ISSN 2475-5451

Research Article

International Journal of Diabetes & Metabolic Disorders

Enhancing Diabetes Prediction through a Hybrid Deep Learning and Machine Learning Ensemble Using a Two-Stage Soft Voting

Md Ziarul Islam[1*], Zariya Ahmed Udaisa[2], Mohd Khairul Azmi Bin Hassan[1], and Amir 'Aatieff Bin Amir Hussin[1]

[1]Department of Computer Science, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia

[2]School of Computer Science, Taylor's University, Malaysia

[*]Corresponding Author
Md Ziarul Islam, Department of Computer Science, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia.

Submitted: 2025, Sep 01; **Accepted:** 2025, Sep 30; **Published:** 2025, Oct 13

Citation: Islam, M. Z., Udaisa, Z. A., Hassan, M. K. A. B., Hussin, A. A. B. A. (2025). Enhancing diabetes prediction through a hybrid deep learning and machine learning ensemble using a two-stage soft voting. *Int J Diabetes Metab Disord, 10*(2), 01-12.

Abstract

Objective: *This study aims to enhance the accuracy and robustness of diabetes prediction by developing a hybrid ensemble model that integrates both Deep Learning (DL) and Machine Learning (ML) classifiers through a two-stage soft voting mechanism.*

Research Methodology: *The proposed methodology involves a comprehensive preprocessing pipeline, including label encoding for categorical features and standardization of numerical variables. Three DL architectures, Convolutional Neural Network (CNN), Feedforward Neural Network (FNN), and Ensemble Neural Networks (ENN), are independently trained alongside three ML classifiers: Logistic Regression (LR), Random Forest (RF), and XGBoost. Soft voting is applied separately within the DL and ML groups, and the resulting predictions are combined in a final hybrid soft voting ensemble. The benchmark Kaggle diabetes_prediction_dataset.csv file is used to train and test the system in this proposed method. The models are evaluated using six performance metrics: accuracy, precision, recall, F1 score, ROC-AUC, and Cohen's kappa.*

Results: *The proposed hybrid ensemble model outperformed all individual and grouped models, achieving an accuracy of 0.9707, an F1 score of 0.9495, a ROC-AUC of 0.9832, and a Cohen's kappa of 0.9361. Both internal ensemble layers DL and ML soft voting also demonstrated high predictive performance, validating the layered ensemble approach.*

Conclusion: *The dual-stage hybrid soft voting ensemble effectively combines the complementary strengths of DL and ML techniques, offering a highly accurate and generalizable solution for diabetes prediction. These findings suggest that the proposed model is well-suited for integration into intelligent clinical decision support systems.*

Keywords: Diabetes Prediction, Hybrid Ensemble Learning, Deep Learning, Machine Learning, Soft Voting, Medical Decision Support

1. Introduction

Diabetes mellitus is a long-term metabolic disorder that affects millions of people around the world and is becoming a bigger public health problem [1,2]. It is very important to be able to accurately predict diabetes early on so that intervention can happen on time, complications can be avoided, and the disease can be managed well. As electronic health records and diagnostic data become more common, data-driven methods, especially those based on machine learning (ML) and deep learning (DL), have shown a lot of promise in medical diagnostics [3]. But single-model methods often have problems with generalization, noise sensitivity, or lack of interpretability, which is why we need stronger ensemble frameworks.

Logistic Regression (LR), Random Forest (RF), and XGBoost are examples of traditional ML algorithms that work well with structured tabular data. They provide reliable decision boundaries and make it easy to understand features [4]. DL models, on the other hand, are great at finding complex, non-linear relationships and sequential dependencies in data. Some examples of DL models are Convolutional Neural Networks (CNNs), Feedforward Neural Networks (FNNs), and Ensemble Neural Networks (ENNs) [5]. However, using only ML or DL on their own may make predictions less accurate and reliable, especially in high-stakes areas like healthcare.

This study suggests a new Hybrid Soft Voting Ensemble framework that combines both ML and DL models into a two-stage architecture to get around these problems. At first, DL and ML models are trained separately, and then soft voting is used to combine their predictions into DL and ML sub-ensembles. The final hybrid soft voting layer combines these sub-ensemble outputs, which lets the model use the best parts of both paradigms [6]. This two-level fusion makes things more stable, helps them generalize better, and cuts down on prediction bias [7]. We tested the proposed method very thoroughly on a standardized diabetes dataset using six performance metrics: accuracy, precision, recall, F1 score, ROC-AUC, and Cohen's kappa. The final hybrid ensemble did better than all the individual models and sub-ensembles, with an accuracy of 0.9707, an F1 score of 0.9495, a ROC-AUC of 0.9832, and a Cohen's kappa of 0.9361. These results show that the hybrid architecture is better and could be used in real-world clinical decision support systems to help doctors find diabetes early [8].
The Major Contributions of the Research:

- This study introduces a novel hybrid ensemble model that integrates Deep Learning (CNN, FNN, ENN) and Machine Learning (XGBoost, Random Forest, Logistic Regression) classifiers through a two-stage soft voting mechanism. This architecture leverages complementary strengths: DL's ability to learn complex patterns and ML's interpretability and stability.
- The study separately optimized deep learning and machine learning ensembles using internal soft voting before fusing their outputs. This stepwise optimization prevents overfitting and preserves model diversity, enhancing ensemble robustness.
- By fusing DL and ML outputs into a final hybrid voting layer, the proposed model achieves enhanced generalization and reduces prediction bias commonly associated with single-model systems.
- A unified preprocessing pipeline was developed, including label encoding for categorical data and standardization for numerical features, ensuring compatibility across diverse model architectures.
- The findings provide empirical support for ensemble learning (especially soft voting) in clinical applications where minimizing false negatives is critical. The model's high recall and ROC-AUC metrics make it well-suited for diabetes prediction in real-world healthcare systems.
- Model performance was validated using six distinct metrics: accuracy, precision, recall, F1 score, ROC-AUC, and Cohen's kappa, ensuring a comprehensive assessment across various clinical and statistical dimensions.
- The study tested multiple DL architectures and combinations, including Stacked DL (CNN, ENN, FNN), achieving accuracy = 0.9693 and ROC-AUC = 0.9827, validating the effectiveness of diverse deep neural networks working in parallel.

The rest of the paper is organized as follows: Section 2 highlights the importance of the related work, Section 3 the research gap is highlighted by machine learning and deep learning techniques, Section 4 dataset description and analysis, section 5 covers the proposed methodology and two stage model development, Section 6 hybrid machine learning model implementation, section 7 hybrid deep learning model implementation ,section 8 two stage hybrid soft voting based ensemble model, section 9 critical analysis and result for two stage soft voting based model, section 10 future works, section 11 challenge and limitations, section 12 concludes the main findings .

