Enhanced Multi-Class Hybrid Machine Learning Techniques Model for Predicting Diabetes Mellitus

Ntia U. J.
Department of Cyber Security
University of Uyo
Akwa, Ibom State, Nigeria

Asagba P. O.
Department of Computer Science
University of Portharcourt
Rivers State, Nigeria

Okengwu U. A.
Department of Computer Science
University of PortHarcourt
Rivers State, Nigeria

Abstract—A widespread chronic condition known as diabetes, poses significant healthcare challenges globally. Machine Learning (ML) algorithms have been shown to aid in enhancing the predictive models of diabetes, yet, it is comparatively difficult for single-model approaches to provide maximum probable predictive accuracy. For early intervention, proper management of the condition and prevention of delayed treatment, accurate prediction of a greater number of true positive cases is necessary. This research aims to create a strong and consistent hybrid model for diabetes occurrence prediction with high recall value. Employing an ensemble of various ML algorithms through the utilization of voting technique, this blended model combines the predictive power of Gradient Boosting, Decision Trees, Naïve Bayes, ANN, Random Forest, and Logistic Regression and addresses missing values through the Simpleimputer function using median imputation, outliers through Winzorisation, class imbalance through Balanced Bagging Classifier, and hyperparameter optimization through Grid-search. It applies object-oriented analysis and design methodology. Programming language, Python, is used to accessed the Jupyter Notebook IDE for implementation. Evaluation metrics used to assess the hybrid model are recall, Micro-average, precision, macro-averages, F1-score, AUC and accuracy. The model provides accuracy of 0.77, a recall value of 0.90. Using a voting approach, the study maximizes the predictive accuracy for prediction of diabetic disease using the exclusive strengths of such heterogeneous algorithms. Diabetes outcome predictions are made easier if ML algorithms are used wisely along with a soft voting mechanism.

Index Terms- Diabetes Disease, Machine Learning Algorithms, Hybrid Model, Feature Selection.

I. INTRODUCTION

When insulin production is inadequate or when the body has trouble metabolizing the insulin that is produced, the result is diabetes, a chronic disease [1]. The hormone insulin controls blood sugar levels. [2] noted the yearly damage rate of 1.6 million caused by a non-communicable disease, Diabetes Mellitus. A chronic illness is one that does not disappear or has lasting effects. It is a disease where no permanent cure is found **[3].** Insulin is a hormone that regulates blood glucose [1]. Lack of this hormone caused Diabetes **[4].** One typical complication of uncontrolled diabetes is hyperglycemia, which can be defined as an elevated blood glucose level or blood sugar level. It happens when glucose concentrations in the bloodstream are unusually high as the result of inability of the body to control blood sugar levels. Among those 18 and up, 8.5% had diabetes in 2014. In 2019, about 1.5 million individuals died due to diabetes, with roughly half of those deaths being in the under-70 age group. The patient's thirst, hunger, heart illness, kidney disease, etc., can all worsen or perhaps cause their death as a result of this sickness [5]. Managing the condition is difficult due to challenges faced to perfectly predict the specific type, an unclear situation for medical professionals [6]. but researchers are working tirelessly to address this problem. Type1, Type2 and Gestational Diabetes are the three most common forms of the chronic disease diabetes [1]. Adults are prone Type 2 diabetes, whereas children are more likely to have Type 1. There is a correlation between pregnancy and gestational diabetes. Curiously, Type 2 diabetes affects about 95% of all diabetics, as World Health Organization (WHO) reported. As a stepping stone to preventing diabetes, there are a number of intermediate disorders, including impaired glucose tolerance (IGT) and impaired fasting glucose (IFG). Although it is not always the case, people with these illnesses exhibit a significant chance of developing full-blown Type 2 diabetes. Recognizing the increasing worldwide burden of diabetes, the WHO is making great strides to encourage and facilitate the implementation of effective strategies for the monitoring, avoidance and management of diabetes and its consequences, with a focus on nations with low or medium income levels. A worldwide effort to enhance diabetes care and prevention, the Global Diabetes Compact was established in April 2021 by the WHO. Its primary goal is to assist low and middle-income nations. The World Health Assembly (WHA) also passed a resolution in May 2021 to better manage and prevent diabetes. After this, in May 2022, the WHO endorsed five worldwide goals for diabetes treatment and coverage that are to be reached by 2030. WHO is responsible for: Facilitating innovative guideline for the prevention of major non-communicable diseases including diabetes, developing norms and standards for diabetes diagnosis and care, building awareness on the global epidemic of diseases, marking World Diabetes Day, Conducting surveillance of Diabetes and its risk factors. According to [6], it is possible to predict diabetes without relying on diagnostic laboratory tests but only on machine Learning (ML) techniques. Moreover, conventional statistical models and standalone machine learning algorithms are incapable of modeling complex relationships and structures in data, resulting in poor performance on key metrics such as recall and precision. This research seeks to solve this issue by designing a novel, high-performance hybrid prototype that can predict the occurrence of various types of diabetes with great recall. The goal is to improve the model's recall ensuring that the maximum number of true positive cases are correctly identified enabling timely intervention and minimizing false negative occurrences that can lead to missed or delayed diagnose. This will help in effective management of patients with diabetes.

