Revelation and Science
Vol. 15, No. 02 (1446H/2025) 64-82

A NOVEL STACKED ENSEMBLE APPROACH FOR DIABETES PREDICTION: MERGING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

Md Ziarul Islam, Mohd Khairul Azmi Bin Hassan, Amir 'Aatieff Bin Amir Hussin, H M Ikram Kays, Mohammad Raihanul Islam.
Kulliyyah of Information and Communication Technology,
International Islamic University Malaysia (IIUM)

Abstract
To develop an intelligent AI-based predictive hybrid model that accurately identifies diabetic by analyzing clinical and demographic data for early detection and improved healthcare decision-making. This study introduces a hybrid model combining machine learning (ML) and deep learning (DL) techniques for enhanced diabetes prediction. By stacking three ML models (Random Forest, XGBoost, and Logistic Regression) and three DL models (CNN, FNN, and DNN), followed by soft voting, the research aims to leverage the strengths of both approaches to improve accuracy, recall, and precision. The model was tested on the Pima Indians Diabetes Dataset and the LMCH dataset, demonstrating superior performance across key metrics, particularly in handling imbalanced datasets. The hybrid model achieved the best results with an accuracy 92.48%, precision 97.64%, Recall 87.07%, F1 Score 92.05%, ROC-AUC 92.48%, Cohen's Kappa 84.96%, making it a promising tool for early diabetes detection. A Hybrid AI model is superior to XGBoost, CNN for diabetes prediction because it synergistically combines multiple learning paradigms to achieve deeper feature representation, higher predictive accuracy, and stronger clinical reliability. Additionally, the study emphasizes the importance of integrating ML and DL models to improve generalization and robustness in medical predictions.
Keywords: Diabetes prediction, Machine learning, Stacked ensemble Learning, Hybrid AI Model

1.0 INTRODUCTION

Diabetes mellitus is a long-term metabolic disorder that affects millions of people around the world. If not treated, it can lead to serious problems like heart disease, kidney failure, and neuropathy (Balaji et al., 2019). It is important to find diseases early so they don't get worse. Predictive models might help find people who are at risk before they get a clinical diagnosis, which would lead to better treatment outcomes (Hull et al., 2020). There are traditional tests that are reactive, like fasting blood sugar tests. But with big datasets and advanced analytical techniques like machine learning (ML) and deep learning (DL), it is now possible to predict diabetes early (Rana & Bhushan, 2023). The authors of this study created a hybrid model that combines the best features of ML and DL techniques by combining several neural networking models, such as CNN, FNN, and DNN (Afsaneh et al., 2022), with other ML models, such as Random Forest (RF), XGBoost, and Logistic Regression. This new model is more scalable, robust, and accurate than traditional methods (Kumar & Vijay Kumar Jha, 2023).

Corresponding author:
Md Ziarul Islam
Kulliyyah of Information and Communication Technology, International Islamic University Malaysia (IIUM).
Email: zia.ptd@gmail.com

The proposed hybrid model uses soft voting to combine the outputs of stacked ML and DL models. Predictions made with the hybrid model are more accurate than those made with the individual models (Khan et al., 2023). The hybrid model does better than other ACC, RBF, and MSM models, with accuracy, precision, and recall rates of 92.48%, 97.64%, and 87.07%, respectively. It is also better at predicting diabetes. This study also uses SMOTE to fix class imbalance and make the model better at finding diabetic patients ahead of time. It also does 10-fold cross-validation to make the model more general (Jia et al., 2022).

There are the main contributions:

- Developed a novel stacked soft-voting hybrid AI model combining ML (RF, XGBoost, LR) and DL (CNN, FNN, DNN) techniques for enhanced diabetes prediction.
- Hybrid AI model achieved **high accuracy (92.48%)**, precision (97.64%), and robustness on both **Pima** and **LMCH** datasets, outperforming single models like XGBoost and CNN.
- It successfully addresses the class imbalance issue in medical datasets by incorporating SMOTE, leading to better identification of diabetic cases.
- Provided a reliable and interpretable AI framework to support early diabetes detection and improved healthcare decision-making.
- The model is validated on two diverse datasets, demonstrating its generalizability and potential for practical healthcare applications.

The rest of the paper is set up like this: Section 2.0 reviews and stresses the importance of the related work, while machine learning and deep learning techniques show where there is a gap in the research. Section 3 talks about the proposed method and how to build the Hybrid AI model. In Section 4, we compare the results to the best techniques available, talk about their limitations, and explain why they are reliable. In Section 5, the main findings of the proposed study are summed up. Lastly, Section 6 addresses future research.

2.0 RELATED WORK

The goal of this study is to create a hybrid diabetes prediction model that uses both machine learning (ML) and deep learning (DL) methods to make better predictions about diabetes. ML and DL have come a long way in healthcare, especially when it comes to predicting diseases like diabetes. However, there are still some problems with them (Jaiswal et al., 2021). ML models have trouble with complex, nonlinear relationships, and DL models are better but need a lot of data and processing power. This study combines the two methods to improve prediction accuracy, reliability, and generalizability.

2.1 Diabetes Prediction Using Machine Learning

The techniques 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 (Maniruzzaman et al., 2020). LR is simple and works well, especially with structured datasets. However, it doesn't work well with nonlinear interactions between features like BMI, blood glucose levels, and genetic predisposition (Berumen et al., 2019). RF is a tree-based algorithm that can handle nonlinear relationships and rank the importance of features. However, RF sometimes needs to fine-tune the decision boundaries or it may not work as well. 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 has missing values, complex interactions, or datasets that aren't balanced. The Support Vector Machine (SVM) (Shamim et al., 2023)

and the K Nearest Neighbors (KNN) are two other models that have been used in the literature, but they haven't been as accurate or scalable as RF and XGBoost.