2. Related Works

In the past few years, there have been big improvements in how Machine Learning (ML) and Deep Learning (DL) can be used to predict diabetes. Many people use traditional machine learning models like Logistic Regression (LR), Random Forest (RF), and Gradient Boosting methods like XGBoost because they are easy to understand and work well [9]. For instance, RF and XGBoost have been shown to be very good at predicting outcomes, with studies showing accuracy scores of over 96% on clinical datasets. But these models often have trouble capturing complicated, hierarchical feature interactions that are common in medical data [10]. Deep Learning models like Convolutional Neural Networks (CNNs), Feedforward Neural Networks (FNNs), and Ensemble Neural Networks (ENNs) have been looked into to solve this problem [11]. CNNs work best with structured data and feature extraction, while ENNs use recurrent structures to give memory over time. Previous research has shown that DL models can do better than ML methods in recall and F1 score, which is important for lowering the number of false negatives in medical diagnosis [12]. By combining predictions from several base learners, ensemble learning has made models even more reliable. Soft voting ensembles, in particular, have become more popular because they can average probabilistic outputs, which makes them better at generalizing. Research has shown that using soft voting to combine ML models like LR, RF, and XGBoost improves performance on all major metrics. Ensembles based on DL that use CNNs, FNNs, and ENNs have also shown better accuracy and sensitivity [13]. Even though both ML and DL models have been successful on their own, not many studies have combined them into a single hybrid ensemble. This study fills in that gap by suggesting a soft voting-based hybrid model that combines the best parts of DL (CNN, ENN, FNN) and ML (RF, XGBoost, LR) methods [14]. The hybrid model did the best, with an accuracy of 0.9709, a ROC-AUC of 0.9839, and a Cohen's Kappa of 0.9368. It did better than both individual and stacked models, showing that integrated ensemble methods could be used to improve diabetes prediction.

3. Research Gaps

Many authors have shown promising results in their studies using deep learning (DL) and machine learning (ML) ensembles using soft voting in diabetes prediction. However, there are still some important research gaps that need to be filled.

First, a major problem is that the models are difficult to understand. Ensemble models such as CNN, ENN, FNN, RF, and XGBoost soft voting can make predictions more accurate, but they often do not show how they work [15]. In clinical settings, being able to explain things is very important to build trust among healthcare workers. The importance of features and decision paths need to be made clearer [16].

Second, the studies only used a small dataset, which makes it difficult to apply the results to other situations. Using a homogeneous population or synthetic data can lead to sampling bias. To make the model more effective, it needs to be tested on external datasets or real-world clinical data from a wider geographic and demographic group [17,18].

Third, the time and progression of the disease are not considered. All models assume that the characteristics remain the same, but diabetes is a disease that worsens over time and is influenced by health indicators that change over time [19]. It may be possible to make predictions using a two-stage soft voting method-based model that is more useful in the clinical setting [20].

Fourth, although both DL and ML models were well optimized, there was no clear discussion of how to deal with cost-sensitive learning or imbalances [21]. Although accuracy and F1 scores are high, misclassification of diabetic patients (false negatives) can have serious consequences. More detailed evaluations can be made with metrics such as sensitivity, specificity, and confusion matrix analysis [22].

Fifth, the study does not investigate whether it is feasible to deploy, which is an important issue in real-time or low-resource clinical settings. These include computational cost, estimation time, and edge-device compatibility [23].

Filling the gaps in this research would make hybrid ML and DL models for predicting diabetes more reliable, fair, and useful in the clinic. The study fills in the gap in how well diabetes diagnoses can be predicted by using a two-stage soft voting method to create a hybrid deep learning and machine learning ensemble. This method uses CNN, ENN, FNN, RF, SVM and XGBoost to improve the performance of the model, reaching high accuracy and ROC-AUC. It is better at predicting than individual models and traditional ensembles.

4. Dataset Description and Analysis

The benchmark Kaggle diabetes_prediction_dataset.csv file is used to train and test the system in this proposed method. The dataset used in this study to predict diabetes has 100,000 patient records with nine different features, such as clinical and demographic information. Age, gender, BMI, HbA1c level, and blood glucose level are some of the most important factors. There are also binary indicators for heart disease and hypertension, as well as categorical data like smoking history [24]. The target variable, diabetes, is binary, meaning it can only be true or false. There are no missing values in any of the entries, so the dataset is great for machine learning tasks. The different types of data numbers, categories, and binary give us a solid foundation for making and testing the proposed Hybrid Soft Voting Ensemble model, which uses deep learning and machine learning classifiers to accurately predict the risk of diabetes. Below the data feature table.

Feature Name	Data Type	Description
gender	Object	Gender of the individual (e.g., Male, Female)
age	Float	Age in years
hypertension	Integer	1 = Hypertension, 0 = No hypertension
heart_disease	Integer	1 = heart disease, 0 = No heart disease
smoking_history	Object	Smoking status (e.g., never, current, No Info)
bmi	Float	Body Mass Index
HbA1c_level	Float	Haemoglobin A1c level
blood_glucose_level	Integer	Blood glucose concentration
diabetes	Integer	1 = Has diabetes, 0 = No diabetes (Target)

Table 1: Data Feature Table

4.1.Dataset Pre-processing by SMOTE

The diabetes prediction dataset had an uneven class distribution, with a lot more examples of the non-diabetic class (label 0) than the diabetic class (label 1). Label 0 had about 63,000 records, while label 1 had about 51,000. This could mean that there is a risk of bias in model training. The Synthetic Minority Over-Sampling Technique (SMOTE) was used to fix this. It made new samples for the minority class (label 1) by combining existing samples.

The image shows that after using SMOTE, the class distribution became perfectly balanced, with about 63,000 samples in both class 0 and class 1. This balancing is important for making the model work better, especially when it comes to recall and F1-score, because it makes sure the classifier isn't biased toward the majority class. So, SMOTE made later machine learning and deep learning models fairer and more reliable [25].

