The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 4, August 2025
https://doi.org/10.33022/ijcs.v14i4.4947

Enhancing Diabetes Prediction Accuracy Using Stacked Machine Learning and Deep Learning Models: A Public Health Approach Malaysia

Md Ziarul Islam*[1], Mohd Khairul Azmi Bin Hassan[2], Amir 'Aatieff Bin Amir Hussin[3], Md Salman Sha[4]
zia.ptd@gmail.com[1], mkazmi@iium.edu.my[2], amiraatieff@iium.edu.my[3], s.salman@live.iium.edu.my[4]
[1.2.3,4] Kulliyyah of Information and Communication Technology, International Islamic University Malaysia

Article Information

Received : 7 Jul 2025
Revised : 19 Jul 2025
Accepted : 1 Aug 2025

Keywords

Diabetes Prediction,
Artificial Intelligence,
Machine Learning,
Ensemble Learning,
Public Health

Abstract

Diabetes mellitus is a growing public health issue in Malaysia, affecting 7 million adults aged 18 and older. By 2025, 20.1% of Malaysians will have diabetes, with the International Diabetes Federation predicting 5 million by 2030. A study aims to improve diabetes prediction accuracy and reliability. The Indian PIMA Diabetes dataset was used to develop stacked machine learning and deep learning models, with 70% ML and 30% DL achieving optimal results. The weighted soft voting ensemble (70% ML, 30% DL) outperformed individual stacking models in terms of reliability and balanced performance, improving diabetes classification with 75.65% accuracy, 67.89% precision, and 81.41% ROC-AUC. The ensemble method, optimized for medical diagnosis tasks, showed improved accuracy, robustness, and generalization. However, ethical considerations, data privacy, and algorithmic biases are crucial for maximizing AI's potential in diabetes care, highlighting the need for scalable solutions.

A. Introduction

1. Diabetes Prevalence and Figures

Diabetes mellitus is on the rise in Malaysia, and it is becoming a major public health problem that is getting a lot of attention around the world. According to the International Diabetes Federation (IDF), Malaysia is one of the 38 countries and territories that make up the IDF Western Pacific region. It is believed that 206 million people in the Western Pacific Region and 537 million people around the world will have diabetes by 2045. There are 4,431,500 adults with diabetes in Malaysia as of 2021. The total number of adults in the country is 22,130,900 [1, 2]. The NHMS 2023 is a survey that looks at a group of people at a single point in time. It shows that 15.6% of adults have diabetes and 9.7% of adults know they have diabetes. It was done using a two-stage stratified random sampling method and complex weighted sample analysis. 5.9% of people who didn't have diabetes had high blood sugar [3]. Twenty percent of adults have diabetes. With a 5% incidence rate, the Malaysian population of about 15 million people means that almost 3.9 million people have diabetes. This number is expected to rise to about 5,024,900 by 2030 [4]. The results show that diabetes is a big public health problem in Malaysia. Type 2 diabetes (T2D) is now affecting 2.8 million adults over the age of 30, which is 20.8% of the population.

Primary and tertiary healthcare providers working in different places are responsible for managing diabetes. Even with these efforts, Malaysia's healthcare system has problems like not having enough resources, having too many patients for each doctor, and having to do diabetes prediction and treatment by hand [6]. Experts say that the high rates of diabetes and prediabetes mean that the country needs full diabetes control programs [7]. Using artificial intelligence (AI) and machine learning (ML) can help make diabetes management better by improving diagnostics, predictive modeling, and personalized care [8, 9].

Adding artificial intelligence (AI) and machine learning (ML) to healthcare could help solve many of the problems it faces. AI and ML technologies have a lot of potential to change diagnostics, improve patient outcomes, make healthcare systems more efficient, and make decision-making processes better [10]. AI and ML can also make administrative tasks easier, improve hospital operations, make better use of resources, and get patients more involved by using virtual assistants and chatbots [11,12,13]. But to fully benefit from AI and ML in healthcare, we need to deal with problems like data privacy, algorithmic biases, and ethical issues.

2. Problem Statement

Diabetes mellitus is a significant and growing public health issue in Malaysia. The number of people with diabetes in the country has been steadily rising, and estimates say that by 2025, more than 7 million adults could be affected. Lifestyle changes, an aging population, urbanization, and eating habits are all causing this rise [14]. The increasing number of cases not only raises the risk of complications like heart disease, kidney failure, and neuropathy, but it also puts a lot of stress on Malaysia's healthcare system. Even though the government is trying to raise awareness and get people screened early, many cases still go undiagnosed or are not handled well because of limited access to healthcare, not enough medical resources, and delays in making clinical decisions [15].

Artificial intelligence (AI) has the potential to improve diabetes care by allowing for early prediction, personalized treatment planning, and decision support based on data. But there are a number of problems that make it hard to use AI successfully in Malaysia's healthcare system [16]. These include the lack of datasets that are specific to the area, the possibility of biases in imported models, worries about data privacy, and the fact that AI tools don't work well with existing clinical workflows. Most AI models that predict diabetes use global datasets, like the Pima Indian Diabetes dataset and Kaggle diabetes dataset. These datasets may not be representative of Malaysia's unique population. Also, there aren't many ethical and regulatory issues that need to be worked out before AI can be used in healthcare.

To improve diabetes care in Malaysia, it is very important to create and test AI-driven models that are specific to the country [5]. To get the most out of AI in terms of improving early diagnosis, treatment outcomes, and healthcare efficiency across the country, these problems need to be solved [17].

3. Aim, Contribution, and Paper Organization

The study looked at the five best deep learning and regular machine learning algorithms using the India Pima diabetes dataset. We combined the five best machine learning models to make a stacking model for machine learning [18]. We also made a stacked deep learning model by combining the five best deep learning models [19]. The stacking ML and DL model did a better job than any of the other models by themselves. Finally, we used a weighted soft voting classifier on the two stacked models, and it worked very well.

So, we are using the India Pima diabetes dataset to build a new framework for Malaysia's healthcare system. The expected results include better digital health tools for Malaysia's healthcare system, lower healthcare costs in the future, and the creation of a healthcare system in Malaysia that can last over time [20].

The main contributions of this article to research are as follows:

- The study presents a new hybrid model that uses weighted soft voting to combine machine learning (ML) and deep learning (DL) stacking methods. This method combines the best parts of ML (like stability and interpretability) and DL (like feature extraction and sensitivity) to create a more balanced and robust model for predicting diabetes diagnosis.
- The study finds that the best overall performance is achieved by giving 70% weight to ML models and 30% weight to DL models in the soft voting mechanism. This results in 75.65% accuracy, 67.89% precision, and a ROC-AUC of 81.41%, which is better than individual and purely stacked models.
- The article thoroughly evaluates model performance using multiple classification metrics: accuracy, precision, recall, F1-score, ROC-AUC, and Cohen's Kappa. This evaluation with multiple metrics gives a complete picture of the model's strengths in both correct classification and class balance, which is crucial in medical situations like predicting diabetes.
- The study uses logistic regression as a meta-learner in both ML and DL stacking architectures. The method makes it possible to combine the outputs of base models in a way that improves generalization and lowers

prediction variance, which is especially useful when working with health datasets that have complex, nonlinear patterns.

- The proposed framework is a scalable and understandable AI-based diagnostic tool made for real-world healthcare systems. It focuses on simple models (rather than more complex meta-learners) and data-driven decision support. It also talks about ethical issues like data privacy and algorithmic bias, which are crucial for using AI responsibly in public health.

The remainder of the paper follows this structure. Section B presented related on the Indian Pima dataset employed and its contents with weighted soft voting, section C presented outlines the study's methodology, section D presented machine learning model stacking application, section E deep learning model stacking application, section F weighted soft voting application on ML and DL stacking application with algorithm, section G Critical Analysis on the Weighted Soft Voting Experimental Results, finally, section H provides concluding remarks on the paper, an overview of the entire article, and some scopes for further research.

B. Related Works

A significant amount of research has looked into how to use artificial intelligence (AI) and machine learning (ML) to predict and control diabetes. Early models frequently utilized conventional statistical techniques, including logistic regression and decision trees, to pinpoint individuals at elevated risk. Recent studies have focused on using advanced AI methods like deep learning and ensemble methods to make predictions more accurate and relevant [21]. The Pima Indian Diabetes dataset is a well-known dataset for these kinds of models. It has been used to test classification algorithms like support vector machines (SVM), random forests, deep learning, and gradient boosting [22].

However, the best results have been found when testing classification algorithms such as deep learning. These models have shown good results in global settings, with high accuracy, sensitivity, and specificity [23]. For instance, research has shown that hybrid and ensemble methods can make predictions with more than 80% accuracy, which shows how useful they could be in clinical settings. However, transferring these models to Malaysia is still very hard because of differences in genetics, lifestyle, and access to healthcare among the people there. There are not many studies that look at the specific needs of Malaysian patients or use local clinical data, which makes it hard to use existing models in the country [24.25].

