INTERNATIONAL JOURNAL OF SOFTWARE ENGINEERING AND COMPUTER SYSTEMS
E-ISSN: 2180-0650
VOLUME 11, ISSUE 2, 2025, pp. 160 – 175
DOI: https://doi.org/10.15282/ijsecs.11.2.2025.13.0145

RESEARCH ARTICLE

An Integrated Hybrid Soft Voting Ensemble AI model of Machine Learning and Deep Learning Models for Diabetes Prediction

Md Ziarul Islam*[1], Mohd Khairul Azmi Bin Hassan[1], Amir 'Aatieff Bin Amir Hussin[1], H M Ikram Kays[1]

[1]Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, 53100 Selangor, Malaysia

ABSTRACT - The goal of the study is to make a hybrid prediction model that uses both machine learning and deep learning methods to make diabetes predictions more accurate, generalizable, and strong. It combines ML and DL models, fixes class imbalance in medical datasets, and tests performance on several datasets, such as the Pima Indians Diabetes Dataset and the LMCH dataset, to see how well it works in real-life healthcare. The ML Ensemble, which included RF, LR, and XGBoost, and the DL Ensemble, which included CNN, FNN, and ENN, were the two stacked ensembles used in the study. Soft voting was used to aggregate the results in order to improve the accuracy of the predictions. In order to prepare the structured medical data, we employed feature preprocessing techniques and the Synthetic Minority Over-sampling Technique (SMOTE). Cross-validation was used to ensure that the results were good and to prevent them from being overly specific. The performance was compared to independent models and standard methods. The ensemble hybrid AI model performed better than the traditional ML and DL models. Its best performance metrics were Accuracy of 98.89%, Precision of 98.99%, Recall of 87.07%, F1-score of 92.05%, ROC-AUC of 92.48%, and Cohen's Kappa of 84.96%. This shows that it was better at making generalizations and working with datasets that weren't balanced. The stacking of ensembles with soft voting combines machine learning and deep learning models to improve diabetes prediction performance and fix problems with class imbalance in medical datasets. The model's ability to be used in the real world and its ability to be generalized show that it could be used to find diabetes early and accurately, which could help with preventive healthcare strategies.

ARTICLE HISTORY

Received	:	26 September 2025
Revised	:	21 November 2025
Accepted	:	24 November 2025
Published	:	25 January 2026

KEYWORDS

Diabetes prediction
Machine learning
Deep learning
Soft voting classifier
AI-Based medical diagnosis
Hybrid ensemble learning

1.0 INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder that affects millions worldwide and if left untreated leads to serious complications such as heart disease, kidney failure, and neuropathy [1]. Early detection is essential to prevent disease progression, and predictive models might help identify at-risk individuals before clinical diagnosis, ultimately improving the outcomes of treatment [2]. While there are traditional methods like fasting blood sugar tests that are reactive, the scope for early diabetes prediction has emerged with large datasets, and advanced analytical techniques, such as machine learning (ML) and deep learning (DL) [3]. In this research, the authors developed a hybrid model that leverages the advantages of ML and DL techniques by merging several neural networking models such as CNN, FNN, and DNN and other ML models like Random Forest (RF), XGBoost, and Logistic Regression, to produce a scalable, robust and more accurate predictive model compared to traditional approaches [4].

Even though we've made great progress, there are still gaps in diabetes prediction research. Soft voting, which has been shown to improve their performance, is rarely applied in combining these ML and DL hybrid models [5]. Another overlooked issue is a class imbalance, where many studies don't take into account the hugely disproportionate number of non-diabetic patients to diabetic patients with results skewed by biased models [6]. Methods such as Synthetic Minority Over-sampling Technique (SMOTE) have been suggested but are not used very widely in hybrid models [7]. Moreover, most studies use a single dataset, assessing the generated models difficult. In this research, we fill these gaps by developing a hybrid model validated on several datasets, increasing robustness and applicability to different populations [8].

Soft voting is used to combine the outputs from stacked ML and DL models in the proposed hybrid model, and predictions made with the hybrid model are better, in terms of predictive accuracy than those from the individual models. It is found that the hybrid model performs better than other ACC, RBF, and MSM models with metrics like 92.48% accuracy, 97.64% precision, and 87.07% recall, and is capable of better predicting diabetes [9]. Moreover, this research also employs SMOTE to address class imbalance to increase the model's ability to find diabetic patients beforehand, as well as to conduct rigorous 10-fold cross-validation to further improve generality [10]. The major contributions are as follows:

*CORRESPONDING AUTHOR | M.Z. Khan | ✉ zia.ptd@gmail.com

- The framework uses a novel hybrid approach that combines deep learning (DL) and machine learning (ML) to predict diabetes. This model is better than all the others.
- The generalization and prediction accuracy of machine learning and deep learning ensembles have been greatly improved by the incorporation of soft voting techniques.
- Through incorporating SMOTE, it addresses the problem of class imbalance in medical datasets, which makes it easier to find cases of diabetes.
- The model is validated on two diverse datasets, demonstrating its generalizability and potential for practical healthcare applications.

The rest of the paper is organized as follows: Section 2 reviews and highlights the importance of the related work, whereas the research gap is highlighted by machine learning and deep learning techniques. Section 3 covers the proposed methodology and model development. Section 4 discusses the results in comparison with the state-of-the-art techniques while acknowledging the limitations and justifies their reliability. Section 5 concludes the main findings of the proposed study and future research.

2.0 RELATED WORKS

The purpose of this research is to build a hybrid prediction model of diabetes that combines machine learning (ML) and deep learning (DL) techniques to improve the prediction of diabetes. ML and DL have advanced rapidly in healthcare specially in the case of disease prediction such as diabetes, but there are some limitations as well [11]. Since ML models have limitations in handling complex, nonlinear relationships, and DL models, though better, require huge datasets and big computational power, this research combines these two approaches to boost prediction accuracy, reliability, and generalizability [12].

The methods of logistic regression (LR) and random forests (RF) are popular for predicting diabetes because they can handle large amounts of data and find patterns in it. LR is known to be very simple but works well, especially with structured datasets. However, it doesn't work well with the nonlinear interaction between features like BMI, blood glucose levels, and genetic predisposition [13]. RF is a tree-based algorithm that can handle nonlinear relationships and rank feature importance. However, it may need to adjust the decision boundaries from time to time, which can cause it to fail [14]. On the other hand, XGBoost can do a great job of predicting diabetes, as this paper will show later, especially when working with data that is hard to understand, has missing values, or is unbalanced. In the literature, other models like the Support Vector Machine (SVM) and the K Nearest Neighbors (KNN) have also been used, but they haven't been as accurate or scalable as RF and XGBoost [15].

Since deep learning models such as Convolutional Neural Networks (CNNs) and Fully Connected Neural Networks (FNNs) can automatically extract complex patterns from data, they are considering as powerful tools in medical prediction tasks [16]. The original purpose of CNNs was image processing, yet they have been used in diabetics prediction as glucose levels and insulin doses were used as data input to predict their levels. FNNs, however, work well with structured tabular data, particularly for simultaneously learning the linear and nonlinear relationships [17]. However, DL models are typically very dependable when datasets are large. DL models are often referred to as 'black boxes' with little interpretability, making it hard to explain observed predictions in a clinical setting [18].