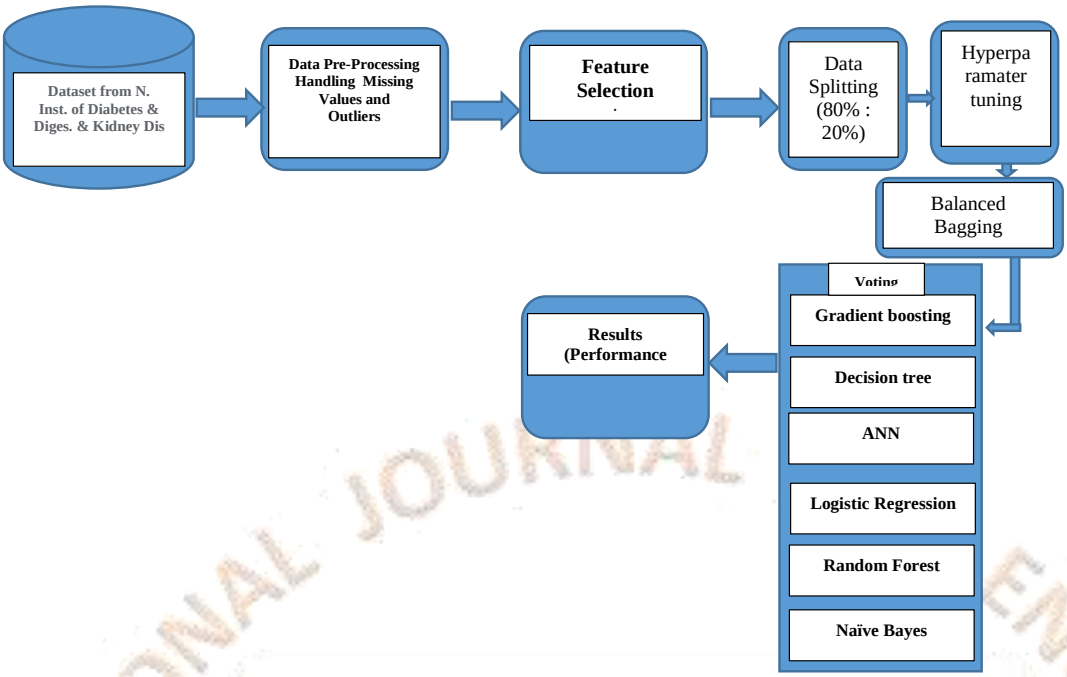


Fig 1. Architecture of the Hybrid Model

II. LITERATURE SURVEY

[7] propounded a hybrid type 2 diabetes prediction model that takes into consideration class imbalance. Logistic Regression, K-Nearest Neighbors, Support Vector Machine (SVM), Gaussian Naïve Bayes, and Random Forest are some of the machine learning algorithms used in the study. There are 390 African-Americans data collected with variables age, weight, glucose, cholesterol, and waist-to-hip ratio that were obtained from the biostatistics program in Vanderbilt University. Oversampling and undersampling were employed to keep the dataset in balanced form, and filter and wrapper techniques were used to select features. The classifier was a voting ensemble model on all five methods, and the model was validated by cross-validation technique known as stratified K-fold. F1-score, Recall, precision and accuracy were used. Improving recall with the appropriate precision was the main objective, as evidenced by comparisons between the sampled and unsampled models, which led to substantial recall gains. The result obtained proves that this work has better recall value of 0.88 compared to other work used for benchmarking.

[8] proposed a type 2 diabetes occurrence prediction model through an ensemble of adaptive boosting and light gradient boosting machines. This research utilized data accessed through Kaggle form National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), There are a total of 768 instances and eight attributes along with one class attribute in each. They are 268 diabetic and 500 non-diabetic patients for a total of 65.1% healthy and 34.9% diabetics in the dataset. They are all female Pima Indians and all are older than the age of 21. There are eight medical predictor variables available in the data: body mass index, pregnancy number, age, insulin level, diabetes family history function, glucose, skin thickness, and blood pressure. A combination of models, including k-NN, RF, AdaBoost, Naive Bayes (Gaussian), and a new model Light Gradient Boosting Machine, is used through an ensemble technique. 90.76% accuracy is achieved when k-NN, AdaBoost, and LightGBM are applied together. The recall value for the ensemble model combining LightGBM, k-NN, and Adaboost was 85.82%.