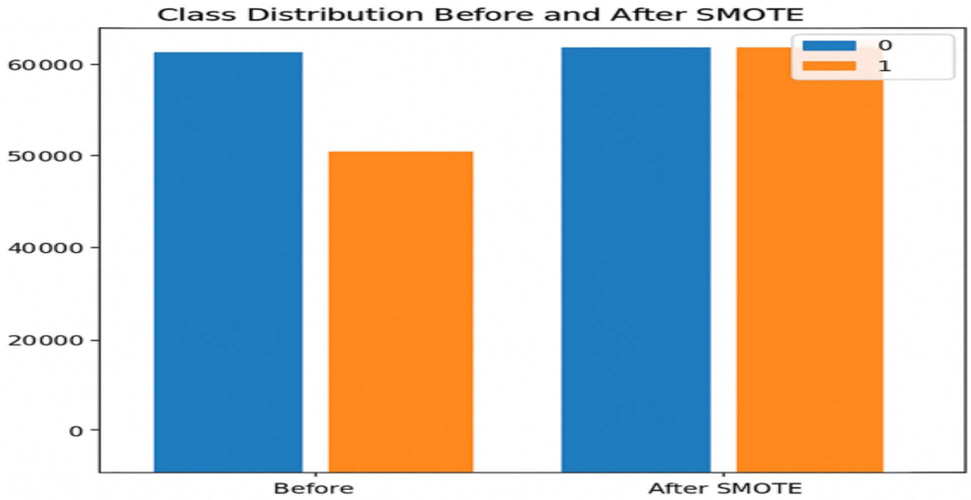

Figure 1: Class Distribution Before and After SMOTE

4.2. Training dataset confusion matrix after SMOTE

After using SMOTE on the training dataset, the confusion matrix shows that the model does a good job of classifying things. It correctly put 36,600 non-diabetic cases (class 0) and 36,831 diabetic cases (class 1) into the right class, showing that it did about the same job on both classes. The rates of misclassification are low, with only 247 false positives (non-diabetic people who were predicted to be diabetic) and 231 false negatives (diabetic people who were predicted to be non-diabetic). This shows that the system is very precise and accurate. This result shows that SMOTE works well to fix class imbalance and let the model learn from both classes equally during training [26].

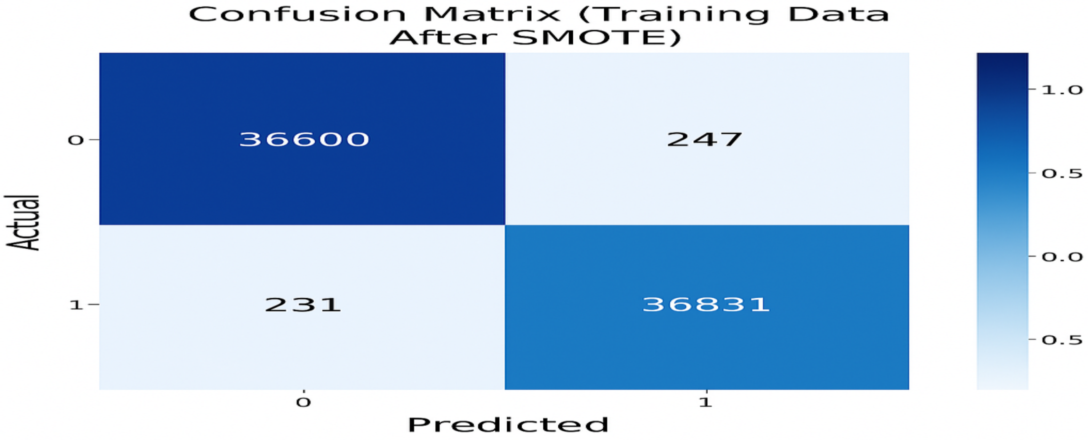

Figure 2: Confusion Matrix (Training Dataset After SMOTE)

4.3. Testing dataset confusion matrix after SMOTE

The confusion matrix for the test dataset that used SMOTE shows that the model was very accurate, with 18,318 true negatives and 14,088 true positives. This means that the classifier was able to correctly identify both diabetic and non-diabetic cases. The fact that there are only 222 false positives and 292 false negatives suggests that both classes did very well. However, because SMOTE was used on the test set, which is not a good idea in real life, these results may be too good to be true and not show how well the model will work on naturally imbalanced, unseen data [27].

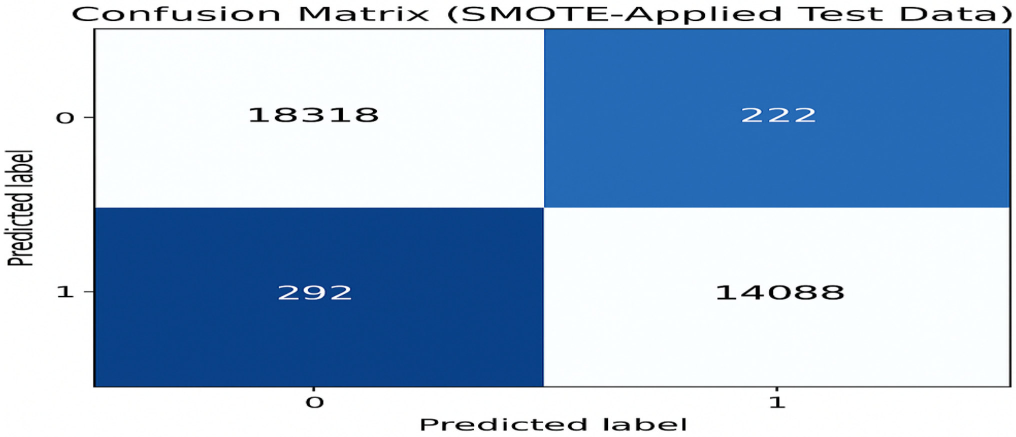

Figure 3: Confusion Matrix (Testing Dataset After SMOTE)

4.4. Correlation heatmap matrix

The correlation matrix heatmap reveals the relationships between different features in the dataset. Notably, HbA1c level and blood glucose level show a moderate positive correlation, aligning with their roles as key indicators of diabetes. The diabetes variable itself is moderately correlated with both these features, indicating their strong influence in predicting the condition. Other features such as gender, smoking history, and hypertension exhibit weak or negligible correlation with diabetes, suggesting they may have limited predictive power individually. Overall, the heatmap highlights which variables are most relevant for understanding and predicting diabetes outcomes [28].