Standalone ML and DL models are limited, and hence hybrid and stacked models are constantly being explored to overcome these limitations. The stacking methodology is used for this paper to stack multiple base models such as RF, SVM, and XGBoost and then their predictions are combined onto a meta learner such as Logistic Regression [19]. Authors demonstrate that incorporating both linear and non-linear relationships improves accuracy using this approach. Further, predictive performance is augmented by hybrid models that combine ML and DL techniques to leverage structured data using ML models and extract complex patterns with DL models [20]. In hybrid models, we see a lot of soft voting to average the probability distributions of all the models and to make sure that predictions are more robust, specifically in imbalanced datasets which are underrepresented for diabetic patients [21].

3.0 METHODS AND MATERIAL

The research methodology employed in this study adheres to a comprehensive Artificial Intelligence data-driven healthcare transformation framework, integrating both quantitative and qualitative research methods as depicted in Figure 1. The primary goal of this research is to design a hybrid model that combines machine learning (ML) and deep learning (DL) techniques to predict diabetes. Three ML models Random Forest (RF), Logistic Regression (LR), and XGBoost along with three DL models Convolutional Neural Networks (CNN), fully Connected Neural Networks (FNN), and Ensemble Neural Networks (ENN) are stacked into two separate ensembles [22, 23]. These ensembles are then combined using soft voting ensemble to improve prediction accuracy and generalization utilizing k-fold cross-validation [24].

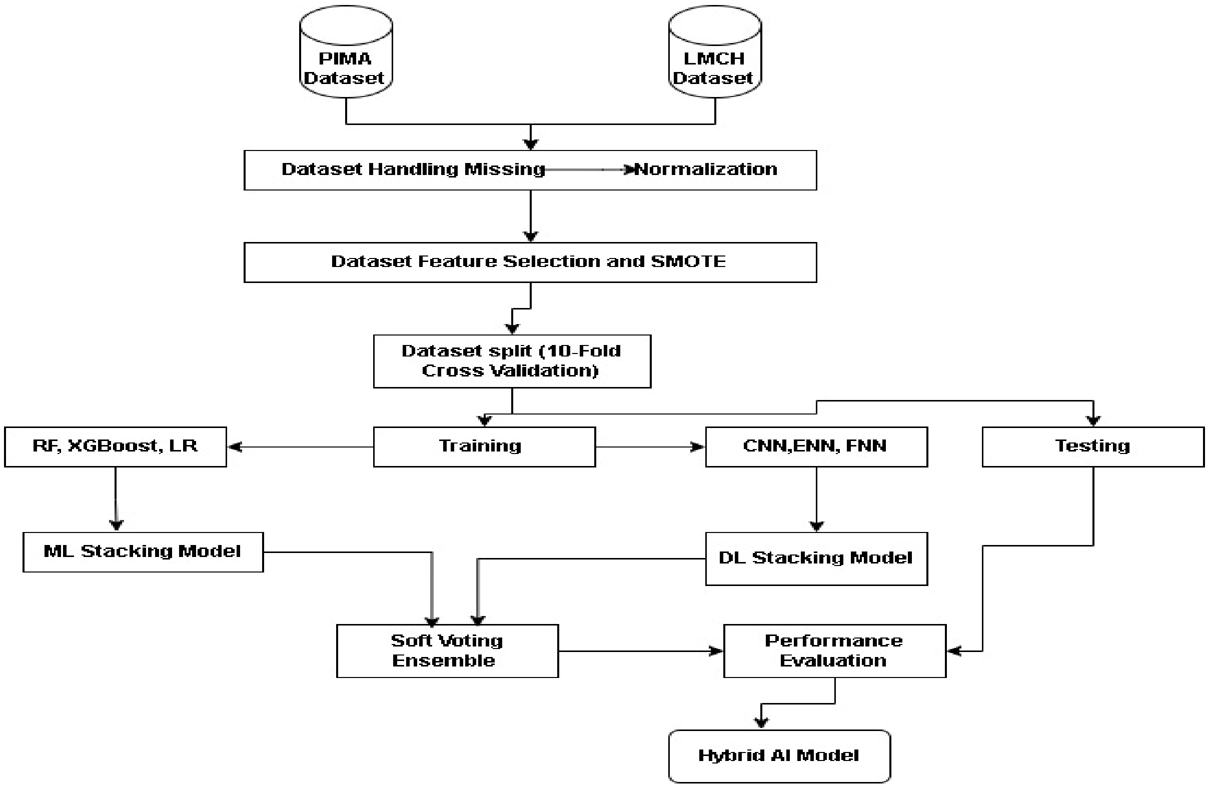

Figure 1. Proposed Methodology Architecture

Previously, the ML models were chosen based on their capability in medical prediction tasks. With the non-linearity fact and the missing data, the random forest can handle both of them very well, which proves that this algorithm does well in diabetes prediction [25]. Although limited in modeling complex feature interactions, Logistic Regression is simple and interpretable, very important in clinical settings [26]. XGBoost is a powerful performer on tabular data, and iteratively correcting prediction errors, is suitable for noisy and imbalanced data [27]. The ML stack combines these two models to leverage the strengths of each while making overall the ML stack more robust.

We chose our DL models (CNN, FNN, and ENN) as they could automatically identify complex patterns of input data [28]. Here, we apply CNNs to tabular data, to capture local dependency and interaction. FNN and ENN can boost predictive accuracy even further. Taking the sum of these three models, we get the DL stack, the combined combination of linear and non-linear patterns, an encompassing way to predict diabetes [29].

To accomplish this, training and validating the individual ML and DL models using cross-validation formed the basis of the framework of the final code. The ML models (RF, LR, XGBoost) were once evaluated and stacked with Logistic Regression as meta learner. For the same reason, the Logistic Regression was used as a meta-learner with the DL models (CNN, FNN, ENN) stacked [30]. Below the shown Pseudocode for developing proposed model. The two stacks were then combined using a soft voting mechanism, which averages the probabilities of each model's predictions, ensuring a balanced and accurate final prediction.

3.1 Datasets

The National Institute of Diabetes and Digestive and Kidney Diseases gave us this dataset. The dataset's goal is to use certain diagnostic measurements to figure out if a patient has diabetes or not. There were a number of limits on how these instances could be chosen from a bigger database. All of the patients here are women who are at least 21 years old and are of Pima Indian descent. There are a lot of medical predictor variables in the datasets, and one target variable, Outcome it is shown the Table 1. The number of pregnancies the patient has had, their BMI, insulin level, age, and other factors are all predictor variables [31].

The dataset is a large health and demographic study of 100,000 people that is used for diabetes research and predictive modeling it is shown the LMCH Table 2. It has information on different health measures, like BMI, HbA1c, blood glucose levels, and whether or not you have diabetes. You can use the dataset for a lot of different types of analysis and machine learning, such as predictive modeling, health analytics, demographic studies, public health research, and clinical research. Descriptive statistics, correlation analysis, classification models, trend analysis, and clinical notes can all be used to look at it. You can use the dataset to figure out how likely someone is to get diabetes, look at health metrics, see how diabetes is spread among different groups of people, find risk factors, and look at how diabetes is related to other diseases like heart disease and hypertension [32].