[9] built a machine learning model for the prediction of diabetes onset based on an integration of explainable AI and machine learning. DT, SVM, Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and ensemble methods were some of the machine learning classification algorithms used by authors to identify the most precise prediction algorithm. These were also processed together with a proprietary dataset as well as with the Pima Indian dataset that is publicly distributed. The latter consisted of 768 people with 268 diabetic patients. Pregnancy, glucose, blood pressure, skin thickness, body mass index (BMI), and diabetes outcome are six parameters in the dataset from Rownak Textile Mills Ltd RTML) in Dhaka, Bangladesh, with 203 females aged between 18 and 77 years. The public dataset has eight parameters with the recall value of 0.80.

[10] built an improved diabetes prediction model based on ensemble models amid COVID-19 pandemic. Among the features of the model are Gradient Boosting, Decision Trees, and Support Vector Machines. The Pima Indians Diabetes Database and includes nine features which are Skin Thickness, Insulin, Diabetes Pedigree Function, Age, Blood Pressure, Glucose, and Outcome and contained 768 records. The ensemble model achieved a recall of 88.5%.

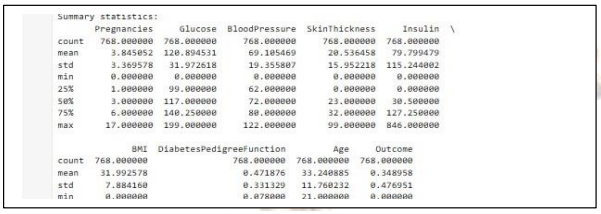
III. MATERIALS AND METHODS

A. Dataset

The dataset was from National Institute of Diabetes, Digestive and Kidney Diseases. It was downloaded from kaggle data repository. The dataset is of size 768 comprises of 8 independent attributes and 1 dependent attribute. Age, Diabetes Pedigree Function (DPF), Skin Thickness, Insulin, Blood Pressure, Glucose, and Pregnancies are the independent attributes. The result is the dependent attribute.