Researchers have also been paying more and more attention to the ethical issues that AI in healthcare raises, such as data privacy, fairness, and algorithmic bias. These issues are especially important in Malaysia, where AI is still being added to public healthcare [26]. Consequently, there is a distinct necessity for Malaysia-centric research that assesses the efficacy of AI models within local contexts, while addressing the overarching ethical and systemic challenges associated with AI implementation in healthcare.

C. Methodology

1. Data Collection and Preprocessing

One of the most well-known datasets in machine learning and deep learning, particularly for binary classification used to predict diabetes, is the Pima Indians Diabetes dataset [27]. The National Institute of Diabetes and Digestive and Kidney Diseases provided the study's data, which included 768 records of female Pima Indian individuals who were 21 years of age or older [28]. Eight characteristics make up the dataset: age, diabetes pedigree function (the ancestry of diabetes), triceps skin fold thickness, two-hour plasma insulin, pregnancy, diastolic blood pressure, and BMI. Because the dependent variable is dichotomous, an answer of "1" indicates that the person has been diagnosed with diabetes, while a response of "0" suggests otherwise. This dataset is especially helpful for evaluating and contrasting the precision with which various machine learning algorithms predict the onset of diabetes [28]. As a result, it can be used as a tool to create models that support early disease identification were preprocessed to handle missing values and outliers through techniques like imputation and normalization. The dataset was split into training and testing sets to assess model robustness and accuracy. The Pima Indian dataset is taken from the URL https://data.world/data-society/pima-indians-diabetes-database. It is split into two sets: 80% for training and 20% for validation [30]. The validation set is 20% of the input dataset that was chosen to help choose the hyperparameters. The Indian PIMA diabetic feature table is presented below.

Table 1. Indian PIMA diabetic feature

Sr. no.	Selected Attributes from PIMA Indian dataset	Description of selected attributes	Range
1	Pregnancy	Number of times a participant is pregnant	0–172
2	Glucose	Plasma glucose concentration a 2 h in an oral glucose tolerance test	0–199
3	Diastolic Blood pressure	It consists of Diastolic blood pressure (when blood exerts into arteries between heart) (mm Hg)	0–122
4	Skin Thickness	Triceps skinfold thickness (mm). It concluded by the collagen content	0–99
5	Serum Insulin	2-Hour serum insulin (mu U/ml)	0–846
6	BMI	Body mass index (weight in kg/(height in m)^2)	0–67.1
7	Diabetes pedigree Function	An appealing attributed used in diabetes prognosis	0.078–2.42
8	Age	Age of participants	21–81
9	Outcome	Diabetes class variable, Yes represent the patient is diabetic and no represent patient is not diabetic	Yes/No

2. Heatmap of persons correlation coefficient for all diabetes feature

The correlation heatmap shows how the features in the diabetes dataset are linearly related to each other. The color scale shows how strong and in what direction correlations are: red for strong positive correlations (values close to 1), blue for strong negative correlations (values close to -1), and neutral colors for

weak correlations (values close to 0). When two features are highly positively correlated, it means that as one goes up, the other does too. When two features are highly negatively correlated, it means that as one goes down, the other goes up. Features that are not very related (close to 0) are more independent. This heatmap can help you find duplicate features and relationships, which can help you make decisions about which features to use and how to design your model to avoid multicollinearity and make predictions more accurately [31].

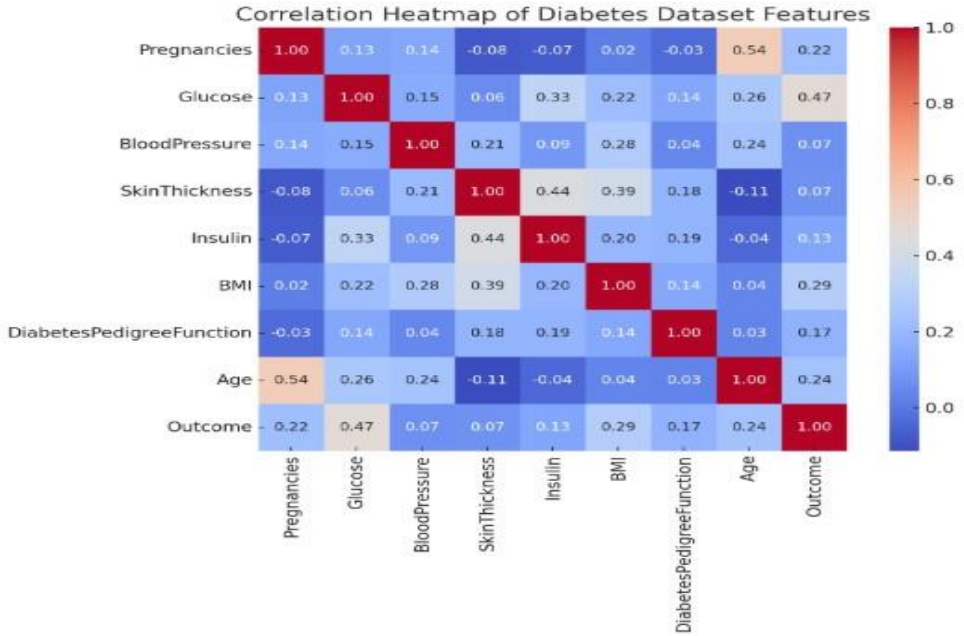

Figure 1. Heatmap of persons correlation diabetes dataset feature

3. An Overview of the Proposed Research Architecture

The first step in the proposed framework for improving diabetes prediction is to load and prepare the Indian Pima dataset. This means dealing with missing values, making sure features are the same, and dividing the data into training and testing sets. We then use the dataset to train five machines learning models, including Random Forest, XG-Boost, and Logistic Regression. Then we design and train five deep learning model that has GANs, GRU, LSTM, MLP, and RNN at the same time [32]. Then, all of the models make predictions about the probabilities on the test set. A weighted soft voting ensemble combines these outputs. Each model's prediction is multiplied by a set weight, and the final prediction is based on the weighted average.

The Indian Pima Diabetes Dataset exhibited accuracy, precision, F1 score, ROC-AUC, and Cohen's Kappa, utilizing five conventional machine learning and five deep learning, evaluated through a 5-fold cross-validation method [33]. Nevertheless, by leveraging clinical, lifestyle, and demographic data, Malaysia's healthcare sector will develop and assess an improved AI model for the prediction of diabetes mellitus with greater accuracy.