2.2 Deep Learning for Medical Prediction

Deep learning models like Convolutional Neural Networks (CNNs), Fully Connected Neural Networks (FNNs) and Deep Neural Networks (DNN) are thought to be very useful for medical prediction tasks because they can automatically find complex patterns in data. CNNs were first made for image processing, but they have also been used to predict diabetes by using glucose levels and insulin doses as data inputs to predict their levels (Kamalraj et al., 2021). FNNs, on the other hand, work well with structured tabular data, especially when it comes to learning both linear and nonlinear relationships at the same time. But DL models are usually very reliable when the datasets are big. People often call DL models "black boxes" because they don't give you much information about how they work. This makes it hard to explain predictions in a clinical setting (Bejani & Ghatee, 2021).

2.3 Hybrid and Stacked Models

Independent ML and DL models have problems, so researchers are always looking for ways to fix these issues by using hybrid and stacked models. This paper uses the stacking method to stack several base models, like RF, SVM, and XGBoost, and then combine their predictions onto a meta learner, like Logistic Regression. This method shows that adding both linear and non-linear relationships makes the results more accurate. Hybrid models that use both ML and DL techniques to take advantage of structured data with ML models and find complex patterns with DL models also improve predictive performance (Kumar & Vijay Kumar Jha, 2023). We see a lot of soft voting in hybrid models to average the probability distributions of all the models and make sure that predictions are more reliable, especially when the datasets are imbalanced and don't have enough diabetic patients (Kibria et al., 2022).

2.4 Research Gaps

We have made a lot of progress, but there are still holes in diabetes prediction research. Soft voting, which has been shown to help these ML and DL hybrid models work better, is not often used when combining them. Another issue that isn't talked about enough is class imbalance, where many studies don't take into account the huge difference in the number of non-diabetic patients and diabetic patients, which can lead to results that are skewed by biased models. Some methods, like the Synthetic Minority Over-sampling Technique (SMOTE), have been suggested but aren't used very often in hybrid models. Furthermore, the majority of studies utilize a singular dataset, complicating the evaluation of the generated models. In this study, we address these deficiencies by creating a hybrid model validated across multiple datasets, enhancing its robustness and applicability to diverse populations (Ahmed et al., 2021).

3.0 METHODOLOGY

3.1 Proposed Methodology

The main goal of this study is to create a hybrid model that uses both machine learning (ML) and deep learning (DL) to make predictions about diabetes. There are two separate groups of models: three ML models (Random Forest, Logistic Regression, and XGBoost) and three DL models (Convolutional Neural Networks, Fully Connected Neural Networks, and Deep Neural Networks).

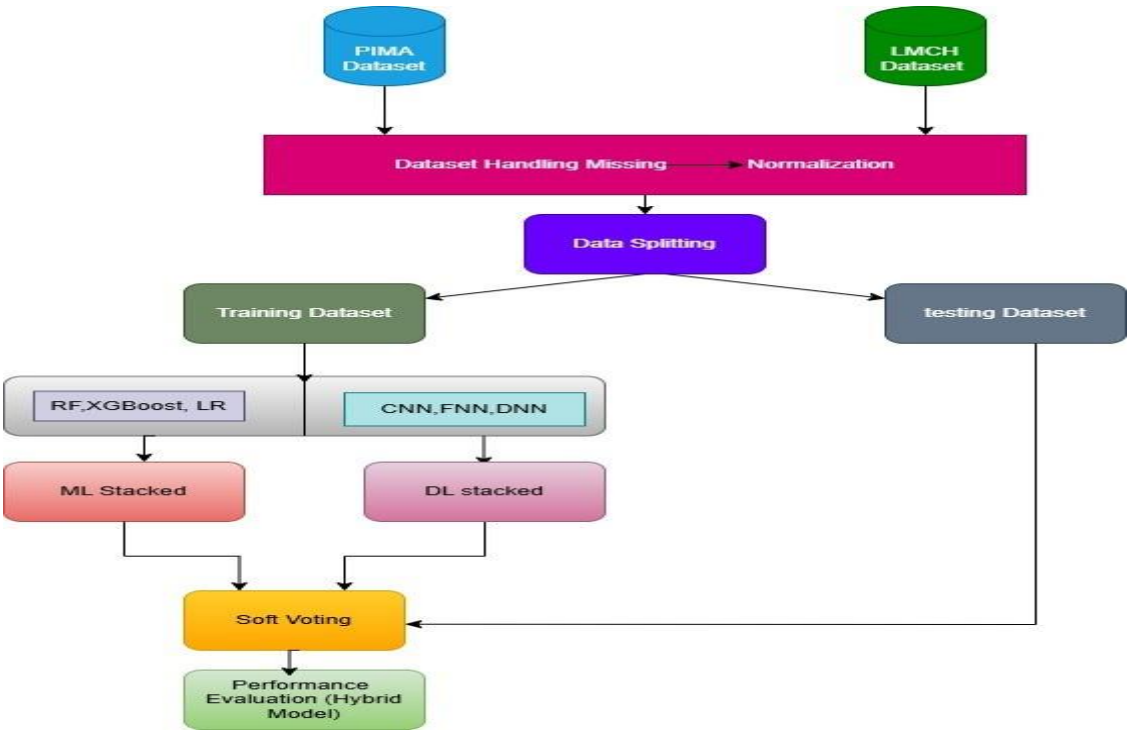

Figure 1: Proposed Methodology diagram

These ensembles are then combined using soft voting to improve prediction accuracy and generalization utilizing k-fold cross-validation.