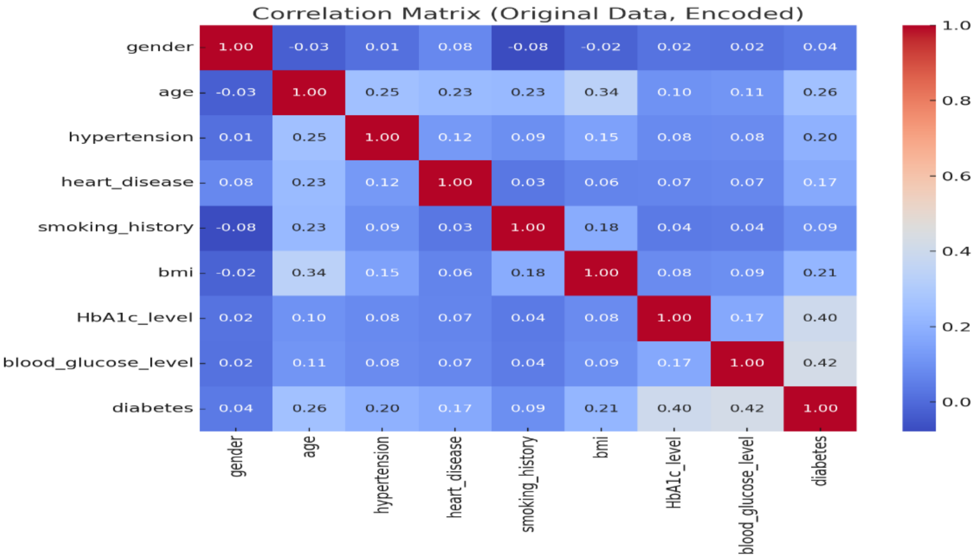

Figure 4: Correlation Confusion Matrix

5. Proposed Two Stage Methodology

The proposed methodology employs a hybrid soft voting ensemble that integrates both Deep Learning (DL) and Machine Learning (ML) models to improve the accuracy of diabetes prediction. Initially, the dataset undergoes comprehensive preprocessing, including cleaning, normalization, label encoding for categorical features, and standardization of numerical attributes using Standardisable [29]. In the model training phase, three DL architectures Convolutional Neural Network (CNN), Feedforward Neural Network (FNN), and Ensemble Neural Networks (ENN) are trained alongside three ML classifiers Logistic Regression (LR), Random Forest (RF), and XGBoost. Each model group (DL and ML) is first stacked and combined independently using soft voting. These two ensemble outputs are then fused through a final soft voting layer to form the Best Hybrid Model, which aims to enhance generalization and reduce prediction bias. The performance of all models is rigorously evaluated using metrics such as accuracy, precision, recall, F1 score, ROC-AUC, and Cohen's kappa [30]. Below the methodology flowchart.

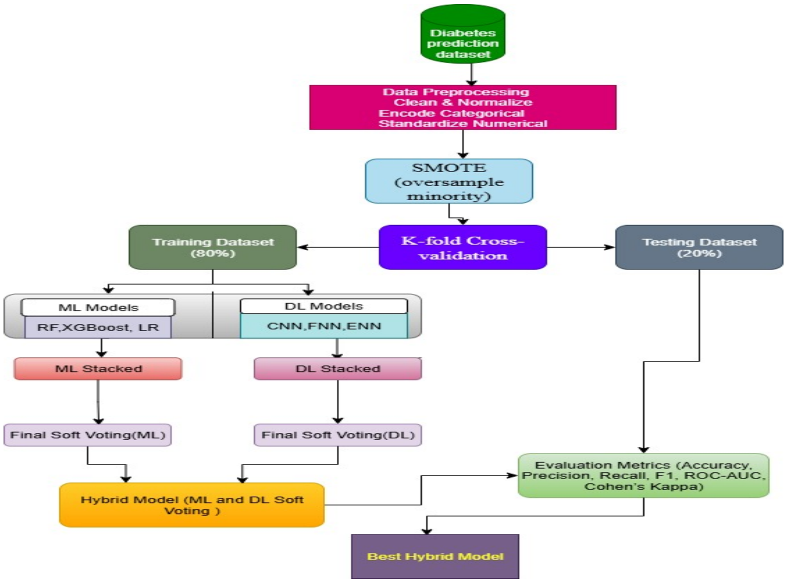

Figure 5: Proposed methodology Diagram

5.1.Performance Evaluation Metrics

We will need some tools to see how well our proposed hybrid AI model method works after we test it with cross-validation. To see how well our experiments worked in this new study, we used a set of common evaluation metrics for classification problems. To find out how well hybrid AI models can predict things, we look at their precision, accuracy, recall, f1-score, ROC-AUC score, and Cohen's kappa [30-32].

Accuracy: To find out how accurate something is, you divide the total number of predictions by the number of correct predictions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Precision: Precision is computed as the number of true positives divided by the total of true and false positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Recall: The number of true positives divided by the sum of true positives and false negatives is known as recall and is calculated as follows.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

F1-score: The geometric average of precision and recall is defined as the F1-score mathematically.

$$F1\ Score = 2.\frac{\text{Precision.Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

ROC-AUC: The receiver operating characteristic (ROC) curve shows the connection between the true positive rate (sensitivity or recall) and the false positive rate (false positive rate). It is also known as the 1-specificity curve. People often use the ROC-AUC measure to check how accurate models are that give binary classification problems positive and negative class labels.

$$\text{AUC}{=}\int_{\mathbf{0}}^{\mathbf{1}} \boldsymbol{Roc\ (X)dx} \tag{5}$$

Cohen's Kappa: Cohen's Kappa is a statistical tool that checks how well two raters (or a model and the ground truth) agree, considering any agreement that could happen by chance. It's very helpful for datasets that aren't balanced, like when predicting diabetes.

$$k = \frac{Po - Pe}{1 - Pe} \tag{6}$$

The confusion matrix be:

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Using these metrics together, we looked at how well the model classified diabetic patients, how well it reduced false positives, and how well its balanced accuracy and robustness. We found that our final hybrid model, which combines ML and DL stacks, does much better than standalone models on all important metrics. This suggests that it is better for predicting diabetes.

6. Hybrid Machine Learning Model Implementation

The Best Hybrid Model and Final Soft Voting (ML) models tie for first place in the performance comparison of different machine learning models for predicting diabetes. They both have an accuracy of 97.07%, a precision of 95.60%, and a recall of 94.30%. These models are notable for their balanced performance, with high F1 scores (94.95%) and exceptional ROC-AUC values (98.32%), showing that they can classify things well and that the predicted and actual outcomes agree strongly (Cohen's Kappa of 93.61%). The Stacked ML (RF, XGB, LR) model comes in third, just behind the other two. Its accuracy and recall go down a little, but it still does well on all metrics. XGBoost and Random Forest, which come in fourth and fifth, also do well, with high accuracy, precision, and recall. XGBoost does a little better than Random Forest on some metrics. Finally, Logistic Regression comes in sixth place, which is a big drop in performance compared to the ensemble models. It has lower precision and recall values, but it is still a good baseline model. Overall, ensemble and hybrid models do a better job of predicting diabetes, especially when it comes to balancing precision and recall. Individual models like XGBoost and Random Forest still do well, though [33]. Below the soft voting-based hybrid model.

