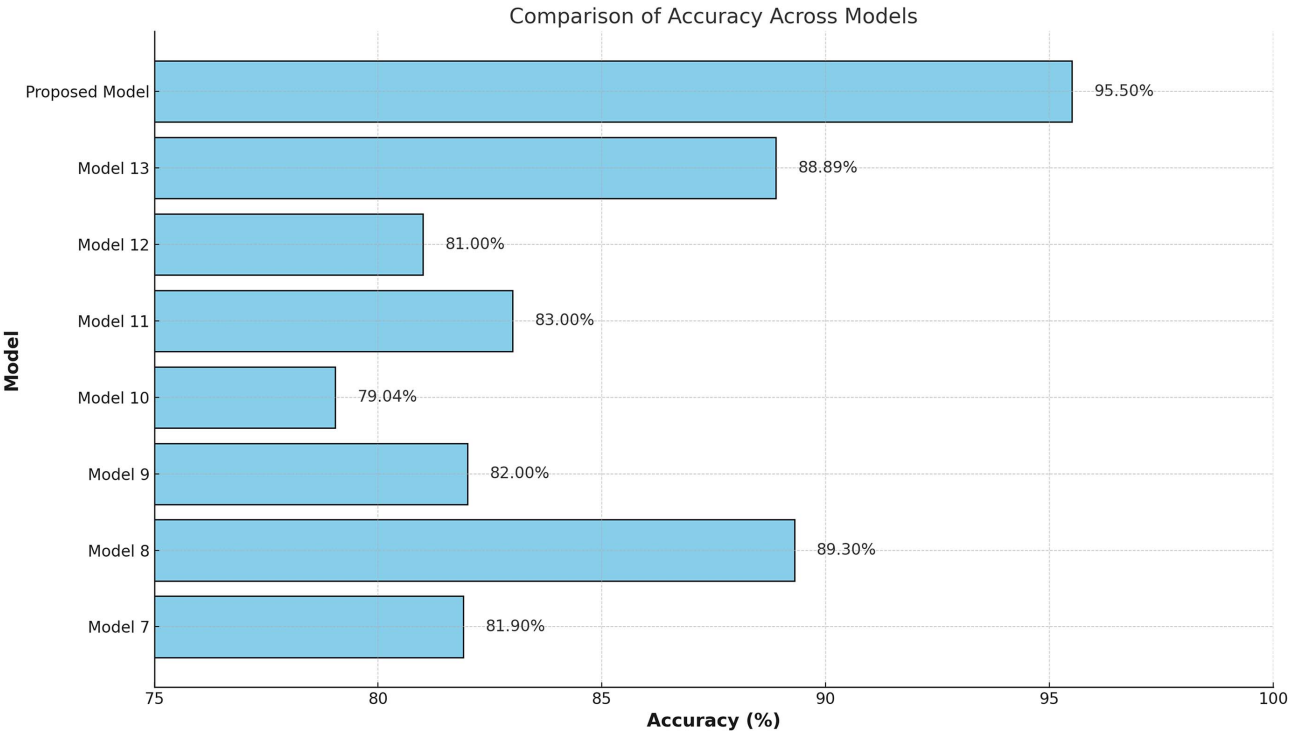

Figure 6. Performance comparison with state-art-the-art studies.

4. Conclusions

This study highlights the effectiveness of a hybrid ensemble model combining Extra Trees Classifier, XGBoost, and k-Nearest Neighbors (k-NN) for early and accurate detection of type 2 diabetes. By employing GridSearch for hyperparameter tuning and SMOTEENN for class balancing, the proposed model achieves remarkable performance, with an accuracy of 95.50%, a recall of 98.21%, and an AUC-ROC of 98.83%, outperforming individual models and existing approaches.

Despite these high-performance levels, a key challenge in medical diagnosis is the integration of multimodal data from various sources, including physiological signals, medical imaging, electronic health records, and genetic data. Leveraging heterogeneous data could significantly enhance diagnostic reliability and personalize predictions by considering multiple dimensions of a patient's health.

For future research, the focus will be on incorporating multimodal data into the model using advanced techniques such as Deep Learning, Convolutional Neural Networks (CNNs) for medical image analysis, and Natural Language Processing (NLP) for electronic health records interpretation. The objective is to improve model generalization and develop a more precise, adaptive, and clinically relevant diabetes prediction system.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., *et al.* (2019) Global and Regional Diabetes Prevalence Estimates for 2019 and Projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th Edition. *Diabetes Research and Clinical Practice*, **157**, Article 107843. https://doi.org/10.1016/j.diabres.2019.107843

[2] World Health Organizations (2016) Global Report on diabètes. World Health Organization. https://apps.who.int/iris/handle/10665/204871

[3] Thammano, A. and Meengen, A. (2005) A New Evolutionary Neural Network Classifier. *Advances in Knowledge Discovery and Data Mining*, Hanoi, 18-20 May 2005, 249-255. https://doi.org/10.1007/11430919_31

[4] Cohen, S., Dagan, N., Cohen-Inger, N., Ofer, D. and Rokach, L. (2021) ICU Survival Prediction Incorporating Test-Time Augmentation to Improve the Accuracy of Ensemble-Based Models. *IEEE Access*, **9**, 91584-91592. https://doi.org/10.1109/access.2021.3091622

[5] Mushtaq, Z., Ramzan, M.F., Ali, S., Baseer, S., Samad, A. and Husnain, M. (2022) Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques. *Mobile Information Systems*, **2022**, Article ID: 6521532. https://doi.org/10.1155/2022/6521532