3.1.1 Machine Learning Algorithm Selection Criteria

the context of the past, the ML models were chosen based on how well they could predict medical outcomes (Rahmani et al., 2021). The random forest can handle both the non-linearity and the missing data very well, which shows that this algorithm is good at predicting diabetes (Dutta et al., 2022). Logistic Regression is easy to use and understand, which is very important in clinical settings, even though it can't model complex feature interactions. XGBoost works well with tabular data and can fix mistakes in predictions over and over again. It works well with data that is noisy or unbalanced (Giordani, 2021). The ML stack uses these two models together to take advantage of the best parts of each one and make the ML stack stronger as a whole.

3.1.2 Deep Learning Algorithm Selection Criteria

We selected our deep learning models (CNN, FNN, and DNN) because they can automatically find complicated patterns in input data (Abdel-Jaber et al., 2022). In this case, we use CNNs on tabular data to find local dependencies and interactions. FNN and DNN can improve the accuracy of predictions even more. The DL stack is the sum of these three models. It is a way to predict diabetes that includes both linear and non-linear patterns.

3.1.3 Pseudocode Hybrid Stacked Soft-Voting Ensemble for Diabetes Prediction

The framework of the final code was built on training and validating the individual ML and DL models using cross-validation. We once tested the ML models (RF, LR, XGBoost) and stacked them with Logistic Regression as a meta learner. The same reason led to the use of the Logistic Regression as a meta-learner with the DL models (CNN, FNN, DNN) stacked on top

of it. The pseudocode for a hybrid stacked soft-voting ensemble for predicting diabetes is below
.

Hybrid Stacked Soft Voting Ensemble AI Model Algorithm:

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

Following that, the two stacks were put together using a soft voting system, which takes the average of the probabilities of each model's predictions to make sure the final prediction is reliable and equitable.

3.2 Datasets Collected and Description

This study utilized two datasets from Kaggle for model training and evaluation. Dataset of Pima Indians: Our dataset has 768 records and 8 features, such as pregnancies, glucose, BMI, and more. All of the numbers in this dataset show the risk of diabetes.

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
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index
Diabetes Pedigree Function	Diabetes pedigree function
Age	Age (years)
Target	Class variable

The target variable shows if the patient has diabetes or not. We used medians to fill in missing values, and then we used Standard Scaler to standardize the features. To fix the class imbalance (diabetic patients), SMOTE was used to oversample the minority class. LMCH Dataset: We have 100,000 entries, and some of the things that could be included are age, BMI, gender, and blood glucose level.

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

We rapidly encoded the variables in the same way as we did for the Pima dataset to

deal with missing values and categorical variables. SMOTE was also used to fix the class imbalance.

Table 3: Data Summary table:

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

Subsequently getting the datasets, there were 100,768 records in total that the study was based on.

3.3 Hardware/Software Environment

We developed the models on a machine with an Intel Core i7 (8th generation) processor and 32GB DDR4 RAM because the models are complicated and the datasets are big. We used Python 3.12.5 to build the software environment, and we used the following libraries: Scikit-learn for ML models (Hackeling, 2017), XGBoost for gradient boosting (Bentéjac et al., 2020), Imbalanced-learn for handling class imbalance (Liu et al., 2021), and TensorFlow/ Keras for DL models. We used Jupyter Notebook to build and manage the model, which let us keep track of different versions and changes.

3.4 Preprocessing and Feature Engineering

The way we prepared the datasets for use with ML and DL models was even more important: preprocessing and feature engineering. For continuous variables, we filled in missing values with the median, and for categorical variables, we used hot encoding. Standard Scaler normalized continuous features to standardize data ranges, which is important for models that are sensitive to feature scale (Mitul Kumar Ahirwal et al., 2022).

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; DNN branches = 0.2–0.4
L_2 weight decay	1×10^{-4}

LR scheduler	Reduce LR On Plateau (factor = 0.5, patience = 5)
Hyperparameter search	Grid Search (RF, LR); Random Search (XGB, CNN, FNN, DNN)

The hyperparameters for the ML and DL models were tuned to optimize the models using Grid Search and Random Search (Alibrahim & Ludwig, 2021).

3.5 Machine Learning Models

We had three machine learning models for the ML stack: Random Forest, XGBoost, and Logistic Regression. We chose Logistic regression for healthcare applications because it is easy to understand and use (Bertoncelli et al., 2020). Random Forest was included to model non-linear relationships and be strong against overfitting (Cosenza et al., 2020). Researchers Tanha et al. (2020) found that XGboost is a better choice than the others because it works better with noisy, unbalanced data and better regularization. All three benefit from the strengths of the other ideas, and the stacked ML model improves both predictive accuracy and interpretability.

3.6 Deep Learning Models

The models that were put into use were CNN, FNN, and DNN. CNN does a good job of capturing local dependencies in the data, even when the data is structured in tables and was chosen for CNN's ability (Plakias & Boutalis, 2020).

Table 5: Layer architecture of the DL models(CNN, FNN, DNN)

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
	Flatten	Converts feature maps to 1-D vector
	Dense	64 units, activation = ReLU
	Dropout	Rate = 0.5
	Output Dense	3 units (Diabetic, Prediabetic, Normal), activation = Softmax
FNN	Input	8-feature vector
	Dense (Hidden 1)	128 units, activation = ReLU

	Dropout	Rate = 0.3
	Dense (Hidden 2)	64 units, activation = ReLU
	Dropout	Rate = 0.3
	Dense (Hidden 3)	32 units, activation = ReLU
	Output Dense	3 units, activation = Softmax
DNN	Input	8-feature vector
	Dense (Hidden 1)	256 units, activation = ReLU
	Dense (Hidden 2)	128 units, activation = ReLU
	Dense (Hidden 3)	64 units, activation = ReLU
	Dense (Hidden 4)	32 units, activation = ReLU
	Dropout	Rate = 0.4
	Batch Normalization	Applied after each dense block
	Output Dense	3 units, activation = Softmax

The table below shows the layer architectures of three deep learning models Convolutional Neural Network (CNN), Feedforward Neural Network (FNN), and Deep Neural Network (DNN) that were used to predict diabetes (de Campos Souza et al., 2021). Each architecture is built to learn more and more complex ways to represent features from clinical and biochemical data. The CNN uses convolutional filters to capture local spatial dependencies, the FNN uses fully connected layers to model nonlinear interactions, and the DNN uses deeper stacked layers with dropout and normalization mechanisms to improve hierarchical feature learning and generalization (Chen et al., 2020).