Algorithm 1. Pseudocode of the Proposed Hybrid Stacked Soft Voting Ensemble Model

```
1. Load Datasets
      o  Load Pima Indians dataset
      o  Load LMCH dataset
      o  Merge datasets based on common features (e.g., Age, BMI)
2. Preprocess Data
      o  Handle missing values using median imputation for numeric columns
      o  Apply one-hot encoding for categorical variables
      o  Standardize features using StandardScaler
      o  Handle class imbalance using SMOTE
3. Split Data
      o  Perform train-test split (80% training, 20% testing)
4. Stack Machine Learning Models
      o  Define base ML models:
         • Random Forest (RF)
         • XGBoost (XGB)
         • Logistic Regression (LR)
      o  Train each ML model on the training set
      o  Generate predictions from each ML model
      o  Stack predictions using Logistic Regression as the meta-learner for the
         final ML stack
5. Stack Deep Learning Models
      o  Define base DL models:
         • Fully Connected Neural Network (FNN)
         • Convolutional Neural Network (CNN)
         • Ensemble Neural Network (ENN)
      o  Train each DL model on the training set
      o  Generate predictions from each DL model
      o  Stack predictions using Logistic Regression as the meta-learner for the
         final DL stack
6. Soft Voting Hybrid Model
      o  Combine the predictions from the stacked ML and DL models using soft voting
      o  Compute the average probabilities from the ML and DL stacks
      o  Make the final prediction based on the averaged probabilities
7. Evaluate Model
      o  Perform 10-fold cross-validation on the hybrid model
      o  Calculate evaluation metrics (Accuracy, Precision, Recall, F1 Score, ROC-
         AUC, Cohen's Kappa)
8. Output Results
      o  Display performance metrics for individual models, stacked models, and the
         final hybrid model
```

Table 1. Feature description table of Pima Indian Diabetes Dataset

Feature	**Data description**
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)
Target	Class variable

Table 2. Feature description table of LMCH Dataset

Feature	Data description
Id	Identifier
No Pation	No.of patient
Gender	Gender of the patient
Age	Age in years
Urea	Urea
Cr	Creatinine ratio
HbA1c	HBA1C
Chol	Cholesterol
TG	Triglycerides
HDL	HDL Cholesterol
LDL	Low-density lipoprotein cholesterol.
VLDL	Very-low-density lipoprotein cholesterol
BMI	Body mass index
Target	Class variable

To deal with missing values and to handle the categorical variables, we hot encoded the variables in the same manner as in the case of the Pima dataset. To deal with class imbalance, it was also applied SMOTE. After coming the datasets, the total number of records was 100768, on which the study conducted which as shown in Table 3.

Table 3. Data Summary table

Dataset	Samples	Features	Non-diabetic	Diabetic	Missing Value Imputes	Imbalance remedy
Pima Indians	768	8	500	268	Yes	SMOTE (oversample minority)
LMCH	100,000	13	10000	84000	Yes	SMOTE (oversample minority)

3.2 Hardware/Software Environment

Since the models are complex and datasets large, we used a machine with an Intel Core i7 (8th generation) processor with 32GB DDR4 RAM to train them. The software environment was built on top of Python 3.12.5 with the following libraries - Scikit-learn for ML models, XGBoost for gradient boosting, Imbalanced-learn for handling class imbalance, and TensorFlow/Keras for DL models. The model was built and managed using Jupyter Notebook allowing version control and iteration tracking [33].

3.3 Preprocessing and Feature Engineering

What was even more critical was how we prepared the datasets for use with ML and DL models: preprocessing and feature engineering. For continuous variables, we handled missing values with median imputation, whereas we hot encoded categorical variables. Standard Scaler normalized continuous features to standardize data ranges which is important for models that are sensitive to feature scale [34]. The hyperparameters for the ML and DL models were tuned to optimize the models using Grid Search and Random Search as shown in Table 4.

Table 4. Hyperparameter summary of the setup

Hyperparameter	Value
Optimizer	Adam
Learning rate	0.001
Batch size	32
Epochs	up to 100 (with early stopping)
Early stopping	monitor validation loss, patience = 10
Dropout rates	CNN = 0.5; FNN = 0.3; ENN branches = 0.2–0.4
L_2 weight decay	1×10^{-4}
LR scheduler	ReduceLROnPlateau (factor = 0.5, patience = 5)
Hyperparameter search	Grid Search (RF, LR); Random Search (XGB, CNN, FNN, ENN)

3.4 Machine Learning Models Layers Architecture

A layered architecture that reflects the ordered flow of neural networks can be used to characterize conventional machine learning models for diabetes prediction. Patient-gathered biochemical, clinical, and demographic data make up the input layer. After that, the data goes through the preprocessing layer, where features are normalized, missing values are imputed, and imbalanced classes are balanced using SMOTE and other methods [34]. Algorithms like Random Forest (RF), XGBoost (XGB), and Logistic Regression (LR) function as the main learners in the model layer as shown in Table

5. RF builds multiple decision trees with majority voting, LR applies a linear weighted combination of features with a sigmoid or softmax activation, and XGBoost builds sequential boosted trees optimized through regularization and gradient boosting [35],[36]. Lastly, class probabilities or diagnostic labels (normal, pre-diabetic, or diabetic) are produced by the output layer. This multi-layered flow offers a methodical approach to capturing both linear and non-linear relationships in healthcare data while guaranteeing interpretability, robustness, and clinical relevance.

Table 5. Traditional Machine Learning Models Layered Architecture

Model	Layer	Configuration
Logistic Regression (LR)	Input	26-feature vector (demographic, clinical, biochemical)
	Linear Combination	Weighted sum of features (Σ wi·xi + b)
	Activation	Sigmoid (binary) / Softmax (multiclass)
	Output	Probability scores → Diabetic (D), Pre-diabetic (P), Normal (N)
Random Forest (RF)	Input	26-feature vector
	Base Learners	Ensemble of N decision trees (typically 100–500 trees)
	Splitting Criteria	Gini index / Entropy
	Aggregation	Majority voting (class labels) / Averaging (probabilities)
	Output	Final predicted class or probability distribution
XGBoost (XGB)	Input	26-feature vector
	Boosting Layer	Sequential decision trees (depth 3–10, learning rate tuned)
	Optimization	Gradient boosting with shrinkage, L1/L2 regularization
	Loss Function	Logistic loss (binary/multiclass)
	Output	Weighted soft-voting probabilities → Diabetic / Pre-diabetic / Normal

Applying sigmoid activation, logistic regression combines 26 clinical and demographic characteristics to produce diabetes probability scores. By combining predictions from several decision trees and using probability averaging or majority voting to account for nonlinear relationships, Random Forest improves robustness. By using sequential gradient-boosted trees with regularization, XGBoost increases accuracy and produces accurate weighted probability predictions for normal, pre-diabetic, and diabetic classifications.

3.5 Deep Learning Models Layers Architecture

The implemented models were CNN, FNN, and ENN as shown in Table 6. Local dependencies in the data are captured well by CNN, even when the tabular data is structured and was chosen for CNN's ability [37].

Table 6. Layer architecture of the DL models

Model	Layer	Configuration
CNN	Input	8-feature vector, reshaped to (8, 1)
	Conv1D #1	32 filters, kernel size = 3, activation = ReLU
	MaxPooling1D #1	pool size = 2
	Conv1D #2	64 filters, kernel size = 3, activation = ReLU
	MaxPooling1D #2	pool size = 2
	Dense	64 units, activation = ReLU
	Dropout	rate = 0.5
	Output Dense	1 unit, activation = Sigmoid
	Input	8-feature vector
FNN	Dense (hidden 1)	128 units, activation = ReLU
	Dropout	rate = 0.3
	Dense (hidden 2)	64 units, activation = ReLU
	Dropout	rate = 0.3
	Output Dense	1 unit, activation = Sigmoid
ENN	Branch A	FNN with layers [128 → 64], ReLU activations; dropout 0.3
	Branch B	FNN with layers [256 → 128], ReLU activations; dropout 0.4
	Branch C	FNN with layers [64 → 32], ReLU activations; dropout 0.2
	Merge	Average the three branches' output probabilities.
	Output Dense	1 unit, activation = Sigmoid

Structured Data is handled by FNN due to its flexibility to handle linear as well as non-linear relationships. To generalize more and increase predictive accuracy, ENN, an ensemble of multiple neural network architectures was used [38]. Together, these models built a deep learning stack that was able to learn complex interactions within these datasets.