B. Data Pre-processing

The presence of many outliers in the dataset, combined with missing or null values, influences the outcomes of diabetes classification [11]. The data pre-processing phase handles outliers and missing data. Outliers was detected through with Inter Quartile Range (IQR) and Z-scores methods. Box plot was used to visualize the Outliers. Winsorization method was used to handle the known outliers. Missing values was handled with Simpleimputer function through median imputation.



Summary statistics:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
count	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479
std	3.369578	31.972618	19.355807	15.952218	115.244002
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000
75%	6.000000	140.250000	80.000000	32.000000	127.250000
max	17.000000	199.000000	122.000000	99.000000	846.000000

	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000
mean	31.992578	0.471876	33.240885	0.348958
std	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.078000	21.000000	0.000000

Fig. 2 Summary Statistics for Outliers Detection

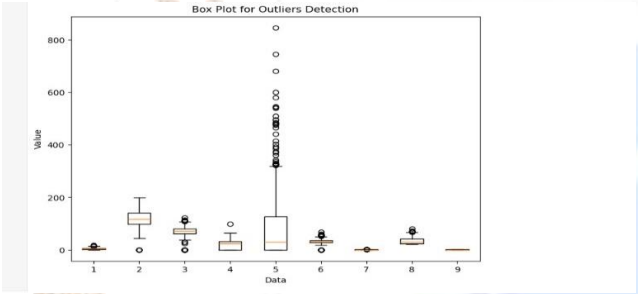


Fig. 3 Box Plot for Outliers Detection

C. Feature Selection

The most influential diabetes risk factors considered in this work are Glucose and Diabetes Pedigree Function (DPF). Blood Sugar level (Glucose) of **125** mg/dL (Milligrams per decilitre) and above and DPF of **0.05** and above indicates the presence of diabetes. Other attributes that help to further classify the specific type of diabetes are BMI, Pregnancies, Age and Insulin.

D. Data Splitting

An integral aspect of machine learning is data splitting, which involves dividing the dataset into distinct groups for testing and training. To make sure your model can generalize well to fresh, unknown data, data splitting is crucial, as it offers a dependable method for assessing its performance. The dataset was splitted in the ratio 80 % for training set and 20% for testing set.

E. Hyperparameter Tuning

The optimization of a model's performance is goal of hyperparameter tuning. Its stability, performance, training time and avoidance of overfitting are all better with its help. This model uses grid search technique for tuning.

F. Sampling

Balanced Bagging Classifier – oversampling and undersampling

Combining bagging and class balancing, the Balanced Bagging Classifier is an ensemble learning approach that enhances classification model performance, especially in the presence of imbalanced datasets. It has two features Bagging and Class Balancing. Bagging (Bootstrap Aggregating) involves training numerous models on separate portions of the training data and then combining their predictions to produce a single forecast. Predictions become more reliable and resilient as a result of less variance among the individual models. Class balancing addresses the issue of poor performance of classification models by applying class balancing techniques to the individual base models. The Common class balancing methods used include over-sampling the minority class, under-sampling the majority class, or using weighted sampling during the bagging process. Classes that have a significant number of instances are called majority classes, while those with fewer instances are referred to as minority classes [12].

IV. ALGORITHMS

A. Gradient Boosting

Gradient boosting is a strong machine learning method where many weak models (usually decision trees) are used and fused into one strong predictive model. The process is designed to add new models to an ensemble in an iterative manner, working to correct errors introduced by past models. The main concept is to construct the models sequentially such that each model in the ensemble must be capable of learning from the errors of the preceding ones. Gradient boosting is specifically beneficial because it can manage different kinds of data and has proven to be effective in competitions like those on Kaggle, where it has consistently achieved strong results. [13]

B. Artificial Neural Network

One form of machine learning techniques is artificial neural network (ANN), which attempts to emulate the way the brain's own neural networks work. It is meant to emulate the way the brain learns and processes information. Artificial neural networks (ANNs) are trained to learn new information. ANN functions by altering the weights of the connections between neurons according to the input data and the associated output, which enables the network to reduce prediction errors. This is typically accomplished using methods like backpropagation, where the network improves by learning from its training errors.[14]