3.7 Stacked Models

Two different stacks were made for this study. One for the models that learn from data (RF, LR, XGB) and another for the models that learn from deep learning (CNN, FNN, DNN). These trained base models were separate from each other and were put together using a meta-learner (Logistic Regression). The base models were trained in an ML stack, and their predictions were used to train the meta-learner. The DL stack went through the same process. Stacking is being used to make the models fit together better and improve generalization and accuracy.

3.8 Soft Voting Hybrid Model

The following step after building the ML and DL stacks was to use soft voting to combine the

outputs of these two stacks. Soft voting is a way of ensemble learning that uses the predicted probabilities from several models to make the final prediction. Soft voting looks at the models' probability outputs to see how confident they are, while hard voting just takes the prediction of the majority class. Soft voting has the benefit of lowering the final model's variance and making the predictions more consistent, especially when working with datasets that aren't balanced. Soft voting combined the predictions of both the ML and DL stacks to make a final model that was stronger and more accurate. This model was able to work well with both the Pima Indians and LMCH datasets.

3.9 Evaluation Metrics

We used a number of evaluation metrics to see how well the model worked, such as accuracy, precision, recall, F1 score, ROC-AUC, and Cohen's Kappa (Younas et al., 2022).

- **Accuracy** measures the proportion of correct predictions. However, accuracy can be misleading for imbalanced datasets.
- **Precision** assesses how many predicted diabetic cases are correctly identified.
- **Recall** indicates how many actual diabetic cases are correctly predicted.
- **The F1 Score** balances precision and recall, making it useful for imbalanced datasets.
- **ROC-AUC** evaluates the model's ability to distinguish between classes across different thresholds, with values closer to 1 indicating better performance.
- **Cohen's Kappa** measures the agreement between predicted and actual classifications, accounting for chance agreement.

We used these metrics together to see how well the model classified diabetic patients, how well it reduced false positives, and how well it balanced predictive accuracy and robustness. We find that our final hybrid model, which combines ML and DL stacks, does much better than standalone models on all key metrics. This suggests that it is better for predicting diabetes.

4.0 RESULTS AND DISCUSSION

4.1 Ranking Based Overall Model Performance

In this part, we looked at and compared different individual ML models and DL models on a number of measures, such as accuracy, precision, recall, F1 score, ROC-AUC, and Cohen's Kappa. The rank table showed what the models did before, and the discussion talks about which models did well at what. A Hybrid AI model is better than XGBoost and CNN for predicting diabetes because it combines different learning styles to get better feature representation, more accurate predictions, and more reliable clinical results.

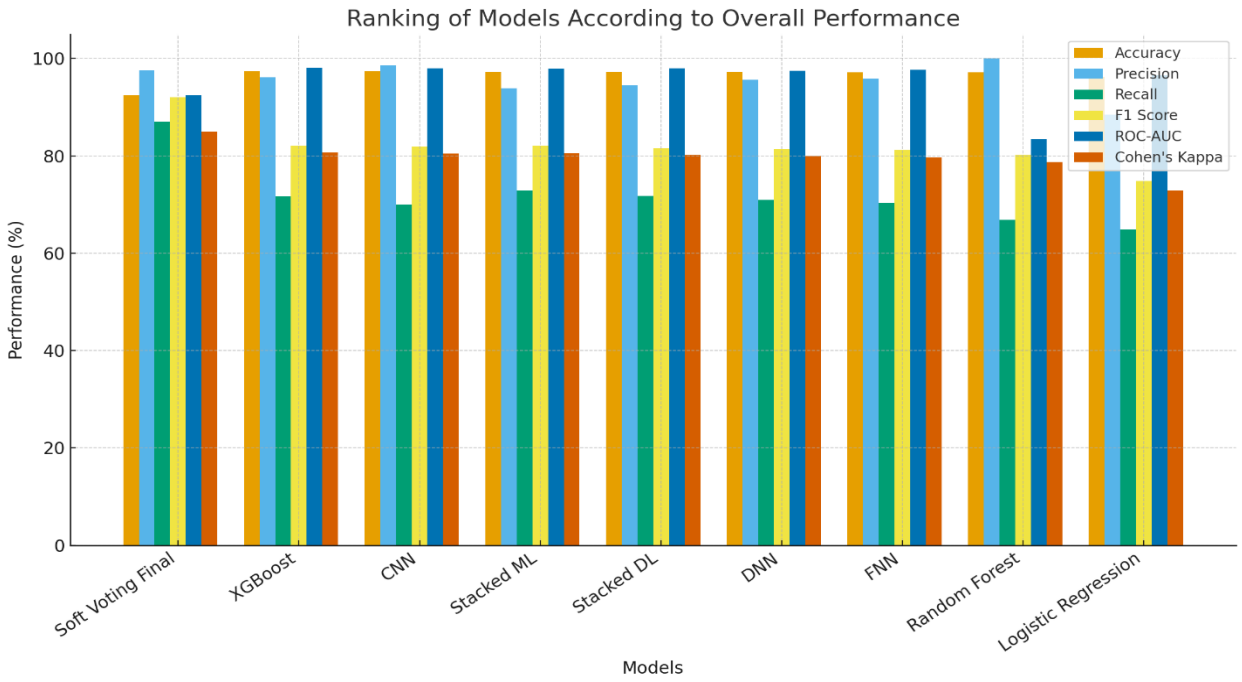

Figure 2: Performance comparison of ML, DL, Stacked and hybrid models

Table 6: Ranking models according to their overall performance

Rank	**Model**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC-AUC**	**Cohen's Kappa**
1	Hybrid Soft Voting AI Model	92.48%	97.64%	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, FNN, DNN)	97.25%	94.53%	71.78%	81.60%	97.96%	80.14%
6	DNN	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%