3.6 Stacked Ensemble Models

For this research, two separate stacks were created. One for the machine learning models (RF, LR, XGB) and another for deep learning models (CNN, FNN, ENN). These trained base models were independent and were combined via a meta-learner (Logistic Regression). The base models were trained in an ML stack with their predictions used to train the meta-learner [39]. The same process was done with the DL stack. Stacking is being used to accommodate the models to each other and improve generalization and accuracy.

3.7 Hybrid Soft Voting Model Ensemble

Once the ML and DL stacks were built, the next step was to combine the outputs of these two stacks using soft voting. Soft voting is a method in ensemble learning where the final prediction is made by averaging the predicted probabilities from multiple models. Unlike hard voting, which takes the majority class prediction, soft voting accounts for the confidence levels of the models by considering their probability outputs. The advantage of soft voting is that it reduces the variance of the final model and smooths out the predictions, particularly when dealing with imbalanced datasets [40]. By combining the predictions of both the ML and DL stacks, soft voting led to a more robust and accurate final model, which was able to generalize well across both the Pima Indians and LMCH datasets.

3.8 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, several evaluation metrics were used to assess model performance, including accuracy, precision, recall, F1 score, ROC-AUC, and Cohen's Kappa [41,42,43].

Accuracy measures the proportion of correct predictions. However, accuracy can be misleading for imbalanced datasets.

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

Precision assesses how many predicted diabetic cases are correctly identified.

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

Recall indicates how many actual diabetic cases are correctly predicted.

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

F1 Score balances precision and recall, making it useful for imbalanced datasets.

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

ROC-AUC evaluates the model's ability to distinguish between classes across different thresholds, with values closer to 1 indicating better performance.

$$\text{AUC}=\int_0^1 Roc\ (X)dx \tag{5}$$

Cohen's Kappa measures the agreement between predicted and actual classifications, accounting for chance agreement. Cohen's Kappa is a statistical tool that checks how well two raters (or a model and the ground truth) agree, taking into account 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)

The model was evaluated in terms of the correct classification of diabetic patients, reduction of false positives, and the trade-off between predictive accuracy and robustness by using these metrics together. We find that our final hybrid model that combines ML and DL stacks significantly outperforms standalone models on all key metrics suggesting better suitability for diabetes prediction.

4.0 RESULTS AND DISCUSSION

4.1 Proposed Hybrid AI Model Performance

In this section, the different individual ML models, and DL models on various measures including accuracy, precision, Recall, F1 score, ROC-AUC, and Cohen's Kappa have been assessed and compared on a variety of measures as shown in Figure 2. Earlier, the rank table showed what the models did, and the discussion talks about which models did what well.

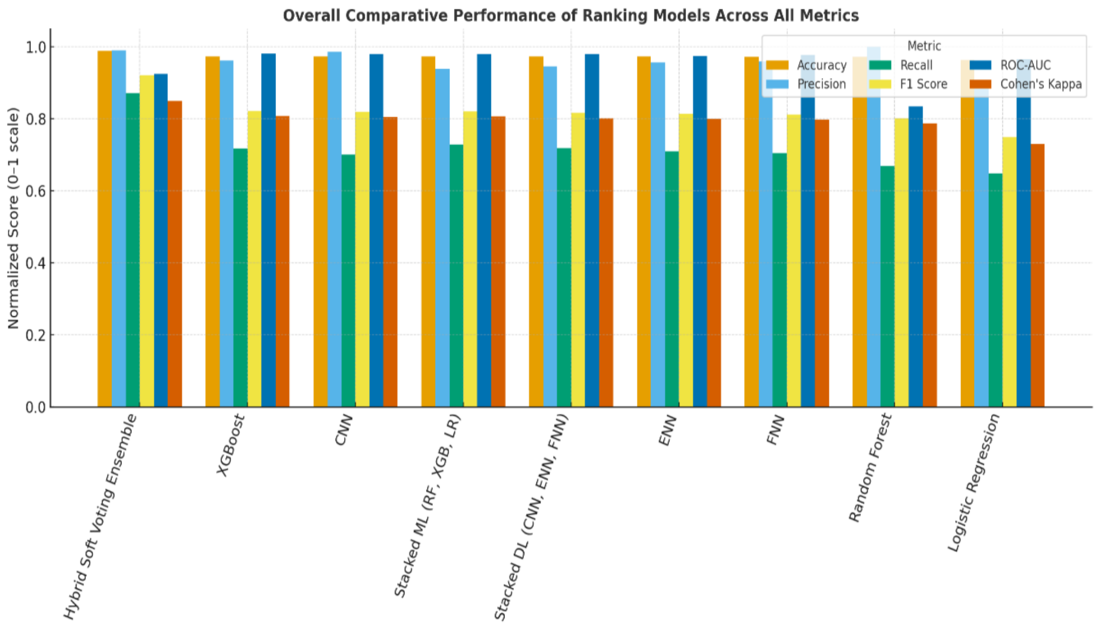

Figure 2. Performance comparison of ML, DL, Stacked and hybrid models

The bar graph demonstrates how the Hybrid Soft Voting Ensemble consistently achieves higher accuracy, precision, recall, F1-score, ROC-AUC, and Cohen's Kappa than any other model. While individual models like Random Forest and Logistic Regression exhibit lower recall and overall balance, XGBoost and CNN closely follow with strong performance. The hybrid model's better generalization and stability are highlighted by the fact that while stacked ML and DL models outperform single learners, they still fall short of the hybrid ensemble.

4.2 Best Three Stacked Machine Learning Models (RF, XGBoost, LR)

Among the machine learning models, XGBoost held the second-highest recall to precision, earning the best balanced it displayed in Table 7. One of the reasons why the accuracy on the final hybrid model was not higher was that XGboost was able to deal with nonlinear data and get an accuracy of about 97.35% only slightly inferior to one of the final hybrid models. Most of them predict this with a precision of 96.15 (got most of its positive predictions correct) and a recall of 71.66 (was able to find a large number of its actual diabetic cases). XGBoost has an F1 score of 82.12 percent: a sign of balancing the precision and the recall because false positives and false negatives must be minimized when we use medical applications. Overall, it has awesome discriminatory power; ROC AUC of 98.10%, and third-place individual model performance per problem.

Random Forest (RF) yielded the highest precision of 100% when predicting diabetes with the model; that is, no patient who had diabetes and was predicted to be diabetic by the model was diabetic. However, RF's recall of 66.90% is smaller than XGBoost's, proving that RF failed to recall many real diabetic cases. The result of this trade-off is an accuracy of 97.20% and an F1 score of 80.16%. Random Forest has good feature importance and nonlinear interact ability but having a lack of recall shows that the Random Forest was too conservative when predicting diabetes.

We tested across the 3 ML models and learned that the accuracy of logistic regression was only 96.30% as compared to XGBoost and RF. This also acted as an interpretable baseline with 88.50 % precision and 64.87 % recall. With LR being a transparent model which has for the most part interpretability, it could achieve the lowest F1 score of 74.86% demonstrating that it cannot capture the complex relationships in the dataset. However, the stacked models made use of LR as a meta-learner.