C. Decision Tree

Decision trees are used as a method for classification and prediction with representation using nodes and internodes [15]. In decision tree, there are nodes and edges, where nodes represent features and branches represent the outcome [16]. Decision trees are supervised machine learning models with two uses: classification and regression. This tree-like model contains nodes, branches, and leaves. The decision tree approach generates predictions or options using a hierarchical tree structure that is created by recursively partitioning the input space according to the data attributes

D. Logistic Regression

A statistical model that can be employed to approximate the chance of a binary or dichotomous response given a collection of predictor variables is logistic regression. The dependent variable for this form of regression analysis can only be one of two different values, i.e., “yes/no” or “0/1”. It can be applied to examine the relationship between a binary outcome and a group of covariates, underscoring its utility in interpreting results and evaluating model performance [17].It uses a set of independent variables to try to forecast the likelihood that the dependent variable will take one of two possible values. Using the dependent variable's log-odds as input, the model approximates the relationship between the independent variables, which are the natural logarithm of the odds. Many areas are reliant on Logistic Regression as a statistical tool, including computer science, public health, and medicine The model estimates the relationship between the independent variables and the log odds (the natural logarithm of the odds) of the dependent variable.

The logistic regression model can be expressed as

$$\text{logit}(p) = \log(p / (1 - p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \qquad \text{Equation 1.1}$$

Where

p is the probability of the dependent variable being equal to one of the two possible outcomes

β_0 is the intercept term $\beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients for the independent variables $x_1, x_2, \ldots, x_n$, respectively. The logit $(p) = \log (p / (1 - p))$ is used to ensure that the predicted probabilities remain within the range of 0 to 1.

E. Naïve Bayes

The Gaussian Naive Probabilistic method is used to determine how likely it is that certain data points belong to a specific class based on existing experience [18]. It is observed that Naive Bayes classifier greatly simplify learning by assuming that features are independent in a given class [**19].** Naïve Bayes can be used for many classification tasks, it is robust to irrelevant features and can handle missing data and it constructs a model of the relationship between data features and their respective labels or classes. It predicts based on observed feature values by selecting the most probable class. Naïve Bayes is also acclaimed for its computational efficiency and simplicity, and is surprisingly known to work well, even though it is making strong independence assumptions about the features.

F. Random Forest

Random forests are a type of ensemble method which makes predictions by averaging over the predictions of several independent base model [20]. It is widely utilized for classification and regression tasks in machine learning. This algorithm builds upon the decision tree classifier and operates using a collection of decision trees, each influenced by a compilation of random variables [21]. Decision trees serve as the foundational components of the Random Forest model. It becomes most popular choice among machine learning practitioners due to its effectiveness in addressing classification and regression challenges, versatility and resilience to handle multidimensional information. It was its top performance and tolerance that made it gain prominence through its introduction in 2001 by Leo Breiman.

G. Hybrid Model

A hybrid model combines many algorithms including ElasticNet, K-means, and Gaussian process regression (GPR) in a judicious way in order to maximize the strengths of each algorithm and reduce their weaknesses [22]. Using this method, experts can maximize the strengths of each algorithm and reduce their weaknesses. Hybrid models are a powerful method of dealing with difficult problems in many ML applications namely deep learning, ensemble learning, and transfer learning. They are priceless in today's data-driven world owing to their adaptability, which is necessary in the management of complexities in modern data environments, which in turn increase innovation and predictive strength in a wide range of applications and industries.

This work combines six ML algorithms explained above to leverage on the strengths of different models, leading to improved performance and generalization of the final predictive model. This work employed soft voting approach for final predictions. The voting approach has two main components: base models and the voting mechanism. First, multiple base models namely decision trees, logistic regression and other suitable algorithms are trained on the same dataset. These base models use different algorithms, feature sets and hyperparameter configurations. Once the base models have been trained, the voting mechanism combines their predictions to make the final prediction. Two main types of voting exist namely majority voting (hard voting), The final prediction is the class or label that receives the most votes from the individual base models. Weighted voting (soft voting), The final prediction is the class or label with the highest weighted average of the base models' probabilities or confidence scores.