[6] Kibria, H.B., Nahiduzzaman, M., Goni, M.O.F., Ahsan, M. and Haider, J. (2022) An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI. *Sensors*, **22**, Article 7268. https://doi.org/10.3390/s22197268

[7] Patil, R.N., Rawandale, S., Rawandale, N., Rawandale, U. and Patil, S. (2023) An Efficient Stacking Based NSGA-II Approach for Predicting Type 2 Diabetes. *International Journal of Electrical and Computer Engineering*, **13**, 1015-1023. https://doi.org/10.11591/ijece.v13i1.pp1015-1023

[8] Bhopte, M. and Rai, M. (2022) Hybrid Deep Learning CNN-LSTM Model for Diabetes Prediction. *International Journal of Scientific Research*, **8**, 444-447.

[9] Qin, L. (2022) A Prediction Model of Diabetes Based on Ensemble Learning. *Proceedings of the* 2022 5*th International Conference on Artificial Intelligence and Pattern Recognition*, Xiamen, 23-25 September 2022, 45-51. https://doi.org/10.1145/3573942.3573949

[10] Kumari, S., Kumar, D. and Mittal, M. (2021) An Ensemble Approach for Classification and Prediction of Diabetes Mellitus Using Soft Voting Classifier. *International Journal of Cognitive Computing in Engineering*, **2**, 40-46. https://doi.org/10.1016/j.ijcce.2021.01.001

[11] Alzubaidi, A.A., Halawani, S.M. and Jarrah, M. (2023) Towards a Stacking Ensemble Model for Predicting Diabetes Mellitus Using Combination of Machine Learning Techniques. *International Journal of Advanced Computer Science and Applications*, **14**, 348-358. https://doi.org/10.14569/ijacsa.2023.0141236

[12] Rashid, M.M., Yaseen, O.M., Saeed, R.R. and Alasaady, M.T. (2024) An Improved Ensemble Machine Learning Approach for Diabetes Diagnosis. *Pertanika Journal of Science & Technology*, **33**, 1335-1350.

[13] Amma N.G., B. (2024) En-RfRsK: An Ensemble Machine Learning Technique for Prognostication of Diabetes Mellitus. *Egyptian Informatics Journal*, **25**, Article 100441. https://doi.org/10.1016/j.eij.2024.100441

[14] Talari, P., N, B., Kaur, G., Alshahrani, H., Al Reshan, M.S., Sulaiman, A., *et al.* (2024) Hybrid Feature Selection and Classification Technique for Early Prediction and Severity of Diabetes Type 2. *PLOS ONE*, **19**, e0292100. https://doi.org/10.1371/journal.pone.0292100

[15] Nagassou, M., Mwangi, R.W. and Nyarige, E. (2023) A Hybrid Ensemble Learning Approach Utilizing Light Gradient Boosting Machine and Category Boosting Model for Lifestyle-Based Prediction of Type-II Diabetes Mellitus. *Journal of Data Analysis and Information Processing*, **11**, 480-511. https://doi.org/10.4236/jdaip.2023.114025

[16] Altaher Taha, A. and Jameel Malebary, S. (2022) A Hybrid Meta-Classifier of Fuzzy Clustering and Logistic Regression for Diabetes Prediction. *Computers, Materials & Continua*, **71**, 6089-6105. https://doi.org/10.32604/cmc.2022.023848

[17] Kibria, H.B., Nahiduzzaman, M., Goni, M.O.F., Ahsan, M. and Haider, J. (2022) An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI. *Sensors*, **22**, Article 7268. https://doi.org/10.3390/s22197268

[18] Dutta, A., Hasan, M.K., Ahmad, M., Awal, M.A., Islam, M.A., Masud, M., *et al.* (2022) Early Prediction of Diabetes Using an Ensemble of Machine Learning Models. *International Journal of Environmental Research and Public Health*, **19**, Article 12378. https://doi.org/10.3390/ijerph191912378

[19] Sarker, I.H., Faruque, F., Alqahtani, H. and Kalim, A. (2018) K-Nearest Neighbor Learning Based Diabetes Mellitus Prediction and Analysis for eHealth Services. *EAI Endorsed Transactions on Scalable Information Systems*, **7**, e4. https://doi.org/10.4108/eai.13-7-2018.162737

[20] Sakagraha Kuspinta, N., Widodo, A.W. and Furqon, M.T. (2018) Penentuan Menu Makanan Untuk Penderita Diabetes Menggunakan Metode Iterative Dichotomizer Tree (ID3). https://j-ptiik.ub.ac.id/index.php/j-ptiik

[21] Tombokan, M., *et al.* (2017) Hubungan dukungan keluarga dengan motivasi dalam men-gontrol kadar gula darah pada penderita diabetes melitus di wilayah kerja puskesmas pampang kecamatan panakkukang kota makassar. *Journal Media Keperawatan: Politeknik Kesehatan Makassar*, **8**, 39-45.

[22] Rustam, F., Ashraf, I., Mehmood, A., Ullah, S. and Choi, G. (2019) Tweets Classification on the Base of Sentiments for US Airline Companies. *Entropy*, **21**, Article 1078. https://doi.org/10.3390/e21111078

[23] Safavian, S.R. and Landgrebe, D. (1991) A Survey of Decision Tree Classifier Methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, **21**, 660-674. https://doi.org/10.1109/21.97458