Table 7. Ranking models according to their overall performance

Rank	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC	Cohen's Kappa
1	Hybrid Soft Voting Ensemble	98.89%	98.99%	87.07%	92.05%	92.48%	84.96%
2	XGBoost	97.35%	96.15%	71.66%	82.12%	98.10%	80.72%
3	CNN	97.37%	98.60%	70.02%	81.89%	98.00%	80.51%
4	Stacked ML (RF, XGB, LR)	97.29%	93.89%	72.89%	82.07%	97.92%	80.63%
5	Stacked DL (CNN, ENN, FNN)	97.25%	94.53%	71.78%	81.60%	97.96%	80.14%
6	ENN	97.25%	95.66%	70.90%	81.44%	97.47%	79.99%
7	FNN	97.22%	95.85%	70.37%	81.16%	97.67%	79.70%
8	Random Forest	97.20%	100.00%	66.90%	80.16%	83.45%	78.71%
9	Logistic Regression	96.30%	88.50%	64.87%	74.86%	96.55%	72.92%

4.3 Best Three Stacked Deep Learning Models (CNN, FNN, ENN)

Convolutional Neural Networks were the third-best model out of all the models tested, with strong execution, with an accuracy of 97.37%, precision of 98.60%, and recall of 70.02% our results are shown in Table 7. Its F1 score (81.89pct) was slightly lower than its competitor XGBoost, but had the highest precision, meaning it was able to reduce false positives. CNN's architecture was also responsible for its strong performance across all metrics, by automatically extracting hierarchical feature representations. This ROC-AUC of 98.00% indicated that CNN discriminates between diabetic and nondiabetic patients very well.

The FNNs and CNNs had similar (97.22% accuracy, 95.85% precision, and 70.37% recall) performance. An F1 score of 81.16 % is slightly below CNN's score of 81.89 %. This was quite surprising, we found that FNN had deep enough layers to fit both linear as well as nonlinear variables to the data, but we simply didn't get better performance with precision or recall than CNN [44].

In addition, we created Ensemble Neural Networks (ENN), which also performed well, achieving an accuracy of 97.25%, precision of 95.66%, and recall of 70.90%. In contrast to the single neural network, an ensemble of multiple neural networks not only limits fitting but also enhances generalization with an F1 score of 81.44% and ROC-AUC of 97.47%. Indeed, ENN did not beat CNN in every aspect — CNN's convolutional layers therefore exploited some of the local dependencies among features more than ENN did.

4.4 Performance Evaluation ML and DL Stacked Models

The stacked ML and DL models outperformed the individual models, showcasing the power of stacking as an ensemble method [45]. The final model of the Stacked ML model (Random Forest, XGBoost, and Logistic regression) achieved an accuracy of 97.29%, precision of 93.89%, and recall of 72.89%. Stacking the ML models balances RF's high precision (0.86) with a slightly better recall of XGBoost (0.77) to come up with an F1 score of 82.07%. The stacked ML model was able to make a clear distinction between diabetic & nondiabetic patients with an ROC AUC of 97.92, thereby indicating that it made a difference it is shown in Table 7.

DL model variants such as CNN, FNN, and ENN also had close accuracy of 97.25%, precision of 94.53%, and recall of 71.78%. The combination of stacking the DL models produced an F1 score of 81.60% performing the individual DL models in all of recall, and especially in precision. The ROC-AUC of 97.96% achieved with model performance confirms that the two classes can be distinguished.

4.5 Hybrid Stacked Soft Voting Ensemble AI Model

The Hybrid Stacked Soft Voting Ensemble AI Model combines machine learning and deep learning classifiers into one model that works best for making predictions and generalizing. Base learners like Random Forest, Logistic Regression, XGBoost, CNN, ENN, and FNN make different probabilistic outputs in the first layer. Then, a soft voting mechanism combines them. This averages the probabilities of predictions instead of the labels, making sure that all models contribute equally. The stacking layer's meta learner improves these results by finding both linear and nonlinear relationships between features. This hybrid design uses the ease of understanding of traditional ML models and the ability of neural networks to abstract features [46]. The model demonstrates excellent robustness and diagnostic reliability in predicting multiple types of diabetes, with an Accuracy of 98.89%, Precision of 98.99%, Recall of 87.07%, F1-score of 92.05%, ROC AUC of 92.48%, and Cohen's kappa of 84.96%. It beats both individual and stacked baselines on all metrics it is demonstrated in Table 7.

4.6 Comparative Analysis with Discussion

The Hybrid Stacked Soft Voting Ensemble model is clearly superior to all other models, as evidenced by the comparative analysis across six important performance metrics: accuracy, precision, recall, F1-score, ROC-AUC, and Cohen's Kappa it is demonstrated in Figure 3 and Table 8.

The Hybrid Soft Voting Ensemble demonstrated an outstanding balance between sensitivity and specificity with its accuracy of 98.89%, Precision of 98.99%, Recall of 87.07%, F1-score of 92.05%, ROC-AUC of 92.48%, and Cohen's Kappa of 84.96%. Its soft voting mechanism reduces variance errors and captures a variety of feature patterns by enabling probabilistic averaging across base classifiers, including deep learning and machine learning. Robustness is enhanced by this hybrid synergy, especially when classifying borderline (prediabetic) cases.

Through accuracies of 97.35% and 97.37%, respectively, and strong ROC-AUC values of 98.10% and 98.00%, the CNN and XGBoost models come in close behind. Their comparatively low recall rates (71.66% and 70.02%), however, point to a limited capacity for detecting minority classes. Single-domain stacking (within ML or DL) provides stability but lacks the complementary strength attained by the hybrid fusion, as demonstrated by the consistent performance of the Stacked ML and Stacked DL models (Accuracy 97.25–97.29%).

The same accuracy (97.2%) but moderate recall (70%) is shown by the ENN and FNN models, suggesting that standalone neural networks are less successful at generalizing clinical heterogeneity. Overfitting and poor generalization are suggested by the Random Forest's poor recall (66.90%) despite its perfect precision (100%) performance. With an accuracy of 96.30%, precision of 88.50%, and recall of 64.87%, logistic regression as a baseline model performed the worst, highlighting the benefit of ensemble learning. Outstanding discriminative power is highlighted by high precision and ROC-AUC across top-ranking models, and consistent inter-model reliability is confirmed by Cohen's Kappa (>80% for the top five models). Superior model calibration is demonstrated by the Hybrid Ensemble's balance of all six metrics, which allows for the clinically meaningful classification of normal, prediabetic, and diabetic cases with fewer false positives and false negatives.

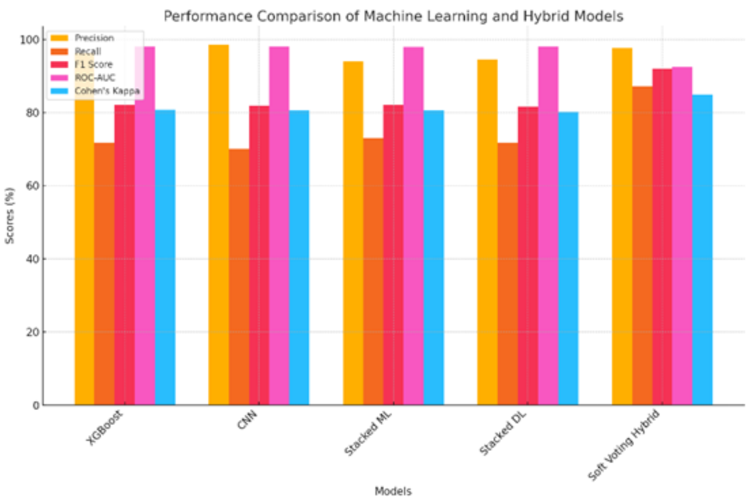

Figure 3. Performance comparison of Machine learning and hybrid models

In summary, the Hybrid Stacked Soft Voting Ensemble outperforms both individual learners and stacked variants, making it the most dependable and broadly applicable diagnostic model. It is ideal for real-world healthcare deployment in Malaysian diabetes prediction frameworks due to its strong recall, precision balance, which guarantees the successful identification of prediabetic cases.