H. Algorithm of the Enhanced Hybrid Model

Step 1 Importation of relevant libraries
Import libraries
Step 2 Load Dataset
Load dataset
Step 3 Data Pre-processing
1. Clean Data
a. Handle outliers through winzorisation.
b. Create imputer object called transformer to enable imputation of mean values.
c. handle missing value with median imputation using fillna.
2. Feature Selection
a. Create ‘outcome’ and ‘type’ columns to determine diabetes and non-diabetes.
b. Identify relevant features and make conditions.
c. Make choices to specify “Gestational diabetes”, “Type2” or “Type1”.
3. Encoding
a. Convert categorical variables to numerical format.
Step 4 Data Splitting
a. Split Dataset by dividing it into training (80%) and testing (20%).
Step 5 Hyperparameter tuning
1. Define hyperparameter search space for each of the machine learning algorithm
2. Create Balanced bagging classifier to perform hyperparameter tuning using search
Step 6 Build the Model
1. Build the hybrid model using Voting approach with the following ML algorithms
Gradient Boosting
ANN
Decision Tree
Logistic Regression
Random Forest
Gaussian Naïve Bayes
2. Train the Hybrid model
3. Fit the hybrid pipeline
4. Make predictions
Step 7 Save the hybrid pipeline model
1. Save the model
2. Load the saved model
3. Make prediction
Step 8 Model Evaluation
1. Calculate the AUC for each class
2. Print AUC for each class
3. Plot the ROC for each class
4. Calculate the macro-average AUC
5. Calculate the micro-average AUC
6. Print the Macro-average and Micro-average results
7. Evaluate and print classification report
8. Evaluate and print confusion matrix
9. Define the scorer using AUC
10. Perform cross-validation
11. Print cross-validated AUC scores and mean AUC score.

V. PERFORMANCE METRICS

A. Accuracy

Accuracy evaluation metrics quantify the overall correctness of a model’s predictions, encompassing both positive and negative instances. A high accuracy score signifies that the model is effectively identifying the target variable across all relevant classes. This model provides an accuracy of 0.77.

B. Precision

Precision assesses a portion of the model’s positive predictions that were accurate. A high precision value signifies the model’s capability to minimize false positive identifications and effectively detect the true positive instances. 0.52 was recorded as precision of this model.

C. Recall

Recall evaluation metrics assess how well a model or system identifies the actual positive cases, highlighting its capacity to detect all relevant instances, with a higher recall value indicating the model's effectiveness at minimizing false negatives. A recall value of 0.90 was achieved.

D. F1 Score

The F1-score is a combined metric that finds the harmonic mean of precision and recall, offering a balanced assessment of a model's performance. It accounts for the model's capacity to accurately identify positive cases (recall) and its precision in making positive predictions, providing a more holistic evaluation of the model's overall effectiveness. This model provides F1-Score of 0.54

E. Area Under Curve (AUC)

The AUC evaluates a model's overall performance by measuring the balance of the true positive rate and false positive rate across different classification thresholds. A higher AUC value implies the model's enhanced ability to differentiate between the positive and negative classes, making it an effective indicator of the model's overall discriminative capability.

F. Micro Average

The micro-average evaluation metric computes the overall precision, recall and F1-score by treating all individual predictions as equally contributing to the final metric. This approach is beneficial for datasets with imbalanced classes, as it ensures that the performance on each class is weighted equal than being disproportionately influenced by the majority class.

G. Macro Average

The macro-average evaluation metric computes the midpoint of the precision, recall and F1-score for each individual class and then takes the average of those values. The method assigns equal weight to the performance on each class, making it beneficial for evaluating models on the datasets with imbalanced class distribution. The goal is to accurately detect every type of diabetes, which is super important because missing a case as a result of a false negative could lead to serious health risks.