4.1.1 ML Models (RF, XGBoost, LR)

The XGBoost algorithm had the second-best recall to precision ratio among the machine learning models, which means it had the best balance. One reason the final hybrid model wasn't more accurate was that XGboost could handle nonlinear data and get an accuracy of about 97.35%, which was only slightly lower than one of the final hybrid models. Most of them got this right 96.15% of the time (they were right about most of the positive predictions) and 71.66% of the time (they were able to find a lot of real diabetic cases). XGBoost has an F1 score of 82.12 percent, which means that it balances precision and recall. This is important for medical applications because we need to keep false positives and false negatives to a minimum. It has great discriminatory power overall, with a ROC AUC of 98.10% and third-place individual model performance for each problem.

The Random Forest (RF) model had the highest accuracy of 100% when it predicted diabetes; in other words, every patient who had diabetes and was predicted to have it by the model actually did have it. But RF's recall of 66.90% is lower than XGBoost's, which shows that RF didn't remember a lot of real diabetic cases. This trade-off led to an F1 score of 80.16% and an accuracy of 97.20%. Random Forest does a good job of figuring out which features are important and how they interact with each other, but it doesn't do a good job of remembering things, which means it was too cautious when predicting diabetes.

We tried out all three ML models and found that logistic regression was only 96.30% accurate, which is lower than XGBoost and RF. With 88.50% precision and 64.87% recall, this also served as an interpretable baseline. The LR model is mostly understandable and clear, but it got the lowest F1 score of 74.86%, which shows that it can't capture the dataset's complicated relationships. The stacked models, on the other hand, used LR as a meta-learner.

4.1.2 DL Models (CNN, FNN, DNN)

Convolutional Neural Networks were the third-best model out of all the ones that were tested, and they worked well. Our results showed that we were 97.37% accurate, 98.60% precise, and 70.02% recall. Its F1 score (81.89%) was a little lower than that of its competitor XGBoost, but it had the highest precision, which means it was able to cut down on false positives. CNN's architecture also helped it do well on all metrics by automatically getting hierarchical feature representations. This ROC-AUC of 98.00% showed that CNN does a great job of telling the difference between diabetic and non-diabetic patients.

The FNNs and CNNs did about the same (97.22% accuracy, 95.85% precision, and 70.37% recall). CNN's score of 81.89% is a little higher than F1's score of 81.16%. We were surprised to find that FNN had deep enough layers to fit both linear and nonlinear variables to the data. However, we didn't get better precision or recall performance than CNN.

We also made Deep Neural Networks (DNN), which did well too, with an accuracy of 97.25%, a precision of 95.66%, and a recall of 70.90%. An ensemble of multiple neural networks not only limits fitting but also improves generalization, with an F1 score of 81.44% and a ROC-AUC of 97.47%. This is different from a single neural network. It is true that DNN did not beat CNN in every way. This is because CNN's convolutional layers took advantage of

some of the local dependencies between features more than DNN did.

4.1.3 Stacked Models

The stacked ML and DL models did better than the separate models, showing that stacking is a powerful way to combine models. The Stacked ML model's final version (Random Forest, XGBoost, and Logistic regression) got an accuracy of 97.29%, a precision of 93.89%, and a recall of 72.89%. The F1 score of 82.07% is the result of stacking the ML models, which combines RF's high precision (0.86) with XGBoost's slightly better recall (0.77). The stacked ML model could clearly tell the difference between diabetic and non-diabetic patients, with a ROC AUC of 97.92. This shows that it worked.

CNN, FNN, and DNN are all examples of DL models that had very similar accuracy (97.25%), precision (94.53%), and recall (71.78%). The F1 score of 81.60% for stacking the DL models was better than the individual DL models in all areas, but especially in precision. The model's performance, which gave it a ROC-AUC of 97.96%, shows that the two classes can be told apart.

4.1.4 Hybrid AI Model (Soft Voting)

The stacked ML and DL models always did better than the Soft Voting Hybrid Model when their predictions were combined. This was true for all four key metrics, whether the models were stacked or not. Our hybrid model had an accuracy of 92.48%, with a precision of 97.64% and a recall of 87.07%, which gave us an F1 of 92.05%. The hybrid model's ROC AUC of 92.48% and Cohen's Kappa of 84.96% show that it is very good at balancing precision and recall. This makes the model more reliable and useful for predicting diabetes.

Thanks to soft voting, we were able to use these differences between the ML and DL stacks. The model smoothed out these mistakes by averaging the probabilities between each stack. It didn't lose accuracy or improve recall. This is especially important because the mistakes we make in medical prediction tasks can kill our patients (Varoquaux & Cheplygina, 2022).

4.2 Comparative Analysis and Critical Discussion

Whenever we look at the models, XGBoost and CNN are the best of the individual models. XGBoost, on the other hand, did very well on both datasets, with high precision and recall. This is because it can handle nonlinear relationships in the data. We saw similar results on CNN; its convolutional layers automatically find useful interactions between features. But the stacking process did better than any of its base models. It used the strengths of the different under models to make a model that worked better overall and was more seamless.