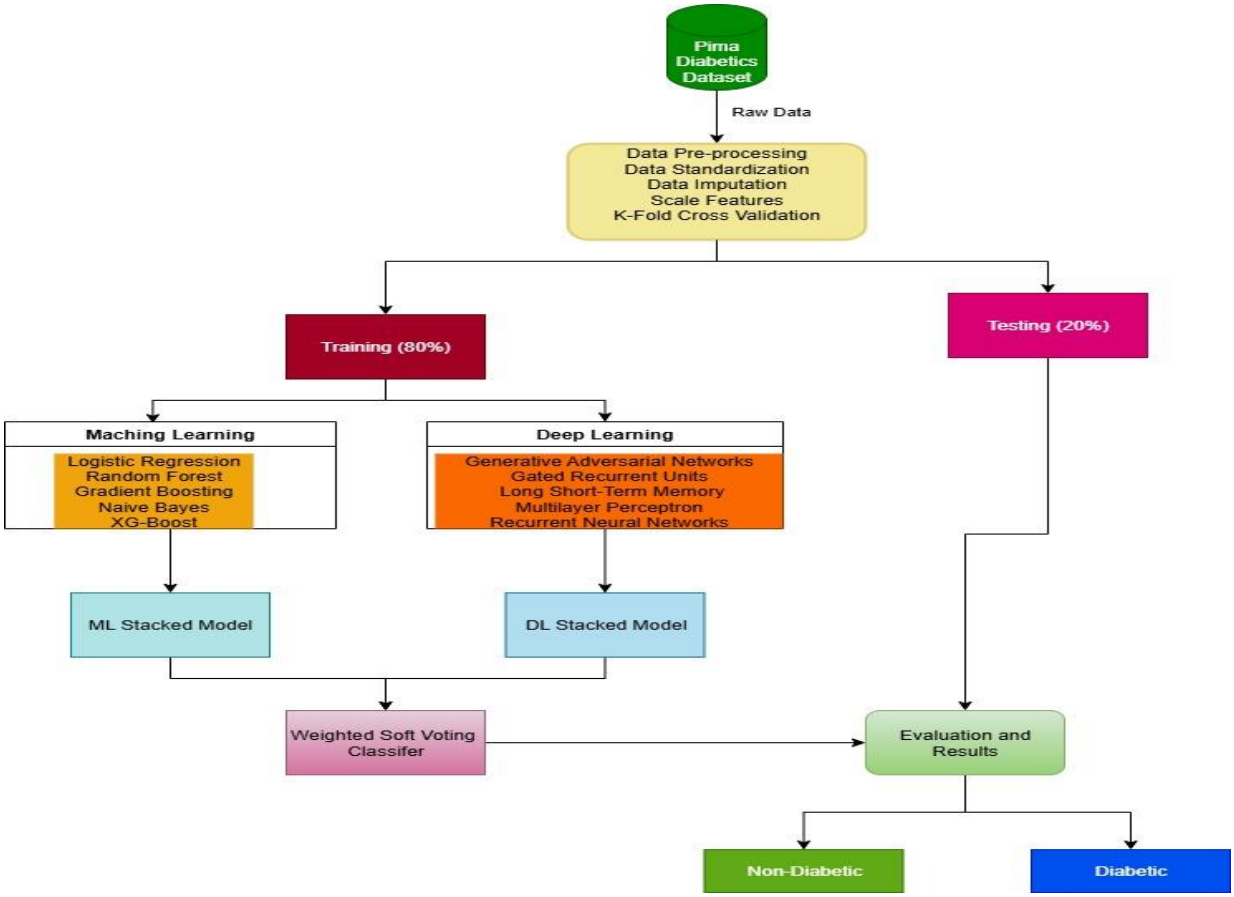

Figure 2. Proposed Research Architecture

4. Performance Parameters

We applied machine learning and deep learning algorithms to the Pima Indian dataset. We performed some preprocessing on the dataset before applying the algorithm. After applying the ML and DL algorithms to the preprocessed dataset, the researchers developed a weighted soft voting ensemble method using a voting classifier [34]. The researchers selected the weights for the ensemble method using a soft voting approach. For the classification of the diabetes disease, six quality parameters have been calculated.
Here are the performance parameters:

4.1 Accuracy: To find the accuracy, divide the total number of predictions below by the number of correct predictions.

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

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

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

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

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

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

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

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

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

4.6 Cohen Kappa: The Cohen's kappa (K) is a number that shows how well a prediction matches the real class. The Kappa statistic says that the best method is one with a value close to 1. In this case, p0 stands for the relative observed agreement, and p0 stands for the hypothetical chance agreement. Cohen's Kappa takes into account chance agreement, which makes it a better way to measure how well a model works than just accuracy.

$$K = \frac{\text{Po} - \text{Pe}}{1 - \text{Pe}}$$

D. ML Model Stacking Application in Pima Database

An ensemble learning method called "model stacking" combines several models to increase prediction accuracy overall. Predictions are made using many base models, which are then fed into a meta-model (also known as a meta-learner) to produce the final forecast. Combining several models might help you make use of each one's advantages while mitigating its shortcomings because different models may identify distinct patterns in the data [35].

1. Machine Learning Model Stacking

Model stacking, often called stacked generalization, is an ensemble learning approach that combines many models to enhance prediction performance in machine learning. This method uses the same dataset to train multiple models (base learners), whose predictions are then fed into a final model (meta-learner) to get the final prediction. By combining the outputs of several models, the goal is to take advantage of their advantages and lessen their disadvantages [36, 37]. Usually, there are two primary processes in the stacking process:

2. Machine Learning Meta Level

In this study, a meta-modeling technique was utilized by training a meta-model with the predictions from five base models as input features, allowing it to

effectively amalgamate their outputs for enhanced final predictions. The five base models used on the Indian Pima Dataset were: Logistic Regression, which is easy to understand and use; Random Forest, which uses a group of decision trees to model complicated relationships and avoid overfitting; Gradient Boosting, which is a sequential method that improves performance by fixing mistakes from the past; Naive Bayes, which is a fast probabilistic model that assumes features are independent; and XGBoost, which is a powerful and efficient implementation of gradient boosting that is known for being very accurate in competitive machine learning tasks [38].

3. Machine Learning Meta-Classifier

Logistic Regression was used as the meta-classifier. This model took the predictions made by the base models as inputs and learned how to best combine these predictions to make the final decision. The meta-classifier essentially learns how to weigh the contributions of each base model to optimize the final prediction [39].

4. Machine Learning Model Stacking Methodology Diagram

Machine Learning Model Stacking, or simply "stacking," is an ensemble learning technique that combines the predictions of multiple models (often called "base models" or "level 0 models") to produce a more accurate and robust final prediction. The key idea is to leverage the strengths of various models to minimize their weaknesses, leading to improved overall performance. The methodology used in this analysis can be represented in a flowchart that includes the following steps [40].

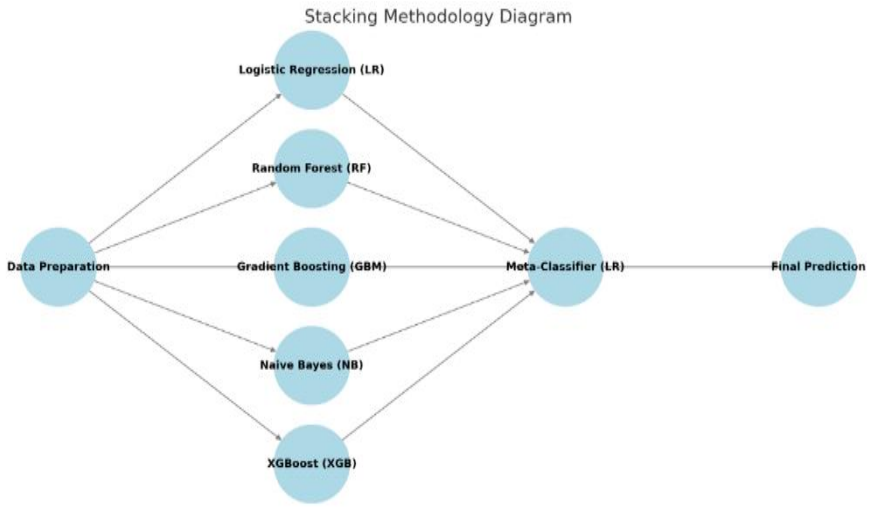

Figure 3. Machine Learning Model Stacking Methodology Diagram

The first step is to get the data ready. This means splitting the dataset into features (X) and the target variable (y). Then, several machine learning base models are used, such as Logistic Regression, Random Forest, Gradient Boosting, Naive Bayes, and XGBoost. Each model has its own strengths when it comes to

working with the data. Each of these base models is trained on its own on the training dataset. Then, a meta-classifier, specifically a Logistic Regression model, uses their predictions as input and learns how to best combine them. Finally, the overall ensemble model is tested using different metrics, such as accuracy, sensitivity, precision, F1 score, ROC-AUC, and Cohen's Kappa, to make sure the results are strong and reliable [41].

5. Initial Performance Metric of Machine Learning Model Stacking

Initial performance metrics on the Indian Pima dataset demonstrate that Machine Learning Model Stacking can significantly improve prediction accuracy for diabetes diagnosis compared to individual models alone. By combining various base models, stacking effectively reduces prediction errors and enhances robustness [42, 43]. After running the stack model on the Pima database, these are the initial findings.

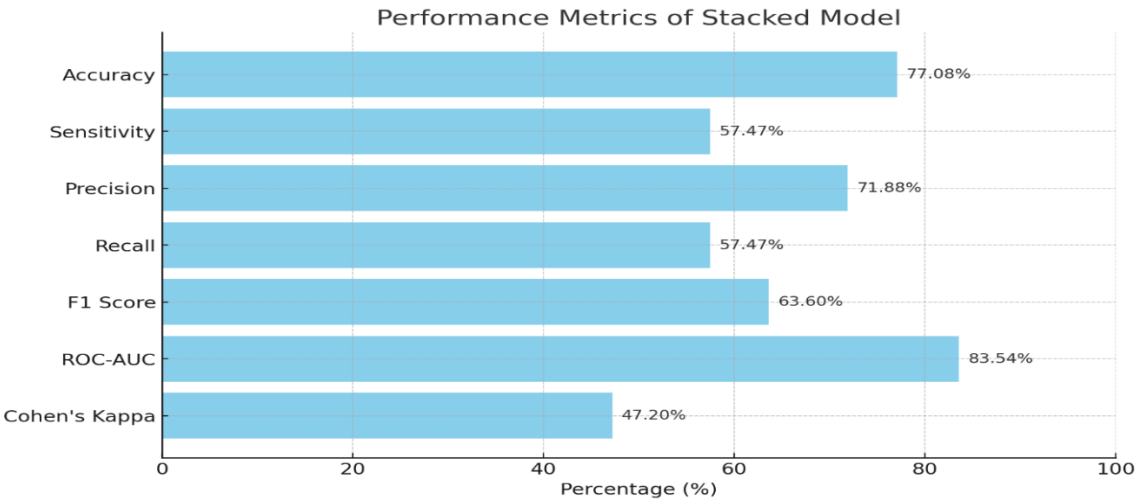

Figure 4. Pima Dataset Performance Metrics of Stack Model

- Accuracy (77.08%): The stacked model correctly predicted the outcome (diabetes vs. no diabetes) 77.08% of the time, indicating strong overall correctness.
- Sensitivity (Recall) (57.47%): Sensitivity measures the model's ability to correctly identify positive cases (i.e., individuals with diabetes). A sensitivity of 57.47% means that the model correctly identified 57.47% of all true positive cases. This suggests that while the model is reasonably accurate overall, there is still room for improvement in correctly identifying all cases of diabetes.
- Precision (71.88%): Precision indicates the proportion of positive identifications that were correct. With a precision of 71.88%, the model made accurate predictions when it identified a case as diabetic 71.88% of the time.
- F1 Score (63.60%): The F1 Score, which is the harmonic mean of precision and recall, is 63.60%. This score provides a balance between precision and