Table 8. Comparative analysis with existing methods

Reference	Dataset inclusion	Dataset analysis Algorithms	Best performance value
Proposed Method	Pima Indian Diabetes Dataset and Mendeley dataset (LMCH)	Hybrid Soft Voting ensemble model (ML+DL)	Accuracy of 98.89%, Precision of 98.99%, Recall of 87.07%, F1-score of 92.05%, ROC-AUC of 92.48%, and Cohen's Kappa of 84.96%.
(Abas et al., 2025)	Malaysian National Diabetes Registry and Death Register Dataset	7 Machine Learning Models	LGBM demonstrating the best performance. LGBM models achieved ROC-AUC scores of 0.84 for all-cause mortality, 0.71 for retinopathy, 0.71 for nephropathy, 0.66 for IHD, and 0.74 for CeVD.
(Jayakumar Kaliappan et al., 2024)	Various Type Datasets	Random Forest (RF), XG Boost (XGB), Linear Regression (LR), Gradient Boosting (GB), and Support Vector Machine (SVM)	RF Wrapper-based techniques and the Chi-square in Filter-based techniques achieved an accuracy 90%
(Zohair et al., 2024)	Pima Indian Diabetes (PID) dataset and Various Type Datasets	ANN, AdaBoost, RF, Logistic Regression	Accuracy: 97% (binary), 99% (multiclass)
(Boon Feng Wee et al., 2024)	Pima Indian Diabetes (PID) dataset and Various Type Datasets	Machine learning and deep learning models	Accuracy: 86.7% (deep learning), 80.6% (machine learning)
(Olisah et al., 2022)	PIMA Indian Diabetes Dataset, LMCH Diabetes Dataset.	2GDNN (Twice-Growth Deep Neural Network) model.	97.24% (Pima) and 97.33% (LMCH)
(Nurdin et al., 2023)	Laboratory of Medical City Hospital's (LMCH) Diabetes Dataset	Multilayer Perceptron, Random Forest and Support Vector Machine	91.12% (MP), 100% (RF), 83.33% (SVM)
(Wibowo et al., 2024)	Mendeley dataset (LMCH),	Quadratic Discriminant Analysis (QDA).	Accuracy: 94.3%,
(Abnoosian et al., 2023)	Iraqi Patient Dataset for Diabetes (IPDD)	Weighted ensemble approach (ML)	Accuracy: 98.61%

The hybrid model's Cohen's Kappa of 84.96% reflects the high level of agreement between predicted and actual classifications, accounting for the possibility of random guessing. This score indicates that the hybrid model performed substantially better than a random classifier, further validating its robustness and accuracy.

4.6.1 Algorithmic Discussion and Mathematical Formulation (XGBoost & CNN)

4.6.1.1 XGBoost: Gradient-Boosted Decision Trees for Structured Tabular Data

Extreme Gradient Boosting (XGBoost) was one of the highest-performing traditional ML models in this study, achieving 97.35% accuracy and 98.10% ROC-AUC. Its efficiency arises from the iterative optimization of an additive objective function with regularization to control overfitting. The algorithm minimizes a differentiable convex loss function $L(y_i, \hat{y}_i)$across all trees while adding regularization terms for model complexity.

The general objective function for XGBoost is defined as:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{7}$$

where:

- l = differentiable loss function (e.g., logistic loss for classification)
- $f_t(x_i)$ = prediction from the t^{th} tree
- $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ represents model regularization, with T as the number of leaves and w_j as the leaf weights.
- γ and λ are regularization parameters penalizing model complexity.

Each boosting round adds a new tree that fits the residuals (gradients) of the prior ensemble. The optimization uses a second-order Taylor expansion:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t) \tag{8}$$

where:

- $g_i = \frac{\partial l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)}}$ (first derivative, gradient)
- $h_i = \frac{\partial^2 l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial\left(\hat{y}_i^{(L-1)}\right)^2}$ (second derivative, Hessian)

The final prediction is the weighted sum of the outputs of all boosted trees:

$$\hat{y}_i = \sigma\left(\sum_{t=1}^{T} f_t(x_i)\right) \tag{9}$$

where σ is the sigmoid (for binary) or softmax (for multiclass) activation.

In diabetes prediction, XGBoost's gradient-based optimization efficiently handles nonlinear dependencies among glucose, BMI, HbA1c, and lipid variables. The L1/L2 regularization controls overfitting and stabilizes learning even when the dataset is imbalanced. Combined with SMOTE preprocessing, XGBoost contributed a strong base learner in the ML stack, offering interpretability via feature importance and SHAP-value explanations in future work.

4.6.1.2. Convolutional Neural Network (CNN): Hierarchical Feature Extraction from Clinical Patterns

The Convolutional Neural Network (CNN) used in the DL stack achieved 97.37% accuracy and 98.00% ROC-AUC, showing superior feature extraction capacity even on structured tabular data. CNN's strength lies in its ability to learn local dependencies and hierarchical representations from clinical attributes such as glucose, cholesterol, and renal biomarkers.

The general forward propagation of a CNN layer can be formulated as:

$$\begin{gathered} Z^{(l)} = W^{(l)} * X^{(l-1)} + b^{(l)} \\ A^{(l)} = f\left(Z^{(l)}\right) \end{gathered} \tag{10}$$

where:

- $W^{(l)}$ = kernel (filter) matrix in the l^{th} convolutional layer
- $*$ = convolution operation
- $b^{(l)}$ = bias term
- $f(\cdot)$ = activation function (ReLU used here)

For 1D convolution (suitable for tabular sequences):

$$Z_i^{(l)} = \sum_{k=0}^{K-1} W_k^{(l)} \cdot X_{i+k}^{(l-1)} + b^{(l)} \tag{11}$$

After convolution and activation, a max-pooling layer down samples the feature map:

$$P_j^{(l)} = \max_{i \in R_j} A_i^{(l)} \tag{12}$$

Finally, the flattened vector from the last convolutional block is passed through dense layers with dropout and sigmoid activation for binary or multiclass prediction:

$$\hat{y} = \sigma\left(W^{(L)} A^{(L-1)} + b^{(L)}\right) \tag{14}$$

In the hybrid soft voting ensemble, CNN acts as a deep pattern recognizer, capturing subtle feature correlations (e.g., between HbA1c, triglycerides, and BMI) that traditional ML models may overlook. Its high precision (98.60%) indicates a strong ability to minimize false positives vital for clinical reliability. However, CNN's lower recall (70.02%) reveals sensitivity to minority diabetic classes, which is later improved when combined with XGBoost and RF via stacking and soft voting, resulting in a balanced hybrid ensemble, as shown in Table 7. Conversely, as demonstrated in Table 9, XGBoost demonstrates robust recall capabilities via regularized gradient boosting, whereas CNNs exhibit elevated precision through the extraction of intricate nonlinear patterns. The Hybrid Soft Voting model, by averaging probabilities, integrates these complementary strengths, thereby attaining superior overall accuracy and generalization performance.