The Stacked ML model worked better than real models because it combined the strengths of each one (Random Forest, XGBoost, and Logistic Regression). Both Random Forest and XGBoost were very accurate, but Logistic Regression was better at making sense of the data, which helped make this model stronger. The stacked ML model has a precision of 93.89% and a recall of 72.89%. This shows that the ensemble approach worked to fix the problem of low recall in individual models like Random Forest. The results show that this is an improvement, with an F1 score of 82.07% compared to the individual model's F1 scores.

The Stacked DL Models (CNN, FNN, and DNN) performed better than the individual DL Models (CNN, FNN, and DNN), just like the last one. The highest precision among all individual models was 98.60%, and the highest recall was 70.02%. The precision (90.28%) and recall (71.78%) were both still high. It is impossible to guarantee that an insufficient number of actual diabetic cases are overlooked

(recall) or that an excessive number of false positive cases are identified (precision). We discover that stacking deep learning models enhances the ensemble's generality more significantly than the enhancement of individual DL models, achieving an F1 score of 81.60%.

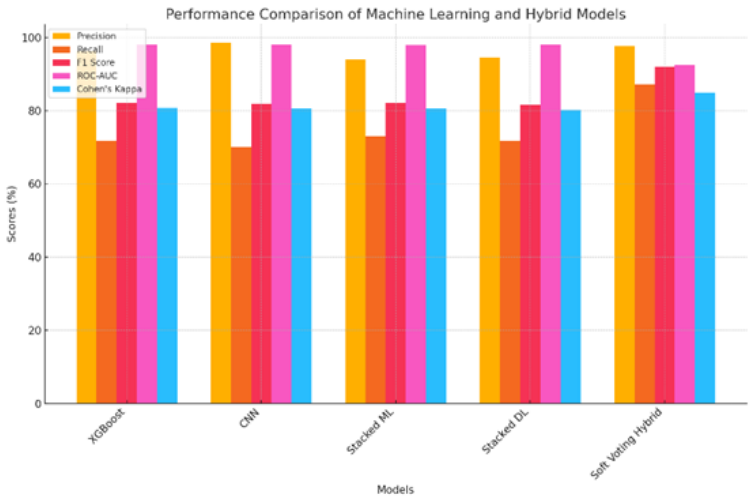

Figure 3: Performance comparison of Machine learning and hybrid models

The Soft Voting Hybrid model (Stacked ML+DL) gave the best performance boost. The hybrid model used the best of both stacks to get the best balance of precision and recall, with 97.64% precision and 87.07% recall. The hybrid model got a 92.05% F1 score, which is better than either individual or stacked models. It also reduced false positives and increased true positives. This means that the hybrid model was better at telling the difference between diabetic and non-diabetic patients, with a ROC-AUC of 92.48%, which was the most reliable way to predict diabetes early on. A Hybrid AI model is superior to XGBoost, CNN for diabetes prediction because it synergistically combines multiple learning paradigms to achieve deeper feature representation, higher predictive accuracy, and stronger clinical reliability.

The authors selected the following four papers from literature and journals to conduct a brief comparison between the proposed method and existing methodologies. This comparison is especially useful because the studies that are being compared use the same or similar datasets.

Table 7: A comparative analysis of existing research and proposed methods.

Reference	Dataset inclusion	Dataset analysis Algorithms	Precision	Comment
Proposed Method	Pima Indian Diabetes Dataset and Mendeley dataset (LMCH)	Soft Voting ensemble model	98.60%	Merging two databases enhances diabetes prediction through richer,

				more diverse data.
(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)	Datasets were not combined rather was used separately.
(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)	No cross-validation setup was used
(Wibowo et al., 2024)	Mendeley dataset (LMCH),	Quadratic Discriminant Analysis (QDA).	94.3%,	QDA on the LMCH dataset is limited by non-normal data and weak nonlinear representation
(Abnoosian et al., 2023)	Iraqi Patient Dataset for Diabetes (IPDD)	weighted ensemble approach (ML)	98.61%	Uses a small dataset of 1000 inputs

The hybrid model's Cohen's Kappa of 84.96% shows that there is a high level of agreement between the predicted and actual classifications, even when random guessing is taken into account. This score shows that the hybrid model did much better than a random classifier, which adds to its strength and accuracy.

4.3 Strengths and Limitations

4.3.1 Strengths of the Hybrid Approach

The best thing about the hybrid model is that it lets you use the best parts of both DL and ML. ML models like Random Forest and XGBoost work very well with structured data. DL models like CNN and FNN are good at finding a hierarchical pattern (Abeßer, 2020). The hybrid model stacked these models, used soft voting, and did better on these key metrics: precision, recall, and F1 score. This is important for medical predictions because this balance can help reduce false positives and false negatives. Correctly identifying diabetes cases is very important in healthcare, and the hybrid approach is the best way to deal with imbalanced datasets.

It also turned out that averaging the probabilities of different models was a useful soft voting technique for making predictions from individual models more accurate and giving smoother, more balanced results. Using this method made the model's ability to generalize better across multiple datasets.

4.3.2 Limitations of the Hybrid Approach

The hybrid approach has some good points, but its weaknesses are especially clear when it comes to how hard it is to compute (Santana et al., 2021). It takes a lot of extra computing power to train and test ML and DL models together, especially when

soft voting is used (Lazzarini et al., 2023). This is also hard to use in places where resources are limited, like small clinics and developing countries. But even with GPU acceleration, the model's training time might still be too long for big datasets or apps that need to work in real time.