recall, indicating that the model performs reasonably well in both identifying and correctly predicting positive cases.

- ROC-AUC (83.54%): The ROC-AUC score of 83.54% measures the model's ability to distinguish between positive and negative cases. A higher ROC-AUC score suggests that the model is very capable of distinguishing between diabetic and non-diabetic individuals.
- Cohen's Kappa (47.20%): Cohen's Kappa accounts for the possibility of agreement occurring by chance. A Kappa of 47.20% suggests moderate agreement beyond chance.

6. Significance of Machine Learning Model Stacking in the Pima Database

The Pima Indian Diabetes dataset is a great example of a machine learning model called Stacking. Stacking improves predictive performance by combining different models, such as Logistic Regression, Random Forest, Gradient Boosting, Naive Bayes, and XGBoost. This method finds different data patterns and makes generalization better by using the best parts of each model. Stacking makes models work better by combining their strengths, which makes them better at predicting and generalizing than any one model alone. Logistic Regression is used for linear relationships, Random Forest and Gradient Boosting are used for more complicated interactions, Naive Bayes is used for probabilistic aspects, and XGBoost is a powerful gradient boosting method that is known for its accuracy and speed [44]. Adding models like XGBoost makes things harder because it can work with big datasets that have a lot of different features.

7. Perceived Improvement of the Stacked Machine Learning Model

The comparison between the stacked model and individual models shows that the stacked model works better because it uses more than one algorithm to make predictions that are more accurate and reliable than any one model alone [45]. The overall model is greatly improved by the addition of XGBoost, which has a performance of 77.08% and advanced boosting features. This ensemble method finds a good balance between precision and recall, which are two metrics that don't always agree with each other in single models. In other words, one model might be better at one thing but worse at the other. So, the stacked model is better at making predictions because it is completer and more reliable.

E. DL Model Stacking Application in Pima Database

Deep Learning Model Stacking is an advanced ensemble method in which several deep learning models are trained on the same dataset independently, and their predictions are combined to make a final predictive model. In this method, base learners like CNNs, RNNs, or transformers make first guesses, which are then used as input features for a meta-model that combines their outputs to make the final classification [46]. This study uses stacked deep learning on the Pima Indians Diabetes dataset by picking the five best models—GANs, GRU, LSTM, MLP, and RNN—based on performance metrics like accuracy, ROC-AUC, and F1 score [47]. A logistic regression model functions as the meta-learner, seeking to improve

predictive accuracy by adeptly amalgamating the advantages of each base model via a meta-learning approach.

1. Deep Learning Model Stacking Methodology Diagram

Stacking is an ensemble learning technique where multiple models (referred to as base learners) are trained, and their predictions are combined using a meta-learner [48]. The meta-learner's role is to learn from the outputs of the base learners and make the final prediction.

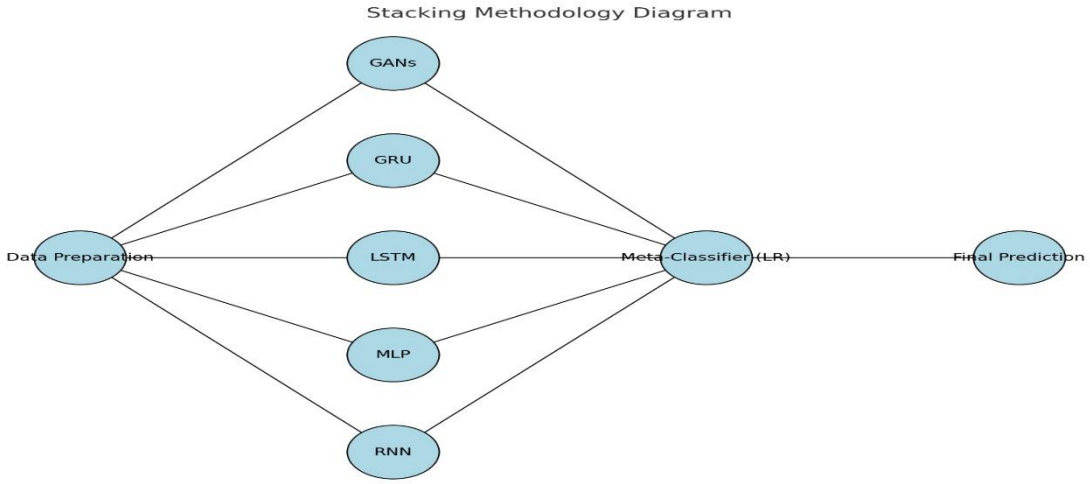

Figure 5. Deep Learning Model Stacking Methodology Diagram

There are several important steps in the Deep Learning Model Stacking method that help make predictions more accurate. First, each base model is trained on the training data by itself, so it can learn to make predictions on its own. After they have been trained, these models can make predictions for both the training and test datasets. Then, these predictions are put together to make new features from the output of each model. A meta-learner, which is usually a logistic regression model, is trained on these stacked predictions to figure out the best way to put them together into a final prediction. Finally, the effectiveness of the stacked model is tested on the test set using a number of performance metrics, such as accuracy, sensitivity, precision, recall, F1 score, ROC-AUC, and Cohen's Kappa [49].

2. Deep Learning Model Layer Architectures

Deep learning models like GAN, GRU, LSTM, MLP, and RNN each have their own strengths that help predict diabetes. GAN (Generator) makes fake data to make training better, which makes the model more robust. LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) are both good at working with sequential data, which is important for predicting how diabetes will get worse over time. LSTM is better at keeping memories for a long time, while GRU is easier on computers. MLP (Multilayer Perceptron) is a great way to predict things with non-sequential, tabular data because it uses fully connected layers. RNN (Recurrent Neural Network) works with sequential data, but it is more likely to lose gradients than GRU and LSTM [50].

Table 2. Deep Learning Model Layer Architectures

Model	Layer	Configuration
GAN (Generator)	Dense (Input)	64 units, ReLU, input_dim=100
	Batch Normalization	momentum=0.9
	Dense	32 units, ReLU
	Dense (Output)	8 units, Sigmoid (generates synthetic feature data)
GRU	GRU (1)	64 units, tanh, return_sequences=True, input_shape=(1, 8)
	Dropout	rate=0.2
	GRU (2)	32 units, tanh, return_sequences=False
	Dense	16 units, ReLU
	Dense (Output)	1 unit, Sigmoid
LSTM	LSTM (1)	64 units, tanh, return_sequences=True, input_shape=(1, 8)
	Dropout	rate=0.2
	LSTM (2)	32 units, tanh, return_sequences=False
	Dense	16 units, ReLU
	Dense (Output)	1 unit, Sigmoid
MLP	Dense (1)	128 units, ReLU, input_shape=(8,)
	Dropout	rate=0.3
	Dense (2)	64 units, ReLU
	Dense (3)	32 units, ReLU
	Dense (Output)	1 unit, Sigmoid
RNN	SimpleRNN (1)	64 units, tanh, return_sequences=True, input_shape=(1, 8)
	Dropout	rate=0.2
	SimpleRNN (2)	32 units, tanh, return_sequences=False
	Dense	16 units, ReLU
	Dense (Output)	1 unit, Sigmoid

3. Initial Performance Metric of Deep Learning Model Stacking

The deep learning stacked model on the Pima Indian Diabetes dataset achieved a 73.16% accuracy, demonstrating competitive performance. Sensitivity and precision were well-balanced, resulting in an F1 score of 60.26%. The model's ROC-AUC of 76.57% indicates strong discriminative power, and a Cohen's Kappa score of 40.02% reflects moderate agreement with actual classifications [51].