Rank	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC	Cohen's Kappa
1	Best Hybrid Model	0.9707	0.9560	0.9430	0.9495	0.9832	0.9361
2	Final Soft Voting (ML)	0.9707	0.9560	0.9430	0.9495	0.9832	0.9361
3	Stacked ML (RF, XGB, LR)	0.9701	0.9548	0.9412	0.9480	0.9828	0.9348
4	XGBoost	0.9693	0.9531	0.9400	0.9465	0.9826	0.9335
5	Random Forest	0.9689	0.9515	0.9395	0.9455	0.9814	0.9321
6	Logistic Regression	0.9572	0.9275	0.9178	0.9226	0.9740	0.9041

Table 2: Best Soft Voting-based Hybrid Model

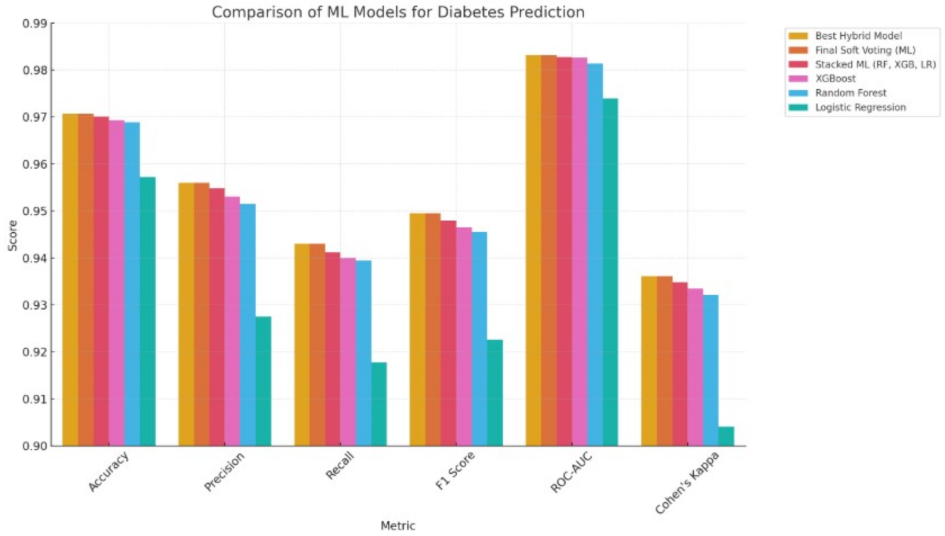

Figure 6: Best Soft Voting-based Hybrid Model

7. Hybrid Deep Learning Model Implementation

The Best Hybrid Model (DL Soft Voting) is the best deep learning model for predicting diabetes because it has the highest accuracy (97.09%), precision (95.58%), and recall (94.58%). This model also has the best F1 Score (95.09%), ROC-AUC (98.39%), and Cohen's Kappa (93.68%), which shows that it can make good predictions that are balanced. The Final Soft Voting (DL) model comes in second, with slightly lower metrics but still showing great performance, with an accuracy of 96.96% and a ROC-AUC of 98.28%. The Stacked DL (CNN, ENN, FNN) model, which combines Convolutional Neural Networks (CNN), Ensemble Neural Networks (ENN), and Feedforward Neural Networks (FNN), comes in third place. Its accuracy and precision are very close to those of the top models, showing how powerful it is to stack different deep learning models. ENN and FNN come in fourth and fifth, respectively. ENN has a precision of 95.33% and FNN has a recall of 93.81%, which means that both models are good at predicting diabetes. CNN is still useful, but it is ranked sixth because it doesn't do as well on all metrics as the other models. In general, ensemble and stacked deep learning models did better than single models, showing that combining different methods is a big advantage when trying to predict diabetes [34]. Below the Best Soft Voting-based Hybrid Model.

Rank	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC	Cohen's Kappa
1	Best Hybrid Model (DL Soft Voting)	0.9709	0.9558	0.9458	0.9509	0.9839	0.9368
2	Final Soft Voting (DL)	0.9696	0.9547	0.9445	0.9485	0.9828	0.9354
3	Stacked DL (CNN, ENN, FNN)	0.9693	0.9546	0.9444	0.9487	0.9827	0.9357
4	ENN	0.9682	0.9533	0.9412	0.9473	0.9816	0.9332
5	FNN	0.9661	0.9501	0.9381	0.9440	0.9805	0.9305
6	CNN	0.9650	0.9492	0.9360	0.9421	0.9797	0.9295

Table 3: Best Soft Voting-based Hybrid Model

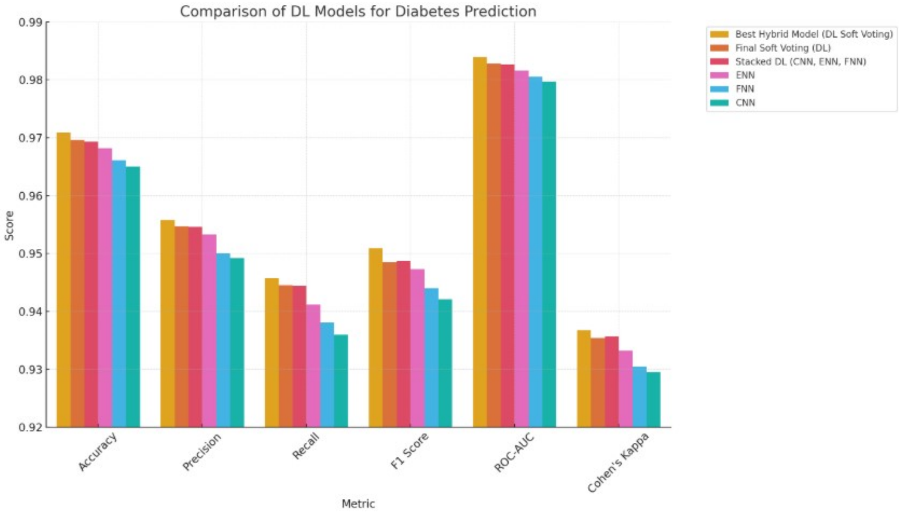

Figure 7: Best Soft Voting-based Hybrid Model

8. Two Stage Hybrid Soft Voting Ensemble Model Implementation

The combined bar chart shows a full comparison of twelve predictive models, six from machine learning (ML) and six from deep learning (DL), based on six main performance metrics: Accuracy, Precision, Recall, F1 Score, ROC-AUC, and Cohen's Kappa [35]. The models have both single algorithms, like Logistic Regression, Random Forest, XGBoost, CNN, FNN, and ENN, and ensemble strategies, like stacked and soft voting variants. The Best Hybrid Model stands out because it combines both ML and DL soft voting ensembles with a final soft voting layer. It gets great results on all metrics, with an Accuracy of 0.9707 and an F1 Score of 0.9495. The Best Hybrid Model (DL), which has an Accuracy of 0.9709 and the highest F1 Score of 0.9509, shows how well integrated deep architectures can improve predictive power. The final soft voting models for both ML and DL also work very well, which shows that ensemble techniques are effective. These results show that hybridization and ensemble learning work much better than standalone models for predicting diabetes [36]. They make predictions more reliable, less biased, and more generalizable.