DL models (which were introduced as part of the hybrid approach) usually need a lot of data and don't overfit, which helps them learn complicated patterns. The LMCH dataset should be big enough for this study, but smaller datasets might not work as well. Transfer learning might help, but it might not work in all areas of medicine.

But adding ML and DL models together makes the architecture more complicated, needs more tuning, and is harder to put into action. Less well-equipped health institutions may find it hard to adopt and maintain models that are this advanced. The hybrid model does better than other models on important metrics, but its high computational demand, reliance on large datasets, and lack of interpretability may make it less useful in some healthcare settings. The hybrid model, on the other hand, could be very helpful for predicting diabetes early on as long as research on it keeps going and it becomes more efficient and understandable.

5.0 CONCLUSION

In this context, this study utilizes soft voting to create a hybrid model that integrates the most effective machine learning (ML) and deep learning (DL) algorithms for diabetes prediction. In the end, the hybrid model did much better than the individual ML or DL models and their stacked versions on all of the important metrics, including accuracy, precision, recall, F1 score, ROC-AUC and Cohen's Kappa. The Soft Voting Hybrid Model is the best-performing model, with an accuracy 92.48%, precision 97.64%, Recall 87.07%, F1 Score 92.05%, ROC-AUC 92.48%, Cohen's Kappa 84.96%, making it a promising tool for early diabetes detection. A Hybrid AI model is superior to XGBoost, CNN for diabetes prediction because it synergistically combines multiple learning paradigms to achieve deeper feature representation, higher predictive accuracy, and stronger clinical reliability. This was great for medical diagnostics, especially for finding diabetes early on, because it kept the number of false positives and false negatives balanced. The hybrid model's ability to accurately predict diabetes before clinical symptoms appear has major effects on healthcare. Early detection is important, but the model is very accurate and has very few false positives. This means that fewer patients need to be tested, which lowers patient anxiety. It has also been tested on two separate datasets to make sure it works in a wide range of healthcare settings and populations. Integrating it with electronic health records (EHR) systems could make it even more useful by allowing it to be used in the routine risk assessment of diabetes.

6.0 FUTURE WORK

The stacking and soft voting approach can be utilized in subsequent studies involving different medical conditions, such as cardiovascular disease or cancer. To make the model better, we need to make it less complex to run, add techniques that make it easier to understand, and train it on more diverse datasets that include people from different ethnic groups. The model could also be used in real time in clinical settings to keep track of patients' risks and take action as soon as possible. The model can also be improved for use in places with few resources, which will make it more useful in healthcare settings all over the world.

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.

Ethical Statement: There are no ethical issues in this study.

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.

REFERENCES

1. Abdel-Jaber, H., Devassy, D., Al Salam, A., Hidaytallah, L., & EL-Amir, M. (2022). A Review of Deep Learning Algorithms and Their Applications in Healthcare. *Algorithms*, *15*(2), 71. https://doi.org/10.3390/a15020071
2. Abeßer, J. (2020). A Review of Deep Learning Based Methods for Acoustic Scene Classification. *Applied Sciences*, *10*(6), 2020. https://doi.org/10.3390/app10062020
3. Ahmed, N., Ahammed, R., Islam, Md. M., Uddin, Md. A., Akhter, A., Talukder, Md. A.-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. https://doi.org/10.1016/j.ijcce.2021.12.001
4. Alibrahim, H., & Ludwig, S. A. (2021). Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization. *2021 IEEE Congress on Evolutionary Computation (CEC)*. https://doi.org/10.1109/cec45853.2021.9504761
5. Bejani, M. M., & Ghatee, M. (2021). A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review*, *54*(8), 6391–6438. https://doi.org/10.1007/s10462-021-09975-1
6. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3). https://doi.org/10.1007/s10462-020-09896-5
7. Bertoncelli, C. M., Altamura, P., Vieira, E. R., Iyengar, S. S., Solla, F., & Bertoncelli, D. (2020). PredictMed: A logistic regression–based model to predict health conditions in cerebral palsy. *Health Informatics Journal*, *26*(3), 2105–2118. https://doi.org/10.1177/1460458219898568
8. Berumen, J., Orozco, L., Betancourt-Cravioto, M., Gallardo, H., Zulueta, M., Mendizabal, L., Simon, L., Benuto, R. E., Ramírez-Campos, E., Marin, M., Juárez, E., García-Ortiz, H., Martínez-Hernández, A., Venegas-Vega, C., Peralta-Romero, J., Cruz, M., & Tapia-Conyer, R. (2019). Influence of obesity, parental history of diabetes, and genes in type 2 diabetes: A case-control study. *Scientific Reports*, *9*(1), 2748. https://doi.org/10.1038/s41598-019-39145-x
9. Chen, Y., Song, L., Liu, Y., Yang, L., & Li, D. (2020). A Review of the Artificial Neural Network Models for Water Quality Prediction. *Applied Sciences*, *10*(17), 5776. https://doi.org/10.3390/app10175776
10. Cosenza, D. N., Korhonen, L., Maltamo, M., Packalen, P., Strunk, J. L., Næsset, E., Gobakken, T., Soares, P., & Tomé, M. (2020). Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of

growing stock. *Forestry: An International Journal of Forest Research*, *94*(2), 311–323. https://doi.org/10.1093/forestry/cpaa034