- Accuracy: The stacked model achieved an accuracy of 73.16%, which is competitive with the individual models.
- Sensitivity and Precision: The sensitivity (recall) and precision of the stacked model are fairly balanced, indicating that the model has a good trade-off between capturing positive cases and minimizing false positives.

- F1 Score: The F1 score of 60.26% reflects the model's balance between precision and recall, making it a reliable metric for assessing overall performance.
- ROC-AUC: With a ROC-AUC of 76.57%, the model demonstrates good discrimination ability, meaning it is effective at distinguishing between positive and negative classes.
- Cohen's Kappa: The Cohen's Kappa score of 40.02% suggests moderate agreement between the predicted and actual classifications, accounting for chance agreement.

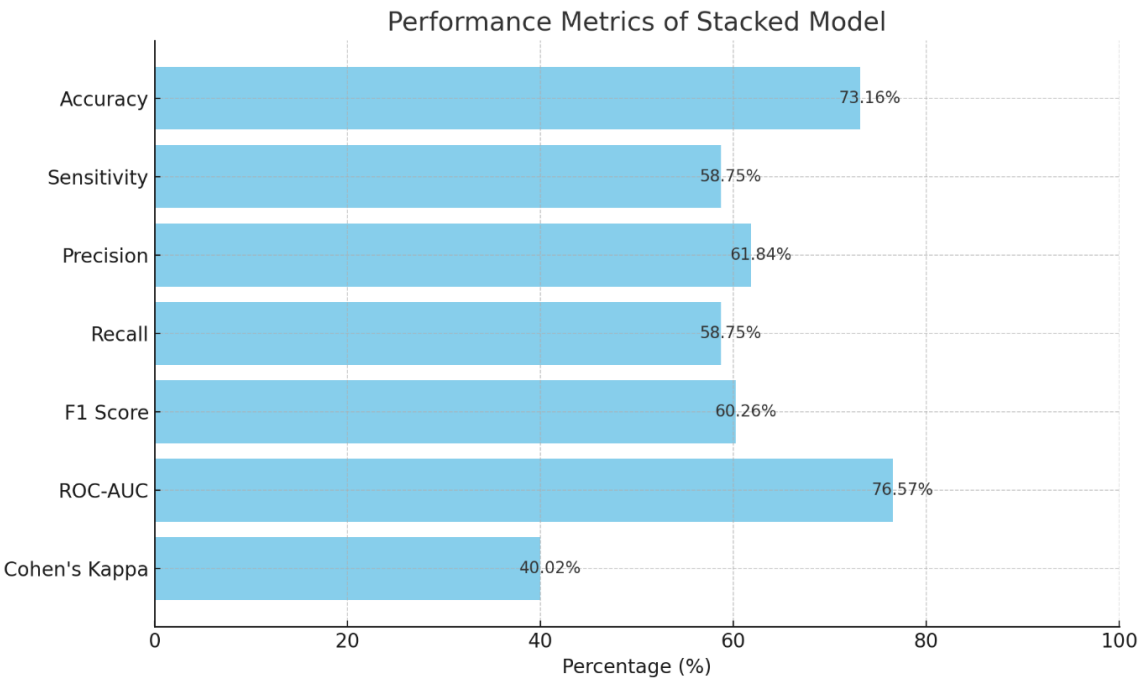

Figure 6. Initial Performance Metric of Deep Learning Model Stacking

4. Perceived Improvement of the Stacked Deep Learning Model

The stacked model built from the top five performing models has demonstrated superior performance compared to individual models across key metrics. The use of a logistic regression meta-learner to combine the predictions of GANs, GRU, LSTM, MLP, and RNN resulted in a model that is better equipped to generalize to new data, providing more accurate and balanced predictions [52]. This analysis underscores the value of ensemble learning techniques, particularly stacking, in enhancing predictive performance in complex datasets like the Pima Indians Diabetes dataset [53]. The results suggest that the stacked model is a more effective approach than relying on a singular model, offering a promising method for future predictive modeling tasks.

F. ML and DL Discussion on Pima Dataset Analysis

In this comprehensive study, we explored various machine learning (ML) and deep learning (DL) models to predict diabetes outcomes using the Pima Indian Diabetes dataset. Our goal was to maximize predictive accuracy by employing different strategies, including individual models, stacking ensembles, and soft voting techniques [54].

1. Stacking Model Approach

The stacking model for diabetes prediction uses Logistic Regression, Random Forest, XGBoost, and SVM as base models to capture diverse data patterns. A deep learning serves as the meta-model, combining these predictions to improve accuracy [55]. This approach enhances prediction accuracy and balances sensitivity and precision effectively. Given the strong performance of individual models, we explored stacking—aiming to combine their strengths in a single ensemble [56]. We created two primary stacking models.

2. Machine Learning and deep learning Stacking

We got this result by combining different machine learning models, such as Logistic Regression, Random Forest, Gradient Boosting, Naive Bayes, and XGBoost, to find different patterns in the data [57]. We used a stacked ensemble approach that combined deep learning architectures like GANs, GRU, LSTM, MLP, and RNN to make performance even better [58]. Logistic Regression was the meta-learner at the end, and it learned how to best combine the outputs from these models to make a strong and accurate prediction.

Table 3. ***Machine Learning and deep learning Stacking***

Machine Learning Stacking Results		**Deep Learning Stacking Results**	
Metric	Value (%)	Metric	Value (%)
Accuracy	77.08	Accuracy	73.16
Recall	57.47	Precision	61.84
Precision	71.88	Recall	58.75
F1 Score	63.60	F1 Score	60.26
ROC-AUC	83.54	ROC-AUC	76.57
Cohen's Kappa	47.20	Cohen's Kappa	40.02

3. Weighted Soft Voting Approaches

Soft voting is an ensemble learning technique used in machine learning to combine the predictions from multiple models to make a final prediction. Unlike hard voting, where the majority class prediction is chosen, soft voting aggregates the probabilistic outputs of each model and makes the final prediction based on the

averaged probabilities. Suppose you have three base models predicting whether a patient has diabetes [59]. Each model outputs probabilities for each class (e.g., 0 and 1). Soft voting calculates the average probability for each class and selects the class with the highest average probability as the final prediction. Following the stacking attempts, we explored soft voting, where the predicted probabilities from multiple models are averaged to make the final prediction. Soft Voting between 5 Stacked Machine Learning Models (Logistic Regression, Random Forest, Gradient Boosting, Naive Bayes, XGBoost) and 5 Stacked Deep Learning Models (Generative Adversarial Networks (GANs), Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Multilayer Perceptron (MLP), Recurrent Neural Network (RNN)). Here, we tried soft voting between the two major stacked models—one composed of ML algorithms and the other of DL algorithms.

Table 4. Weighted Soft Voting Results

Metric	ML (100%)	DL (100%)	80% ML / 20% DL	70% ML / 30% DL	60% ML / 40% DL	50% ML / 50% DL
Accuracy	77.08	73.16	76.32	75.65	74.99	75.12
Precision	71.88	61.84	69.27	67.89	66.50	66.86
Recall	57.47	58.75	57.73	57.97	58.22	58.11
F1 Score	63.60	60.26	62.76	62.01	61.25	61.93
ROC-AUC	83.54	76.57	82.15	81.41	80.68	80.06
Cohen's Kappa	47.20	40.02	45.76	44.76	43.76	43.61

4. Pseudocode for Diabetes Prediction Using ML and DL Stacking Models with Weighted Soft Voting

The pseudocode describes a method for predicting diabetes by using a mix of Machine Learning (ML) and Deep Learning (DL) models with weighted soft voting. The first step in the workflow is to load and preprocess the diabetes dataset by dealing with missing values and making the features the same. After that, the data is split into two groups: one for training and one for testing. Logistic Regression, Random Forest, Gradient Boosting, Naive Bayes, and XGBoost are just a few of the ML models that are trained on the data to make probability predictions [60]. Several deep learning models, including MLP, LSTM, GRU, RNN, and GANs, are trained at the same time and make their predictions [61]. Then, the ML and DL predictions are combined using weighted soft voting, with ML models making up 70% of the final decision and DL models making up 30%. The final predictions are given a threshold to determine whether they are positive (1) or negative (0). Then, evaluation metrics like accuracy, precision, recall, F1 score, ROC-AUC, and Cohen's Kappa are used to measure how well the model works.