Table 9. Comparison table XGboost, CNN and hybrid soft voting ensemble

Model	Type	Key Strength	Equation Type	Role in Ensemble
XGBoost	Machine Learning	Gradient-boosted additive learning with regularization	Second-order Taylor optimization	High recall and interpretability
CNN	Deep Learning	Hierarchical nonlinear feature extraction	Convolution + ReLU + Pooling	High precision and deep representation learning
Hybrid (Soft Voting)	ML + DL	Probability averaging to balance bias and variance	Weighted probability aggregation	Highest All matrics and generalization

4.7 Strengths of the Proposed Hybrid AI Model

The proposed Hybrid Stacked Soft Voting Ensemble AI Model has a number of important strengths that make it a better way to predict diabetes classification. Because it uses both machine learning (Random Forest, Logistic Regression, and XGBoost) and deep learning (CNN, ENN, and FNN) models, it can capture both linear dependencies and complex nonlinear feature interactions. The stacking layer improves the predictions made by these base learners, and the soft voting mechanism makes sure that the best decision is made by averaging the probabilities instead of using a strict majority vote. This architecture improves generalization, reduces overfitting, and strikes a good balance between bias and variance [47]. The model is very reliable, sensitive, and easy to understand, with an accuracy of 98.89%, a precision of 98.99%, a recall of 87.07%, an F1-score of 92.05%, a ROC-AUC of 92.48%, and a Cohen's Kappa of 84.96%. Also, its ability to handle unbalanced data and its high agreement with ground truth make it especially useful in clinical settings where accuracy is very important.

4.8 Limitations of the Hybrid AI Approach

Although the hybrid approach has its merits, its weaknesses are particularly pronounced in terms of its computational complexity. It requires very large extra computational resources for training and evaluation with ML and DL models stacked together, and combined with soft voting. In resource-constrained environments like that of small clinics and developing countries, this also makes it hard to apply [48]. But even with GPU acceleration, the model's training time may still be too long for large datasets or real-time applications.

DL models (introduced as part of the hybrid approach) tend to need a large dataset and do not overfit, as it may lead to the ability to learn complex patterns. If this study could be performed using the LMCH dataset, the size should be sufficient; however, smaller datasets may not show the same performance. Transfer learning might help but may not work for all medical domains.

Yet combining ML and DL models brings with it extra complexity to the architecture, it has to be tuned more and is harder to implement. Such sophistication of models can create hurdles in adoption and maintenance for the less equipped health institutions. While the hybrid model performs superiorly in key metrics as compared with other models, it is likely hindered in some healthcare settings by its high computational demand, reliance on large datasets, and poor interpretability. The hybrid model, however, stands to make an important contribution to early diabetes prediction as long as related research continues and efficiency and explainability increase.

5.0 CONCLUSIONS

Against this backdrop, this research employs soft voting to develop a hybrid model based on the best ML and DL algorithms to predict diabetes. Finally, across all of the key metrics of interest such as accuracy, precision, recall, F1 score, ROC-AUC, and Cohen's Kappa, the hybrid model outperformed the individual ML or DL models and their stacked versions in a significant manner. We find that the Soft Voting Hybrid Model is the best-performing model, with an accuracy of 98.89%, Precision of 98.99%, Recall of 87.07%, F1-score of 92.05%, ROC-AUC of 92.48%, and Cohen's Kappa of 84.96%. Since this kept the level of false positives and false negatives balanced, it was perfect for medical diagnostics, especially when it was early diabetes detection. The predictive ability of the hybrid model to predict diabetes with great accuracy before clinical symptoms occur has significant implications for healthcare. Early detection is key, However, the model is very precise with few false positives, meaning the model reduces the number of patients being tested, as well as patient anxiety. Furthermore, validated on two independent datasets to generalize, it applies to many healthcare settings and populations. Further improvement in utility could be achieved by integrating it with electronic health records (EHR) systems so that it may be used in the routine risk assessment of diabetes.

The same stacking and soft voting methodology can be applied to future research with other medical conditions like cardiovascular disease or cancer. Key to improvement is reducing the model's computational complexity, integrating explainability techniques, and expanding the datasets for training to more ethnically diverse populations. In addition, the model could be applied in real-time to clinical settings to perform a continuous risk assessment and early intervention of patients. Furthermore, the model can be further optimized for deployment in resource-constrained settings to allow for wider use in healthcare settings around the world.

ACKNOWLEDGEMENTS

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.

CONFLICT OF INTEREST

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. The authors declare that they have no conflict of interest. This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit organizations. The authors declare that no direct data were collected from any patients. Instead, they utilized secondary data from publicly available datasets.

AUTHORS CONTRIBUTION

Md. Ziarul Islam: Conceptualization, Methodology, Validation, Formal Analysis, Visualization, Data Curation, Validation, Writing Original Draft, and Writing Review and Editing.

Mohd Khairul Azmi Bin Hassan: Mentoring and insightful guidance, writing review & editing, and validation.

Amir 'Aatieff Bin Amir Hussin: Review & Editing, Investigation, Formal Analysis.

H M Ikram Kays: Review & Editing, Formal Analysis; Data curation; Visualization,

REFERENCES

[1] World Health Organization. Diabetes [Internet]. Geneva: World Health Organization; 2024 Nov 14 [cited 2026 Jan 22]. Available from: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] Tran TT, Yun G, Kim S. Artificial intelligence and predictive models for early detection of acute kidney injury: transforming clinical practice. BMC Nephrol. 2024 Oct;25(1). doi:10.1186/s12882-024-03793-7.

[3] Katiyar N, Thakur HK, Ghatak A. Recent advancements using machine learning and deep learning approaches for diabetes detection: a systematic review. e-Prime Adv Electr Eng Electron Energy. 2024 Sep;9:100661. doi:10.1016/j.prime.2024.100661.

[4] Allal Z, Noura HN, Chahine K. Machine learning algorithms for solar irradiance prediction: a recent comparative study. e-Prime Adv Electr Eng Electron Energy. 2024 Mar;7:100453. doi:10.1016/j.prime.2024.100453.

[5] Altamimi A, et al. An automated approach to predict diabetic patients using KNN imputation and effective data mining techniques. BMC Med Res Methodol. 2024 Sep;24(1). doi:10.1186/s12874-024-02324-0.

[6] Ansari GA, Shafi S, Ansari MD, Shadab A. Advanced supervised machine learning methods for precise diabetes mellitus prediction using feature selection. Front Med. 2025 Sep;12. doi:10.3389/fmed.2025.1620268.

[7] Khokhar PB, Gravino C, Palomba F. Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review. Artif Intell Med. 2025 Apr;164:103132. doi:10.1016/j.artmed.2025.103132.

[8] Talukder MA, Talaat AS, Kazi M. HXAI-ML: a hybrid explainable artificial intelligence-based machine learning model for cardiovascular heart disease detection. Results Eng. 2025 Feb;104370. doi:10.1016/j.rineng.2025.104370.

[9] Abdelbaky I, Ahmed M, Taha M. Machine learning classification approaches for prediction of effective diabetes drugs. Egypt Inform J. 2025 Sep;31:100786. doi:10.1016/j.eij.2025.100786.

[10] Li W, Peng Y, Peng K. Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm. PLoS One. 2024 Sep;19(9):e0311222. doi:10.1371/journal.pone.0311222.

[11] Amer AA, Ravana SD, Ahamed R. Effective k-nearest neighbor models for data classification enhancement. J Big Data. 2025 Apr;12(1). doi:10.1186/s40537-025-01137-2.

[12] Shamshirband S, Fathi M, Dehzangi A, Chronopoulos AT, Alinejad-Rokny H. A review on deep learning approaches in healthcare systems: taxonomies, challenges, and open issues. J Biomed Inform. 2020 Nov;113:103627. doi:10.1016/j.jbi.2020.103627.

[13] Zafar MM, Khan ZA, Javaid N, Aslam M, Alrajeh N. From data to diagnosis: a novel deep learning model for early and accurate diabetes prediction. Healthcare. 2025 Aug;13(17):2138. doi:10.3390/healthcare13172138.