11. de Campos Souza, P. V., Lughofer, E., & Guimaraes, A. J. (2021). An interpretable evolving fuzzy neural network based on self-organized direction-aware data partitioning and fuzzy logic neurons. *Applied Soft Computing*, *112*, 107829. https://doi.org/10.1016/j.asoc.2021.107829
12. Dutta, A., Hasan, Md. K., Ahmad, M., Awal, Md. A., Islam, Md. A., Masud, M., & Meshref, H. (2022). Early Prediction of Diabetes Using an Ensemble of Machine Learning Models. *International Journal of Environmental Research and Public Health*, *19*(19), 12378. https://doi.org/10.3390/ijerph191912378
13. Giordani, P. (2021). Smartboost Learning for Tabular Data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3975543
14. Hackeling, G. (2017). *Mastering machine learning with scikit-learn : learning to implement and evaluate machine learning solutions with scikit-learn*. Packt Publishing Ltd.
15. Jaiswal, V., Negi, A., & Pal, T. (2021). A review on current advances in machine learning based diabetes prediction. *Primary Care Diabetes*, *15*(3), 435–443. https://doi.org/10.1016/j.pcd.2021.02.005
16. Jia, L., Wang, Z., Lv, S., & Xu, Z. (2022). PE_DIM: An Efficient Probabilistic Ensemble Classification Algorithm for Diabetes Handling Class Imbalance Missing Values. *IEEE Access*, *10*, 107459–107476. https://doi.org/10.1109/access.2022.3212067
17. Kamalraj, R., Neelakandan, S., Ranjith Kumar, M., Chandra Shekhar Rao, V., Anand, R., & Singh, H. (2021). Interpretable filter based convolutional neural network (IF-CNN) for glucose prediction and classification using PD-SS algorithm. *Measurement*, *183*, 109804. https://doi.org/10.1016/j.measurement.2021.109804
18. Karlo Abnoosian, Rahman Farnoosh, & Mohammad Hassan Behzadi. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. *BMC Bioinformatics*, *24*(1). https://doi.org/10.1186/s12859-023-05465-z
19. Khan, M. A., Iqbal, N., Imran, Jamil, H., & Kim, D.-H. (2023). An optimized ensemble prediction model using AutoML based on soft voting classifier for network intrusion detection. *Journal of Network and Computer Applications*, *212*, 103560. https://doi.org/10.1016/j.jnca.2022.103560
20. Kibria, H. B., Nahiduzzaman, M., Goni, Md. 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. https://doi.org/10.3390/s22197268
21. Kumar, S., & Vijay Kumar Jha. (2023a). Diabetes prediction model using machine learning techniques. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-023-16745-4
22. Lazzarini, R., Tianfield, H., & Charissis, V. (2023). A stacking ensemble of deep learning models for IoT intrusion detection. *Knowledge-Based Systems*, *279*, 110941. https://doi.org/10.1016/j.knosys.2023.110941
23. Liu, Z., Wei, P., Wei, Z., Yu, B., Jiang, J., Cao, W., Bian, J., & Chang, Y. (2021). *Handling Inter-class and Intra-class Imbalance in Class-imbalanced Learning*. ArXiv.org. https://arxiv.org/abs/2111.12791
24. Maniruzzaman, Md., Rahman, Md. J., Ahammed, B., & Abedin, Md. M. (2020).

Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems*, *8*(1). https://doi.org/10.1007/s13755-019-0095-z

25. Mitul Kumar Ahirwal, Londhe, N. D., & Kumar, A. (2022). *Artificial intelligence applications for health care*. Taylor And Francis.
26. Nurdin, A., Tane, M. M., Tumewu, R. W. T., Suryaningrum, K. M., & Saputri, H. A. (2023). Using Machine Learning for the Prediction of Diabetes with Emphasis on Blood Content. *Procedia Computer Science*, *227*, 990–1001. https://doi.org/10.1016/j.procs.2023.10.608
27. Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, *220*, 106773. https://doi.org/10.1016/j.cmpb.2022.106773
28. Plakias, S., & Boutalis, Y. S. (2020). Fault detection and identification of rolling element bearings with Attentive Dense CNN. *Neurocomputing*, *405*, 208–217. https://doi.org/10.1016/j.neucom.2020.04.143
29. Prasetyo Wibowo, H., Anshori, M., & Syauqi Haris, M. (2024). THE DISCRIMINANT ANALYSIS FUNCTION WAS IMPLEMENTED TO PREDICT THE PRESENCE OF DIABETES. *Journal of Enhanced Studies in Informatics and Computer Applications*, *1*(2), 47–55. https://doi.org/10.47794/jesica.v1i2.10
30. Rahmani, A. M., Yousefpoor, E., Yousefpoor, M. S., Mehmood, Z., Haider, A., Hosseinzadeh, M., & Ali Naqvi, R. (2021). Machine Learning (ML) in Medicine: Review, Applications, and Challenges. *Mathematics*, *9*(22), 2970. https://doi.org/10.3390/math9222970
31. Sansana, J., Joswiak, M. N., Castillo, I., Wang, Z., Rendall, R., Chiang, L. H., & Reis, M. S. (2021). Recent trends on hybrid modeling for Industry 4.0. *Computers & Chemical Engineering*, *151*, 107365–107365. https://doi.org/10.1016/j.compchemeng.2021.107365
32. Shamim, M., Hafsha, U., Amin, R., Yasmin, R., & Sabba Ruhi. (2023). Improving SVM Performance for Type II Diabetes Prediction with an Improved Non-Linear Kernel: Insights from the PIMA Dataset. *Computer Methods and Programs in Biomedicine Update*, *4*, 100118–100118. https://doi.org/10.1016/j.cmpbup.2023.100118
33. Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00349-y
34. Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: methodological failures and recommendations for the future. *Npj Digital Medicine*, *5*(1), 1–8. https://doi.org/10.1038/s41746-022-00592-y
35. Younas, F., Usman, M., & Yan, W. Q. (2022). A deep ensemble learning method for colorectal polyp classification with optimized network parameters. *Applied Intelligence*. https://doi.org/10.1007/s10489-022-03689-9.

Article History
Received: 29-10-2025
Accepted: 04-12-2025