Weighted Soft Voting Algorithm:

1. Load the diabetes dataset
   - Input: CSV file (features + Outcome column)
2. Preprocess the data
   a. Handle missing or zero values (replace with column median)
   b. Split into features (X) and target (y)

```
  c. Standardize features using StandardScaler
3. Split the data into training and testing sets (e.g., 80% training, 20% testing)
4. Train Machine Learning Models
  - Initialize models:
    - Logistic Regression
    - Random Forest
    - Gradient Boosting
    - Naive Bayes
    - XGBoost
  - For each ML model:
    - Train on training data
    - Predict probability scores on test data
  - Stack predictions → ML_Preds (average probabilities)
5. Train Deep Learning Models
  a. Reshape input data if required (e.g., for RNN, LSTM, GRU)
  b. Define architectures:
    - MLP: Dense layers with ReLU and sigmoid
    - LSTM: LSTM + Dense
    - GRU: GRU + Dense
    - RNN: SimpleRNN + Dense
    - GANs: Train generator-discriminator for synthetic data augmentation or prediction
  c. For each DL model:
    - Compile with binary cross-entropy, use Adam optimizer
    - Train on training data
    - Predict probability scores on test data
  - Stack predictions → DL_Preds (average probabilities)
6. Combine ML and DL predictions using Weighted Soft Voting
  - Set alpha = weight for ML predictions (e.g., 0.7)
  - Set beta = 1 - alpha (DL weight, e.g., 0.3)
  - Final_Preds = (alpha * average (ML_Preds)) + (beta * average (DL_Preds))
7. Threshold final predictions
  - If Final_Preds > 0.5 → Predict 1
  - Else → Predict 0
8. Evaluate the final predictions using:
  - Accuracy
  - Precision
  - Recall
  - F1 Score
  - ROC-AUC
  - Cohen's Kappa
9. Print or store the evaluation metrics
END
```

G. Critical Analysis on the Weighted Soft Voting Experimental Results

When you use weighted soft voting, you can get better results than with individual ML or DL stacking models because it uses the best parts of each. The machine learning ensemble (ML Stacking) is better at precision, accuracy, ROC-AUC, and Cohen's Kappa, which are all metrics that show how reliable and confident a classification is [62]. On the other hand, the deep learning stack (DL Stacking) gives stronger recall, which is important for finding positive cases that are sensitive, especially in medical diagnosis like predicting diabetes [63].

The 70% ML and 30% DL weight setting is the best for all metrics. It raises accuracy to 75.65%, which is only a little lower than the ML-only peak (77.08%) and a lot higher than the DL-only peak (73.16%). More importantly, it keeps a high

precision of 67.89%, which helps reduce false positives, and its recall stays competitive at 57.97%, which is only slightly lower than DL's 58.75%. When both precision and recall are important, this balance gives an F1 Score of 62.01%, which is a very important number. Also, ROC-AUC is 81.41%, which shows that the ensemble is very sure about separating the classes, which is important for assessing medical risk. Finally, Cohen's Kappa of 44.76% shows that the labels agree more than they do with either model on its own, which supports the ensemble's consistency.

From a critical point of view, the ensemble works because ML and DL work together: ML gives decisions stability, while DL gives decisions sensitivity [63, 64]. Equal weighting (50/50) leads to diminished advantages—superior to DL alone, yet inferior to strategically biased combinations. Lastly, weighted soft voting, especially with 70% ML and 30% DL, improves the ability to make predictions by balancing discrimination power, reliability, and sensitivity. It works especially well in high-stakes fields like healthcare, where both precision and recall are very important. Using meta-model optimization or probabilistic calibration of outputs could lead to even more improvements.

Finally, the results show that hybrid models, particularly those with a majority ML weighting (e.g., 80% ML / 20% DL), provide a convincing compromise, even though pure ML models outperform them in the majority of metrics. These setups take advantage of DL's sophisticated feature extraction capabilities while maintaining excellent performance. The 80/20 ensemble, which strikes a balance between interpretability, sensitivity, and generalizability, thus seems to be the best approach [65].

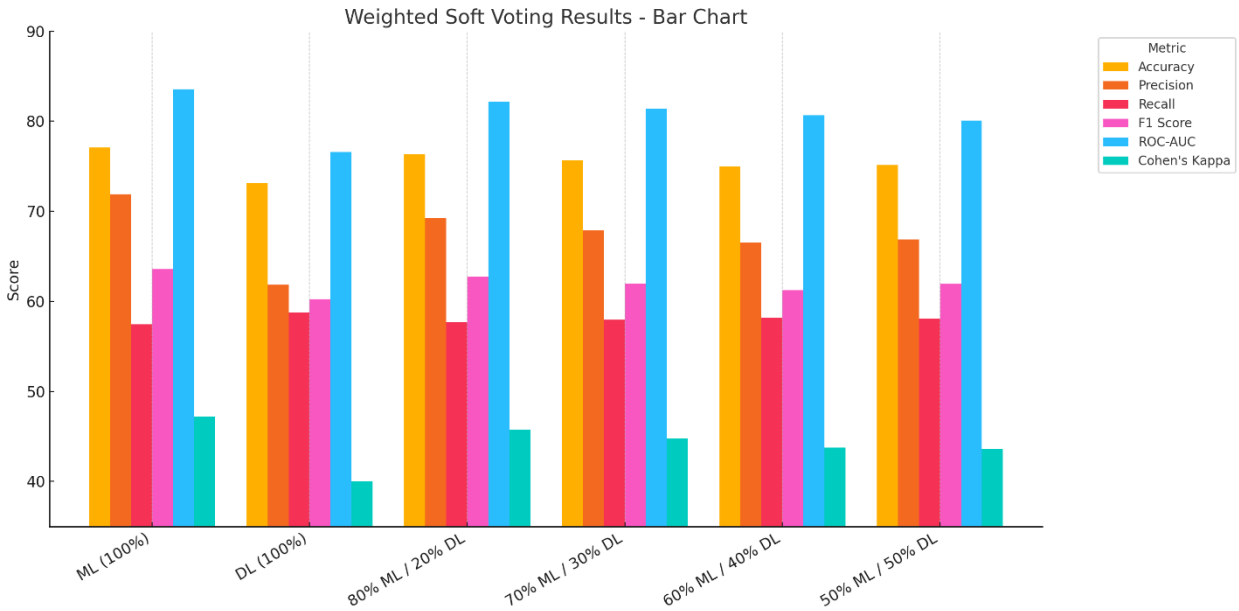

Figure 7. Weighted Soft Voting Experimental Results

H. Conclusion and Future Research

This article showed how powerful it is to combine machine learning (ML) and deep learning (DL) models using a weighted soft voting ensemble approach to improve diabetes prediction. The weighted soft voting method made a strong and

reliable predictive model by combining the best parts of both ML models, which are excellent at being accurate and stable, and DL models, which are better at finding differences. By adjusting the influence of each model type—specifically using 70% machine learning and 30% deep learning—the combined approach reached a good balance in performance, achieving 75.65% accuracy, 67.89% precision, and an ROC-AUC of 81.41%. These numbers all point to the model being well-rounded and able to make accurate predictions while keeping a good balance between precision and recall. The ROC-AUC score of 81.41% as shown in Table 4 that the model can reliably tell the difference between diabetes-positive and diabetes-negative patients. The Cohen's Kappa score of 44.76% shows that the model is mostly accurate, but there is still room for improvement.

The model works well, but it is evident that improving recall remains a crucial goal, particularly in Malaysia healthcare settings, where missing a positive diabetes case could have serious consequences. The results show that using a weighted ensemble of machine learning and deep learning is very useful in important fields like medical diagnostics, where accuracy and sensitivity are both crucial. The weighted soft voting method not only makes diabetes models more accurate, but it also gives you a flexible way to improve the balance between how well the model separates cases, how reliable it is, and how sensitive it is. Future research could improve this technique by optimizing the meta-model, adjusting probabilities, or adding new advanced methods to make the model perform even better. The results show that these combined models, especially when the weights are chosen carefully, could be very useful in the real world, such as in healthcare and other fields that need very accurate predictive models. Future research may use dynamic weighting strategies and meta-learning to improve ensemble effectiveness, apply to larger datasets, and explore potential enhancements for diverse populations.

Authorship Contribution Statement: Md. Ziarul Islam: Conceptualization, Methodology, Validation, Formal Analysis, Visualization, Writing – Original Draft, and Writing— Review and Editing. Mohd Khairul Azmi Bin Hassan: All supervision and mentoring, Conceptualization, Writing – Review & Editing, Validation, Data Curation. Amir 'Aatieff Bin Amir Hussin: Writing – Review & Editing, Investigation, Validation. Md Salman Sha: Writing – Review & Editing, Investigation, Validation. All authors have read and agreed to the submitted version of the manuscript.

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.

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

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

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.

Data Availability: The works used a publicly available dataset from Pima Indian Diabetes dataset [27].