H. Area under the Receiver Operating Characteristic curve (AUC).

AUC is an important metric for assessing the performance of multi-class classification models in machine learning. It evaluates how effectively a model separates positive from negative classes. ROC curve illustrates the True Positive Rate (TPR) against the False Positive Rate (FPR) across different threshold levels. AUC is especially valuable for imbalanced datasets, where accuracy alone can be misleading. AUC basically represents the likelihood that the model will rank one sample higher than another. For instance, an AUC of 0.75 indicates a 75% chance for the model to correctly rank a positive sample above a negative one.

VI. RESULTS

The model was implemented in Jupyter notebook framework using python through scikit-learn. The result was recorded in a tabular format. The results started from individual machine learning algorithms to the hybrid model. The result for Gradient Boosting, ANN, Decision Tree, Naïve Bayes, Logistic Regression and Random Forest showed outstanding performance for Decision Trees with recall of 1.00, precision of 0.84 and accuracy of 0.90. After leveraging on the strength of the six machine learning algorithms listed above, through a voting approach and applying grid search method for hyperparameter tuning, the hybrid model gave a recall of 0.90 as opposed to 0.88 by that of existing system [7]. The results are shown below

Table 1 shows the results of individual ML algorithms and that of hybrid model of the proposed system. The recall scores range from 0.67 to 1.00, accuracy from 0.77 to 0.98 for individual ML algorithms. Recall ranges from 0.74 to 1.00, Macro average recall is 0.90 and accuracy of 0.77 for the hybrid model.

Table 1 Results of Individual Machine Learning Algorithms

Gradient Boosting					
Type	**Precision**	**Recall**	**F1-score**	**Accuracy**	**AUC**
Gestational	**0.57**	**1.00**	0.72	**0.818**	**1.00**
Type1	**0.08**	**1.00**	0.14		**1.00**
No diabetes	**1.00**	**0.79**	0.88		**0.97**
ANN					
Type	**Precision**	**Recall**	**F1-score**	**Accuracy**	**AUC**
Gestational	**0.43**	**1.00**	0.60	**0.98**	**0.99**
Type1	**0.06**	**1.00**	0.12		**0.99**
No diabetes	**1.00**	**0.67**	0.81		**0.97**
Decision Tree					
Type	**Precision**	**Recall**	**F1-score**	**Accuracy**	**AUC**
Gestational	**0.51**	**1.00**	0.68	**0.811**	**0.997**
Type1	**0.10**	**1.00**	0.18		**1.00**
No diabetes	**1.00**	**0.78**	0.88		**0.977**
Logistic Regression					
Type	**Precision**	**Recall**	**F1-score**	**Accuracy**	**AUC**
Gestational	**0.38**	**1.00**	0.55	**0.77**	**0.96**
Type1	**0.00**	**0.00**	0.00		**0.91**
No diabetes	**0.99**	**0.74**	0.85		**0.94**

Random Forest					
Type	**Precision**	**Recall**	**F1-score**	**Accuracy**	**AUC**
Gestational	**0.55**	**1.00**	0.71	**0.84**	**0.99**
Type1	**0.12**	**1.00**	0.22		**1.00**
No diabetes	**1.00**	**0.82**	0.90		**0.97**
Naïve Bayes					
Type	**Precision**	**Recall**	**F1-score**	**Accuracy**	**AUC**
Gestational	**0.50**	**0.86**	0.63	**0.85**	**0.95**
Type1	**0.33**	**1.00**	0.50		**1.00**
No diabetes	**0.97**	**0.85**	0.91		**0.94**
Hybrid Model					
	Precision	**Recall**	**F1-score**	**Accuracy**	**AUC**
Gestational	**0.51**	0.95	0.67	**0.77**	**0.98**
Type1	**0.06**	1.00	0.12		**1.00**
No diabetes	**0.99**	0.74	0.85		**0.95**
Macro Ave.	**0.52**	0.90	0.54		
Micro Ave.	**0.92**	0.77	0.82		