[14] Qamar T, Bawany NZ. Understanding the black-box: towards interpretable and reliable deep learning models. PeerJ Comput Sci. 2023 Nov;9:e1629. doi:10.7717/peerj-cs.1629.

[15] Rondón-Cordero VH, Montuori L, Alcázar-Ortega M, Siano P. Advancements in hybrid and ensemble ML models for energy consumption forecasting: results and challenges of their applications. Renew Sustain Energy Rev. 2025 Jul;224:116095. doi:10.1016/j.rser.2025.116095.

[16] Rustam F, et al. Enhanced detection of diabetes mellitus using novel ensemble feature engineering approach and machine learning model. Sci Rep. 2024 Oct;14(1). doi:10.1038/s41598-024-74357-w.

[17] Abousaber I, Abdallah HF, El-Ghaish H. Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets. Front Artif Intell. 2025 Jan;7. doi:10.3389/frai.2024.1499530.

[18] Kibria HB, Nahiduzzaman M, Goni MOF, Ahsan M, Haider J. An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. Sensors. 2022 Sep;22(19):7268. doi:10.3390/s22197268.

[19] Ijaz M, Alfian G, Syafrudin M, Rhee J. Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, SMOTE, and random forest. Appl Sci. 2018 Aug;8(8):1325. doi:10.3390/app8081325.

[20] Olorunfemi BO, et al. Efficient diagnosis of diabetes mellitus using an improved ensemble method. Sci Rep. 2025 Jan;15(1). doi:10.1038/s41598-025-87767-1.

[21] Ahmed N, et al. Machine learning-based diabetes prediction and development of smart web application. Int J Cogn Comput Eng. 2021 Jun;2:229–41. doi:10.1016/j.ijcce.2021.12.001.

[22] Fan Y. Diabetes diagnosis using a hybrid CNN-LSTM-MLP ensemble. Sci Rep. 2025 Jul;15(1). doi:10.1038/s41598-025-12151-y.

[23] Kumari S, Kumar D, Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. Int J Cogn Comput Eng. 2021 Jun;2:40–6. doi:10.1016/j.ijcce.2021.01.001.

[24] Afsaneh E, Sharifdini A, Ghazzaghi H, Ghobadi MZ. Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. Diabetol Metab Syndr. 2022 Dec;14(1). doi:10.1186/s13098-022-00969-9.

[25] Zhang W, Xia Z, Cai G, Wang J, Dong X. Enhancing diabetes risk prediction through focal active learning and machine learning models. PLoS One. 2025 Jan;20(7):e0327120. doi:10.1371/journal.pone.0327120.

[26] Dash S, et al. Privacy-preserving diabetes and heart disease prediction via federated learning and WCO. Int J Comput Intell Syst. 2025 Aug;18(1). doi:10.1007/s44196-025-00956-8.

[27] Bejugam SK, Vankara J. An efficient model for diabetic detection using heuristic approach-based serial cascaded convolutional ensemble network. Artif Intell Rev. 2025 Aug;58(10). doi:10.1007/s10462-025-11334-3.

[28] Alzubaidi L, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data. 2021 Mar;8(1):1–74. doi:10.1186/s40537-021-00444-8.

[29] Rihan A, Anbar M, Alabsi BA. Meta-learner-based approach for detecting attacks on Internet of Things networks. Sensors. 2023 Sep;23(19):8191. doi:10.3390/s23198191.

[30] Kaggle. Pima Indians diabetes database [Internet]. 2016 [cited 2026 Jan 22]. Available from: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[31] Ziya. Diabetes clinical dataset (100k rows) [Internet]. Kaggle; 2025 [cited 2026 Jan 22]. Available from: https://www.kaggle.com/datasets/ziya07/diabetes-clinical-dataset100k-rows

[32] Python Software Foundation. Python [Internet]. 2025 [cited 2026 Jan 22]. Available from: https://www.python.org/

[33] Scikit-learn. Importance of feature scaling [Internet]. [cited 2026 Jan 22]. Available from: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html

[34] Sonia JJ, Jayachandran P, Md AQ, Mohan S, Sivaraman AK, Tee KF. Machine-learning-based diabetes mellitus risk prediction using multi-layer neural network No-Prop algorithm. Diagnostics. 2023 Jan;13(4):723. doi:10.3390/diagnostics13040723.

[35] Kim GI, Kim S, Jang B. Classification of mathematical test questions using machine learning on datasets of learning management system questions. PLoS One. 2023 Oct;18(10):e0286989. doi:10.1371/journal.pone.0286989.

[36] Singh K, Mahajan A, Mansotra V. 1D-CNN-based model for classification and analysis of network attacks. Int J Adv Comput Sci Appl. 2021;12(11). doi:10.14569/ijacsa.2021.0121169.

[37] Han Y, Kim DY, Woo J, Kim J. Glu-Ensemble: an ensemble deep learning framework for blood glucose forecasting in type 2 diabetes patients. Heliyon. 2024 Apr;10(8):e29030. doi:10.1016/j.heliyon.2024.e29030.

[38] Liu Q, et al. Development and validation of a meta-learner for combining statistical and machine learning prediction models in individuals with depression. BMC Psychiatry. 2022 May;22(1). doi:10.1186/s12888-022-03986-0.

[39] Khan AA, Chaudhari O, Chandra R. A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation. Expert Syst Appl. 2023 Dec;244:122778. doi:10.1016/j.eswa.2023.122778.

[40] Kuo AT, Chen H, Tang L, Ku W, Qin X. ProbSky: efficient computation of probabilistic skyline queries over distributed data. IEEE Trans Knowl Data Eng. 2022 Jan. doi:10.1109/TKDE.2022.3151740.

[41] Liu X, Yang DN, Ye M, Lee WC. U-Skyline: a new skyline query for uncertain databases. IEEE Trans Knowl Data Eng. 2013 Apr;25(4):945–60. doi:10.1109/TKDE.2012.33.

[42] Chen B, Zhu D, Wang Y, Zhang P. An approach to combine the power of deep reinforcement learning with a graph neural network for routing optimization. Electronics. 2022 Jan;11(3):368. doi:10.3390/electronics11030368.

[43] Islam MZ. Enhancing diabetes prediction accuracy using stacked machine learning and deep learning models: a public health approach. Indones J Comput Sci. 2025 Aug;14(4). doi:10.33022/ijcs.v14i4.4947.

[44] Kumar P. A comparison of CNN, RNN, and FNN algorithms to investigate effective diabetes prediction. Arch Comput Methods Eng. 2025 Sep. doi:10.1007/s11831-025-10333-5.

[45] Oliullah K, Rasel M, Islam MM, Islam MR, Anwar M, Whaiduzzaman M. A stacked ensemble machine learning approach for the prediction of diabetes. J Diabetes Metab Disord. 2023 Nov. doi:10.1007/s40200-023-01321-2.

[46] Barton M, Lennox B. Model stacking to improve prediction and variable importance robustness for soft sensor development. Digit Chem Eng. 2022 Jun;3:100034. doi:10.1016/j.dche.2022.100034.

[47] Halder RK, Lima MA, Uddin MN, Islam MA, Saha A. Integrated feature selection-based stacking ensemble model using optimized hyperparameters to predict breast cancer with smart web application. Clin eHealth. 2025 Aug. doi:10.1016/j.ceh.2025.08.001.

[48] Razzaq K, Shah M. Next-generation machine learning in healthcare fraud detection: current trends, challenges, and future research directions. Information. 2025 Aug;16(9):730. doi:10.3390/info16090730.