I. References

[1] International Diabetes Federation, "Malaysia," *International Diabetes Federation*, 2021. https://idf.org/our-network/regions-and-members/western-pacific/members/malaysia/

[2] International Diabetes Federation, "Diabetes around the World in 2025," *IDF Diabetes Atlas*, 2025. https://diabetesatlas.org/

[3] Ministry of Health Malaysia, "NHMS 2023," *Institute for Public Health*, 2023. https://iku.gov.my/nhms-2023

[4] "Fact Sheet National Health and Morbidity Survey (NHMS) 2023," 2023. Available: https://iku.gov.my/images/nhms2023/fact-sheet-nhms-2023.pdf

[5] Z. Hussein, S. Wahyu Taher, H. K. Gilcharan Singh, and W. C. Siew Swee, "Diabetes Care in Malaysia: Problems, New Models, and Solutions," *Annals of Global Health*, vol. 81, no. 6, p. 851, Apr. 2016, doi: https://doi.org/10.1016/j.aogh.2015.12.016.

[6] Muhammad Shoaib Farooq, S. Riaz, Rabia Tehseen, U. Farooq, and K. Saleem, "Role of Internet of things in diabetes healthcare: Network infrastructure,

taxonomy, challenges, and security model," *Digital health*, vol. 9, p. 205520762311790-205520762311790, Jan. 2023, doi: https://doi.org/10.1177/20552076231179056.

[7] S. Akhtar, J. A. Nasir, A. Ali, M. Asghar, R. Majeed, and A. Sarwar, "Prevalence of type-2 diabetes and prediabetes in Malaysia: A systematic review and meta-analysis," *PLoS One*, vol. 17, no. 1, p. e0263139, Jan. 2022, doi: https://doi.org/10.1371/journal.pone.0263139.

[8] Mohamad Zulfikrie Abas, K. Li, Noran Naqiah Hairi, Wan Yuen Choo, and Kim Sui Wan, "Machine learning based predictive model of Type 2 diabetes complications using Malaysian National Diabetes Registry: A study protocol," *Journal of Public Health Research*, vol. 13, no. 1, Jan. 2024, doi: https://doi.org/10.1177/22799036241231786.

[9] Z. Guan *et al.*, "Artificial intelligence in diabetes management: Advancements, opportunities, and challenges," *Cell Reports Medicine*, vol. 4, no. 10, pp. 101213–101213, Oct. 2023, doi: https://doi.org/10.1016/j.xcrm.2023.101213.

[10] I. Dankwa-Mullan, M. Rivo, M. Sepulveda, Y. Park, J. Snowdon, and K. Rhee, "Transforming Diabetes Care Through Artificial Intelligence: The Future Is Here," *Population Health Management*, vol. 22, no. 3, pp. 229–242, Jun. 2019, doi: https://doi.org/10.1089/pop.2018.0129.

[11] K. J. Prabhod, "The Role of Artificial Intelligence in Reducing Healthcare Costs and Improving Operational Efficiency," *Quarterly Journal of Emerging Technologies and Innovations*, vol. 9, no. 2, pp. 47–59, Apr. 2024, Available: https://vectoral.org/index.php/QJETI/article/view/111

[12] A. A. Kuwaiti *et al.*, "A review of the role of artificial intelligence in healthcare," *Journal of Personalized Medicine*, vol. 13, no. 6, Jun. 2023, doi: https://doi.org/10.3390/jpm13060951.

[13] Tariq Osman Andersen, F. Nunes, L. Wilcox, Enrico Coiera, and Y. Rogers, "Introduction to the Special Issue on Human-Centred AI in Healthcare: Challenges Appearing in the Wild," *ACM Transactions on Computer-Human Interaction*, vol. 30, no. 2, pp. 1–11, Apr. 2023, doi: https://doi.org/10.1145/3589961.

[14] S. Kassim, "Winning The Battle Against Diabetes," *Gleneagles*. https://gleneagles.com.my/kuala-lumpur/articles/winning-the-battle-against-this-lifestyle-disease

[15] S. P. Chan, "Diabetes Care Model in Malaysia," *Journal of the ASEAN Federation of Endocrine Societies*, vol. 30, no. 2, p. 100, Nov. 2015, Available: https://www.asean-endocrinejournal.org/index.php/JAFES/article/view/255/663

[16] M. Khalifa and M. Albadawy, "Artificial intelligence for diabetes: enhancing prevention, diagnosis, and effective management," *Computer Methods and Programs in Biomedicine Update*, vol. 5, no. 100141, pp. 1–14, Feb. 2024, doi: https://doi.org/10.1016/j.cmpbup.2024.100141.

[17] B.-H. Chew *et al.*, "Efficient and Effective Diabetes Care in the Era of Digitalization and Hypercompetitive Research Culture: A Focused Review in the Western Pacific Region with Malaysia as a Case Study," *Health Systems &*

Reform, vol. 11, no. 1, Jan. 2025, doi: https://doi.org/10.1080/23288604.2024.2417788.

[18] Khondokar Oliullah, M. Rasel, M. M. Islam, Muhammad Rashedul Islam, M. Anwar, and Md Whaiduzzaman, "A stacked ensemble machine learning approach for the prediction of diabetes," *Journal of Diabetes & Metabolic Disorders*, Nov. 2023, doi: https://doi.org/10.1007/s40200-023-01321-2.

[19] M. F. Aslan and K. Sabanci, "A Novel Proposal for Deep Learning-Based Diabetes Prediction: Converting Clinical Data to Image Data," *Diagnostics*, vol. 13, no. 4, p. 796, Feb. 2023, doi: https://doi.org/10.3390/diagnostics13040796.

[20] K. Swee, E. Khor, W. Pei, C. Chua, and Fried, "Sustainability and Resilience in the Malaysian Health System MALAYSIA," 2024. Available: https://www3.weforum.org/docs/WEF_PHSSR_CAPRI_Malaysia_2024.pdf

[21] I. Z. Sadiq *et al.*, "Data-driven diabetes mellitus prediction and management: a comparative evaluation of decision tree classifier and artificial neural network models along with statistical analysis," *Scientific Reports*, vol. 15, no. 1, Jun. 2025, doi: https://doi.org/10.1038/s41598-025-03718-w.

[22] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, Feb. 2021, doi: https://doi.org/10.1016/j.icte.2021.02.004.

[23] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *Journal of Diabetes & Metabolic Disorders*, vol. 19, no. 1, pp. 391–403, Apr. 2020, doi: https://doi.org/10.1007/s40200-020-00520-5.

[24] H. Mohd, S. A. Sulaiman, A. Murad, N. Ahmad, C. C. Lang, and R. Jamal, "Genetics of type 2 diabetes (T2D) in Malaysia: a review," *Egyptian Journal of Medical Human Genetics*, vol. 26, no. 1, Jun. 2025, doi: https://doi.org/10.1186/s43042-025-00736-1.

[25] W. T. Sze and S. G. Kow, "Perspectives and Needs of Malaysian Patients With Diabetes for a Mobile Health App Support on Self-Management of Diabetes: Qualitative Study," *JMIR Diabetes*, vol. 8, no. 1, p. e40968, Oct. 2023, doi: https://doi.org/10.2196/40968.

[26] S. Gerke, T. Minssen, and G. Cohen, "Ethical and Legal Challenges of Artificial intelligence-driven Healthcare," *Artificial Intelligence in Healthcare*, vol. 1, no. 1, pp. 295–336, Jun. 2020, doi: https://doi.org/10.1016/B978-0-12-818438-7.00012-5.

[27] Kaggle, "Pima Indians Diabetes Database," *www.kaggle.com*, 2016. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[28] M. Kumar, K. Bajaj, B. Sharma, and S. Narang, "A Comparative Performance Assessment of Optimized Multilevel Ensemble Learning Model with Existing Classifier Models," *Big Data*, Dec. 2021, doi: https://doi.org/10.1089/big.2021.0257.

[29] Muhammad and N. S. Suriani, "Development of Diabetes Diagnosis Tool Using Machine Learning," *Evolution in Electrical and Electronic Engineering*, vol. 5, no. 1, pp. 1–9, 2024, Accessed: Oct. 16, 2024. [Online]. Available: https://penerbit.uthm.edu.my/periodicals/index.php/eeee/article/view/13194

[30] G. R. Ashisha, X. A. Mary, G. Mary, J. Andrew, and R. J. Eunice, "Random Oversampling-Based Diabetes Classification via Machine Learning Algorithms," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, Nov. 2024, doi: https://doi.org/10.1007/s44196-024-00678-3.

[31] H. B. Kibria, M. Nahiduzzaman, Md. O. F. Goni, M. Ahsan, and J. Haider, "An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI," *Sensors*, vol. 22, no. 19, p. 7268, Sep. 2022, doi: https://doi.org/10.3390/s22197268.

[32] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies*, vol. 13, no. 3, p. 88, Feb. 2025, doi: https://doi.org/10.3390/technologies13030088.