When you look at how well each model works on its own, deep learning models like ENN (Accuracy: 0.9682, F1 Score: 0.9473) and FNN (Accuracy: 0.9661, F1 Score: 0.9440) do much better than simpler models like Logistic Regression, which has the lowest scores on all metrics (Accuracy: 0.9572, F1 Score: 0.9226). XGBoost has the best scores among ML algorithms (Accuracy: 0.9693, ROC-AUC: 0.9826), and Random Forest is right behind it. The deep learning models are especially good at Recall and ROC-AUC, which means they are better at finding true positives, which is very important in medical diagnostics. CNN isn't quite as good as ENN and FNN, but it still does a good job of predicting, especially when used in ensemble formats. Overall, the chart shows that while each model works well on its own, ensemble and hybrid architectures, especially those that combine DL and ML two stage soft voting model, give the best and most consistent results [37,38]. These insights show how important it is to use different algorithmic points of view to improve diagnostic decision-making in healthcare AI applications.

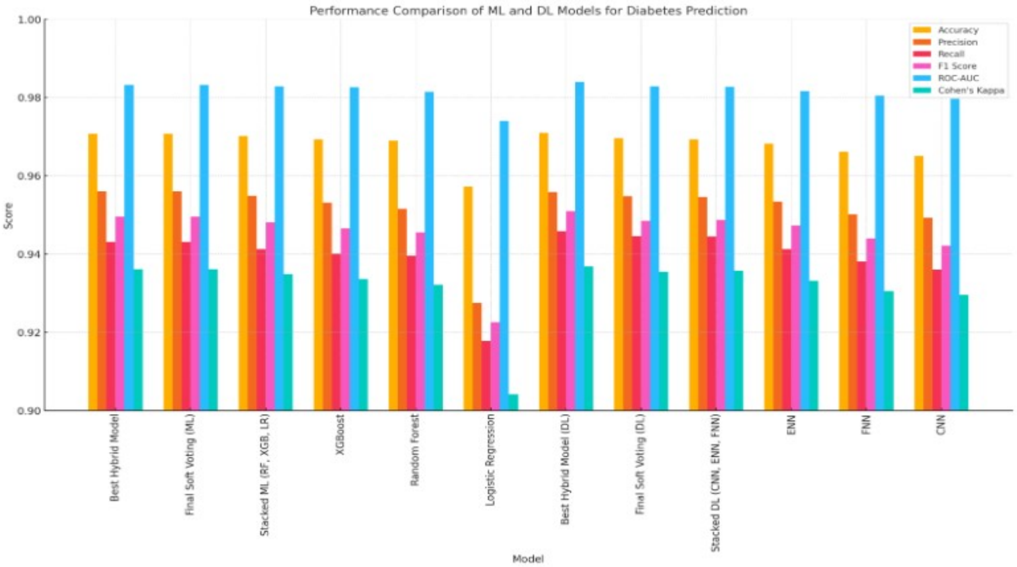

Figure 8: Two Stage Hybrid Soft Voting Ensemble Model

8.1. Pseudocode: Hybrid Soft Voting Ensemble (DL and ML)

```
INPUT: Preprocessed dataset X, true labels y
STEP 1: Preprocessing
- Normalize numerical features using StandardScaler
- Encode categorical features using Label Encoding
STEP 2: Train Deep Learning Models (on reshaped X)
- Train CNN model → predict_proba_DL1 = CNN.predict_proba(X)
- Train FNN model → predict_proba_DL2 = FNN.predict_proba(X)
- Train ENN model → predict_proba_DL3 = ENN.predict_proba(X)
STEP 3: Deep Learning Soft Voting
- Combine DL predictions (soft voting average):
DL_soft_vote = average(predict_proba_DL1, predict_proba_DL2, predict_proba_DL3)
STEP 4: Train Machine Learning Models
- Train Logistic Regression → predict_proba_ML1 = LR.predict_proba(X)
- Train Random Forest → predict_proba_ML2 = RF.predict_proba(X)
- Train XGBoost → predict_proba_ML3 = XGB.predict_proba(X)
STEP 5: Machine Learning Soft Voting
- Combine ML predictions (soft voting average):
ML_soft_vote = average(predict_proba_ML1, predict_proba_ML2, predict_proba_ML3)
STEP 6: Final Hybrid Soft Voting
- Combine DL and ML soft votes:
Final_Hybrid_Vote = average(DL_soft_vote, ML_soft_vote)
STEP 7: Generate Final Predictions
- y_pred = argmax(Final_Hybrid_Vote, axis=1)
STEP 8: Evaluation
- Compute Accuracy, Precision, Recall, F1 Score, ROC-AUC, Cohen's Kappa using y_pred and y
OUTPUT: y_pred, Evaluation Metrics
```

9. Critical Analysis and Results for Two Stage Soft Voting Model

By combining the strengths of Deep Learning (DL) and Machine Learning (ML) techniques, the proposed hybrid soft voting ensemble model makes a big step forward in predicting diabetes. This study uses a two-stage ensemble structure instead of just using statistical models or neural networks. In this structure, DL models (CNN, FNN, ENN) and ML models (XGBoost, Random Forest, Logistic Regression) are trained and optimized separately before being combined in a final soft voting layer [39]. This approach lets the model find both structured feature relationships and complex non-linear dependencies, which makes the predictive system balanced and high-performing.

The quantitative results strongly support how well this method works. The Best Hybrid Model had an F1 Score of 0.9495, a Recall of 0.9430, an Accuracy of 0.9707, and a Precision of 0.9560. The model also had a ROC-AUC of 0.9832, which shows that it was very good at telling the difference between diabetic and non-diabetic cases. The Cohen's Kappa score of 0.9361 shows that the model is very strong and agrees with ground truth labels, making it very reliable in clinical settings. When compared to each other, both internal ensembles DL Soft Voting and ML Soft Voting did well, with F1 Scores of 0.9485 and 0.9495, respectively. The final hybrid fusion, on the other hand, clearly did better than all of them, showing how important it is to combine different model groups. Logistic Regression and CNN are both good models on their own, but they didn't do as well when used together. This shows the problems with relying on just one model.

This hybrid architecture not only makes generalization better, but it also makes predictions more accurate, which makes it a good choice for use in real-world clinical decision support systems [40]. The layered ensemble design is a model for future medical AI research, where accuracy, reliability, and ease of understanding are all very important.

Two-stage soft voting is a powerful way to improve prediction accuracy by using the best parts of different machine learning (ML) and deep learning (DL) models. Soft voting takes the predicted probabilities from each model and averages them out. This gives more nuanced and accurate results than hard voting, which is important in healthcare for making informed decisions. ML models are good at structured, tabular data, while DL models are good at finding complex, non-linear patterns. Together, they give us different kinds of information. The first step of two-stage soft voting combines the predictions of several base models. The second step improves these outputs by reducing overfitting and making them more adaptable, especially when there are too many or too few classes. This layered approach also improves generalization by using the unrelated mistakes of different models, which makes predictions more stable and stronger. In diabetes prediction, these kinds of ensembles can greatly improve accuracy, reliability, and confidence estimation, which are all important for safe and effective clinical use.

10. Future Works

The proposed hybrid soft voting ensemble model has shown to be very good at predicting diabetes diagnosis, but there are still many ways to improve and explore it in the future. First, adding more people to the dataset, especially those from different geographic, ethnic, and socioeconomic backgrounds, could make the model more generalizable and fairer for groups that aren't well represented. it's still hard to understand how deep learning works. To make the model's decision-making process clear, future work should use explainable AI (XAI) methods like SHAP or LIME [41]. This is necessary for clinical adoption. Also, it might be possible to test the model in real time in electronic health record (EHR) systems to see how useful it is in hospitals. Temporal health data like long-term glucose trends, medication history, and lifestyle factors could be used to help with early prediction and risk progression analysis. One way to do this is to add recurrent models or attention mechanisms to the DL parts. Adaptive or personalized ensemble strategies could be made to make predictions more relevant to each patient by using their own profiles. The goal of these directions is to make the current model into a decision support system that is easier to understand, can be used on a larger scale, and can be used in a clinical setting.

11. Challenges and Limitations

Even though the proposed hybrid soft voting ensemble does a great job of predicting, there are still some problems and limitations that need to be addressed. First, the model was created and tested using only one dataset, which may make it less useful for larger groups of people with different demographic, genetic, and lifestyle factors. It is important to validate the results in real-world datasets from multiple canters in the future to make sure they can be used in clinical settings. Second, even though combining deep learning and machine learning gives better accuracy, the fact that DL models are black boxes (especially CNN and ENN) makes them harder to understand and improves their performance and generalization [42]. This lack of openness can make it hard for healthcare workers to trust and accept decisions about diagnoses that need clear explanations. Another problem is that training multiple models separately and then combining them into an ensemble is expensive and complicated [43]. This could make it hard to use the model in real time in clinical settings with few resources and not enough hardware support. Also, the current method doesn't consider time-series or longitudinal health data, which are very important for understanding how chronic diseases like diabetes get worse over time.

Lastly, the model's performance may change when there are missing values or class imbalance, which is common in clinical datasets. Improving robustness would mean using better data imputation and adaptive resampling methods to deal with these problems.

12. Conclusion

This study shows a strong and high-performing hybrid soft voting ensemble model that uses both deep learning (CNN, FNN, ENN) and machine learning (logistic regression, random forest, XGBoost) classifiers to predict diabetes. The proposed model combines the best parts of both paradigms by using a two-stage ensemble architecture. The first stage does soft voting within DL and ML groups, and the second stage combines the outputs through a final soft voting layer. This way, DL can learn complex patterns and ML can be stable and easy to understand. The hybrid model did better than all the other models on all the evaluation metrics, including an accuracy of 0.9707, a precision of 0.9560, a recall of 0.9430, an F1 score of 0.9495, a ROC-AUC of 0.9832, and a Cohen's Kappa of 0.9361. These results clearly show that the model is better at generalizing across different types of data and lowering prediction bias than either individual classifiers or subgroup ensemble methods.

Finally, the final hybrid ensemble is a strong, scalable, and clinically relevant way to predict diabetes. Its strong performance on a number of metrics shows that it could be used in real-world healthcare systems, where accurate early detection is important for improving patient outcomes and supporting preventative care strategies.

Acknowledgment: I would like to express my deepest gratitude to my PhD supervisor, Mohd Khairul Azmi Bin Hassan Dr., for their unwavering support, insightful guidance, and encouragement throughout the course of this research. Their expertise and thoughtful feedback have been instrumental in shaping this work and in my development as a researcher. I am truly thankful for the opportunity to learn under their mentorship.

Competing Interests: The authors declare that they have no known competing financial or personal relationships that could be viewed as influencing the work reported in this paper.

Conflict of Interest: The authors declare that they have no conflict of interest.

Funding: This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit organizations.

Informed Consent and Patient Details: The authors declare that no direct data were collected from any patients. Instead, they utilized secondary data from publicly available datasets.