[33] Md. Alamin Talukder *et al.*, "Toward reliable diabetes prediction: Innovations in data engineering and machine learning applications," *Digital Health*, vol. 10, Jan. 2024, doi: https://doi.org/10.1177/20552076241271867.

[34] A. Hassan, S. G. Ahmad, T. Iqbal, E. U. Munir, K. Ayyub, and N. Ramzan, "Enhanced Model for Gestational Diabetes Mellitus Prediction Using a Fusion Technique of Multiple Algorithms with Explainability," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, Mar. 2025, doi: https://doi.org/10.1007/s44196-025-00760-4.

[35] Md Shamim Reza, R. Amin, R. Yasmin, Woomme Kulsum, and Sabba Ruhi, "Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data," *Heliyon (Londen)*, vol. 10, no. 2, pp. e24536–e24536, Jan. 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e24536.

[36] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022, Available: https://ieeexplore.ieee.org/abstract/document/9893798

[37] P. Yadav, S. C. Sharma, Rajesh Mahadeva, and S. P. Patole, "Exploring Hyper-parameters and Feature Selection for Predicting Non-communicable Chronic Disease using Stacking Classifier," *IEEE Access*, vol. 11, pp. 80030–80055, Jan. 2023, doi: https://doi.org/10.1109/access.2023.3299332.

[38] R. P. Tripathi *et al.*, "Timely Prediction of Diabetes by Means of Machine Learning Practices," *Augmented Human Research*, vol. 8, no. 1, Dec. 2023, doi: https://doi.org/10.1007/s41133-023-00062-4.

[39] K. Zhao and Z. Wang, "Research on Diabetes Prediction Based on Machine Learning," vol. 2016, pp. 29–33, Oct. 2023, doi: https://doi.org/10.1145/3635638.3635643.

[40] M. Zohair, R. Chandra, S. Tiwari, and S. Agarwal, "A model fusion approach for severity prediction of diabetes with respect to binary and multiclass classification," *International Journal of Information Technology*, vol. 16, no. 3, pp. 1955–1965, Oct. 2023, doi: https://doi.org/10.1007/s41870-023-01463-9.

[41] M. Imani, A. Beikmohammadi, and Hamid Reza Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN,

and GNUS Upsampling under Varying Imbalance Levels," Feb. 2025, doi: https://doi.org/10.20944/preprints202501.2274.v2.

[42] C.S. Manikandababu, S. IndhuLekha, J. Jeniefer, and T. Annie Theodora, "Prediction of Diabetes using Machine Learning," *2022 International Conference on Edge Computing and Applications (ICECAA)*, Oct. 2022, doi: https://doi.org/10.1109/icecaa55415.2022.9936375.

[43] J. Abdollahi and B. Nouri-Moghaddam, "Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction," *Iran Journal of Computer Science*, Mar. 2022, doi: https://doi.org/10.1007/s42044-022-00100-1.

[44] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, "Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review," *Diabetology & Metabolic Syndrome*, vol. 14, no. 1, Dec. 2022, doi: https://doi.org/10.1186/s13098-022-00969-9.

[45] T. Ekemini, J. Agbogun Phd, N. Benjamin, and Godfrey, "DEVELOPMENT OF A PREDICTIVE MACHINE LEARNING MODEL FOR DIABETES USING STACKED ENSEMBLE APPROACH," *International Journal of Artificial Intelligence Trends (IJAIT)*, vol. 2, no. 47, pp. 495–507, 2023, Available: https://www.ijortacs.com/uploads/papers/a3335bea41e25b62537f9d2e9001e58c-ijortacs.pdf

[46] S. K. Kalagotla, S. V. Gangashetty, and K. Giridhar, "A novel stacking technique for prediction of diabetes," *Computers in Biology and Medicine*, vol. 135, p. 104554, Aug. 2021, doi: https://doi.org/10.1016/j.compbiomed.2021.104554.

[47] N. Li, L. Ma, G. Yu, B. Xue, M. Zhang, and Y. Jin, "Survey on Evolutionary Deep Learning: Principles, Algorithms, Applications and Open Issues," *ACM Computing Surveys*, Jun. 2023, doi: https://doi.org/10.1145/3603704.

[48] N. Singh and P. Singh, "Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 1–22, Jan. 2020, doi: https://doi.org/10.1016/j.bbe.2019.10.001.

[49] A. A. Alzubaidi, S. M. Halawani, and M. Jarrah, "Integrated Ensemble Model for Diabetes Mellitus Detection," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 4, Apr. 2024, doi: https://doi.org/10.14569/IJACSA.2024.0150423.

[50] O. S. Alshehri, O. M. Alshehri, and Hussein Samma, "Blood Glucose Prediction Using RNN, LSTM, and GRU: A Comparative Study," pp. 1–5, Apr. 2024, doi: https://doi.org/10.1109/ic_aset61847.2024.10596176.

[51] A. M. Carrington *et al.*, "Deep ROC Analysis and AUC as Balanced Average Accuracy to Improve Model Selection, Understanding and Interpretation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1–1, 2022, doi: https://doi.org/10.1109/TPAMI.2022.3145392.

[52] O. R. Olaniran, A. O. Sikiru, J. Allohibi, A. A. Alharbi, and N. M. Alharbi, "Hybrid Random Feature Selection and Recurrent Neural Network for Diabetes Prediction," *Mathematics*, vol. 13, no. 4, p. 628, Feb. 2025, doi: https://doi.org/10.3390/math13040628.

[53] Blessing Oluwatobi Olorunfemi *et al.*, "Efficient diagnosis of diabetes mellitus using an improved ensemble method," *Scientific Reports*, vol. 15, no. 1, Jan. 2025, doi: https://doi.org/10.1038/s41598-025-87767-1.
[54] B. Motamedi and B. Villányi, "A predictive analytics approach with Bayesian-optimized gentle boosting ensemble models for diabetes diagnosis," *Computer Methods and Programs in Biomedicine Update*, vol. 7, p. 100184, 2025, doi: https://doi.org/10.1016/j.cmpbup.2025.100184.
[55] V. O. Khilwani, V. Gondaliya, S. Patel, J. Hemnani, B. Gandhi, and S. K. Bharti, "Diabetes Prediction, using Stacking Classifier," *IEEE Xplore*, Sep. 01, 2021. https://ieeexplore.ieee.org/abstract/document/9670920 (accessed May 17, 2023).
[56] M. Gollapalli *et al.*, "A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM," *Computers in Biology and Medicine*, vol. 147, p. 105757, Aug. 2022, doi: https://doi.org/10.1016/j.compbiomed.2022.105757.
[57] A. Dutta *et al.*, "Early Prediction of Diabetes Using an Ensemble of Machine Learning Models," *International Journal of Environmental Research and Public Health*, vol. 19, no. 19, p. 12378, Sep. 2022, doi: https://doi.org/10.3390/ijerph191912378.
[58] S. Xie, Z. Yu, and Z. Lv, "Multi-Disease Prediction Based on Deep Learning: A Survey," *Computer Modeling in Engineering & Sciences*, vol. 128, no. 2, pp. 489–522, 2021, doi: https://doi.org/10.32604/cmes.2021.016728.
[59] Nahid Hossain Taz, A. Islam, and I. Mahmud, "A Comparative Analysis of Ensemble Based Machine Learning Techniques for Diabetes Identification," *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, vol. 15, pp. 1–6, Jan. 2021, doi: https://doi.org/10.1109/icrest51555.2021.9331036.
[60] S. Kumar and Vijay Kumar Jha, "Diabetes prediction model using machine learning techniques," *Multimedia Tools and Applications*, Oct. 2023, doi: https://doi.org/10.1007/s11042-023-16745-4.
[61] A. Zarghani, "Comparative Analysis of LSTM Neural Networks and Traditional Machine Learning Models for Predicting Diabetes Patient Readmission," *arXiv.org*, 2024. https://arxiv.org/abs/2406.19980
[62] Emmanuel Imuede Oyasor and A. D. Gbadebo, "Using the Machine Learning Algorithms for Accurate Prediction of Diabetes," *The Indonesian Journal of Computer Science*, vol. 13, no. 6, Dec. 2024, doi: https://doi.org/10.33022/ijcs.v13i6.4488.
[63] M. Saleh *et al.*, "An Innovative Ensemble Deep Learning Clinical Decision Support System for Diabetes Prediction," *IEEE Access*, vol. 12, pp. 106193–106210, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3436641.
[64] Md. K. Hasan, Md. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020, doi: https://doi.org/10.1109/access.2020.2989857.
[65] P. Sampath *et al.*, "Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique," *Scientific*

Reports, vol. 14, no. 1, Nov. 2024, doi: https://doi.org/10.1038/s41598-024-78519-8.