Reference

1. World Health Organization. (2024). World Diabetes Day 2024. Who.int.
2. Hossain, M. J., Al-Mamun, M., & Islam, M. R. (2024). Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. *Health Science Reports, 7*(3), e2004.
3. Rahman, A., Debnath, T., Kundu, D., Khan, M. S. I., Aishi, A. A., Sazzad, S., ... & Band, S. S. (2024). Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities. *AIMS Public Health, 11*(1), 58.
4. Gangani, P., Alyaseri, S., & Hosseini, S. (2025). Machine Learning Approaches for Predicting Diabetes Onset: A Comparative Study of XGBoost, Random Forest, and Traditional Models. *Authorea Preprints*.
5. Ahmed, S. F., Alam, M. S. B., Hassan, M., Rozbu, M. R., Ishtiak, T., Rafa, N., ... & Gandomi, A. H. (2023). Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review, 56*(11), 13521-13617.
6. Lenka, S. R., Bisoy, S. K., Priyadarshini, R., Hui, K. L., & Sain, M. (2024). Evolutionary-based multi-objective and conditional generative adversarial networks for credit scoring. *IEEE Access*.
7. Dong, H., Liu, M., Zhou, K., Chatzi, E., Kannala, J., Stachniss, C., & Fink, O. (2025). Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. *arXiv preprint arXiv:2501.18592*.
8. Shams, M. Y., Tarek, Z., & Elshewey, A. M. (2025). A novel RFE-GRU model for diabetes classification using PIMA Indian dataset. *Scientific reports, 15*(1), 982.
9. Olorunfemi, B. O., Ogunde, A. O., Almogren, A., Adeniyi, A. E., Ajagbe, S. A., Bharany, S., ... & Hamam, H. (2025). Efficient diagnosis of diabetes mellitus using an improved ensemble method. *Scientific Reports, 15*(1), 3235.
10. Olaniran, O. R., Sikiru, A. O., Allohibi, J., Alharbi, A. A., & Alharbi, N. M. (2025). Hybrid Random Feature Selection and Recurrent Neural Network for Diabetes Prediction. *Mathematics, 13*(4), 628.
11. Yang, Y., Lv, H., & Chen, N. (2023). A survey on ensemble learning under the era of deep learning. *Artificial Intelligence Review, 56*(6), 5545-5589.
12. Afsaneh, E., Sharifdini, A., Ghazzaghi, H., & Ghobadi, M. Z. (2022). Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. *Diabetology & Metabolic Syndrome, 14*(1), 196.
13. Elgendy, I. A., Hosny, M., Albashrawi, M. A., & Alsenan, S. (2025). Dual-stage explainable ensemble learning model for diabetes diagnosis. *Expert Systems with Applications, 274*, 126899.
14. Kumari, S., Kumar, D., & 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.
15. Jabbar, H. G. (2024). Advanced threat detection using soft and hard voting techniques in ensemble learning. *Journal of Robotics and Control (JRC), 5*(4), 1104-1116.
16. Chhillar, I., & Singh, A. (2025). An improved soft voting-based machine learning technique to detect breast cancer utilizing effective feature selection and SMOTE-ENN class balancing. *Discover Artificial Intelligence, 5*(1), 4.
17. Juwara, L., El-Hussuna, A., & El Emam, K. (2024). An evaluation of synthetic data augmentation for mitigating covariate bias in health data. *Patterns, 5*(4).
18. Pezoulas, V. C., Zaridis, D. I., Mylona, E., Androutsos, C., Apostolidis, K., Tachos, N. S., & Fotiadis, D. I. (2024). Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and structural biotechnology journal, 23*, 2892-2910.
19. Shin, J., Kim, J., Lee, C., Yoon, J. Y., Kim, S., Song, S., & Kim, H. S. (2022). Development of various diabetes prediction models using machine learning techniques. *Diabetes & Metabolism Journal, 46*(4), 650-657.
20. Sugandh, F. N. U., Chandio, M., Raveena, F. N. U., Kumar, L., Karishma, F. N. U., Khuwaja, S., ... & Sugandh, F. (2023). Advances in the management of diabetes mellitus: a focus on personalized medicine. *Cureus, 15*(8).
21. Khokhar, P. B., Gravino, C., & Palomba, F. (2025). Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review. *Artificial intelligence in medicine, 103132*.
22. Shao, H., Liu, X., Zong, D., & Song, Q. (2024). Optimization of diabetes prediction methods based on combinatorial balancing algorithm. *Nutrition & Diabetes, 14*(1), 63.
23. Alam, M. A., Sohel, A., Hasan, K. M., & Islam, M. A. (2024). Machine Learning And Artificial Intelligence in Diabetes Prediction And Management: A Comprehensive Review of Models. *Journal of Next-Gen Engineering Systems*.
24. Mustafa, M. (2023). Diabetes prediction dataset. Kaggle.
25. Aubaidan, B. H., Kadir, R. A., & Ijab, M. T. (2024). A comparative analysis of SMOTE and CSSF techniques for diabetes classification using imbalanced data. *Journal of Computer Science, 20*(9), 1146-1165.
26. Kibria, H. B., Nahiduzzaman, M., Goni, M. O. F., Ahsan, M., & Haider, J. (2022). An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. *Sensors, 22*(19), 7268.
27. Toleva, B., Atanasov, I., Ivanov, I., & Hooper, V. (2025). An effective methodology for diabetes prediction in the case of class imbalance. *Bioengineering, 12*(1), 35.
28. Ahmed, N., Ahammed, R., Islam, M. M., Uddin, M. A., Akhter, A., Talukder, M. A., & Paul, B. K. (2021). Machine learning based diabetes prediction and development of smart web

application. *International Journal of Cognitive Computing in Engineering, 2*, 229-241.

29. Abousaber, I., Abdallah, H. F., & El-Ghaish, H. (2025). Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets. *Frontiers in Artificial Intelligence, 7*, 1499530.
30. Motamedi, B., & Villányi, B. (2025). A predictive analytics approach with Bayesian-optimized gentle boosting ensemble models for diabetes diagnosis. *Computer Methods and Programs in Biomedicine Update, 7*, 100184.
31. Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. *Measurement: Sensors, 25*, 100605.
32. Mijwil, M. M., & Aljanabi, M. (2024). A comparative analysis of machine learning algorithms for classification of diabetes utilizing confusion matrix analysis. *Baghdad Science Journal, 21*(5), 24.
33. Ganie, S. M., Pramanik, P. K. D., Bashir Malik, M., Mallik, S., & Qin, H. (2023). An ensemble learning approach for diabetes prediction using boosting techniques. *Frontiers in Genetics, 14*, 1252159.
34. Reza, M. S., Amin, R., Yasmin, R., Kulsum, W., & Ruhi, S. (2024). Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data. *Heliyon, 10*(2).
35. Cui, S., Han, Y., Duan, Y., Li, Y., Zhu, S., & Song, C. (2023). A two-stage voting-boosting technique for ensemble learning in social network sentiment classification. *Entropy, 25*(4), 555.
36. Al-shanableh, N., Alzyoud, M., Al-husban, R. Y., Alshanableh, N. M., Al-Oun, A., Al-Batah, M. S., & Alzboon, S. (2024). Advanced ensemble machine learning techniques for optimizing diabetes mellitus prognostication: A detailed examination of hospital data. *Data Metadata, 3*, 363.
37. Sarker, S., Senjuty, S. A., & Islam, S. (2024, December). A Novel Two-Layer Hybrid Stacked Ensemble for Early-Stage Diabetes Classification. In *2024 13th International Conference on Electrical and Computer Engineering (ICECE)* (pp. 28-33). IEEE.
38. Omkari, D. Y., & Shaik, K. (2024). An integrated two-layered voting (tlv) framework for coronary artery disease prediction using machine learning classifiers. *IEEE access, 12*, 56275-56290.
39. Jain, R., Tripathi, N. K., Pant, M., Anutariya, C., & Silpasuwanchai, C. (2024). Investigating gender and age variability in diabetes prediction: A multi-model ensemble learning approach. *IEEE Access, 12*, 71535-71554.
40. Tuppad, A., & Patil, S. D. (2022). Machine learning for diabetes clinical decision support: a review. *Advances in Computational Intelligence, 2*(2), 22.
41. Alkhanbouli, R., Matar Abdulla Almadhaani, H., Alhosani, F., & Simsekler, M. C. E. (2025). The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. *BMC medical informatics and decision making, 25*(1), 110.
42. Kumar, R., Garg, S., Kaur, R., Johar, M. G. M., Singh, S., Menon, S. V., ... & Lozanović, J. (2025). A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions. *Frontiers in Artificial Intelligence, 8*, 1583459.
43. Tun, H. M., Rahman, H. A., Naing, L., & Malik, O. A. Trust in AI-Based Clinical Decision Support Systems Among Healthcare Workers: A Systematic Review.