This is a comparison of several machine learning models namely Gradient Boosting, ANN, Decision Tree, Logistic Regression, Random Forest, Naïve Bayes, and a Hybrid Model used to classify cases based on three diabetes types namely Gestational diabetes, Type 1 diabetes and No diabetes since there is no Type 2 diabetes sample in the testing dataset. Each model's performance is measured using 1. Precision which estimates how many predicted positives were actually correct. 2. Recall which measures how many actual positives were correctly predicted. 3. F1-score which is the balance between precision and recall 4. Accuracy which is the overall correctness of the predictions 5. AUC which shows how well the model distinguishes between classes, especially useful with imbalanced data. Results for individual machine learning algorithms indicates good performances but varies for the different evaluation metrics. Some have perfect scores which can be as a result of inherent deficiencies since there are disparities in the results when merged together to form hybrid.

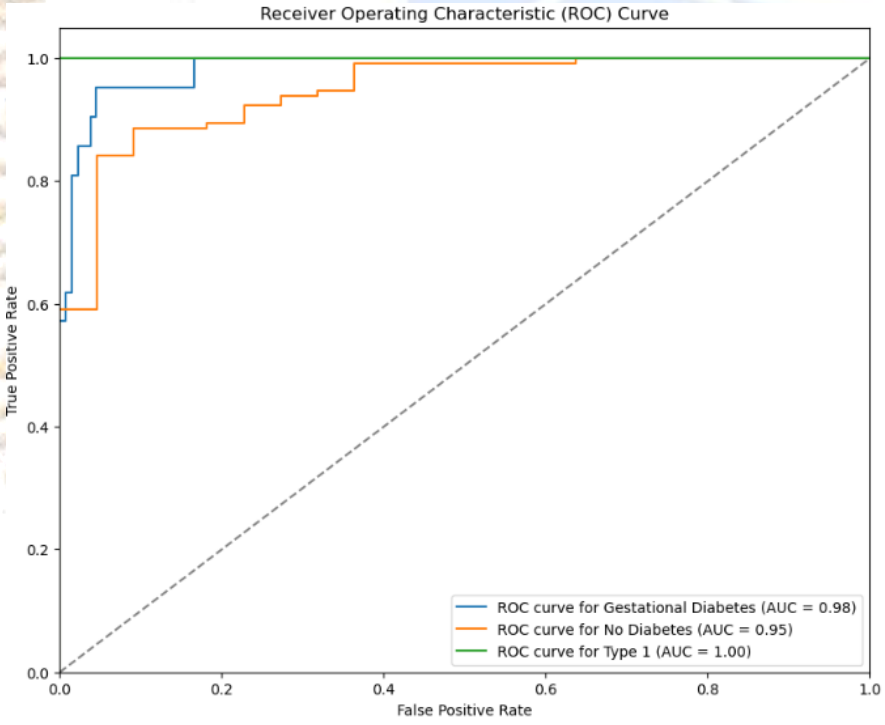


Fig. 4 Receiver Operating Characteristic (ROC) Curve of the Hybrid System

Table 2 Area Under the ROC Curve (AUC)

Diabetes Class	AUC Scores	Interpretations
Gestational Diabetes	0.98	Excellent performance, almost perfect distinction from other classes.
No Diabetes	0.95	Very strong, the model reliably identifies people without diabetes.
Type 1 Diabetes	1.00	Perfect score, model made flawless predictions for this class.
Type 2 Diabetes	Nil	There is sample of Type2 diabetes in the testing dataset .

AUC estimates how well a model can distinguish between classes. It ranges from 0 to 1. 1.00 means perfect separation of classes, 0.50 implies no better than random guessing and <0.50 suggest that model is misclassifying.

```
Tuning hyperparameters for rf...
Best parameters for rf: {'classifier__class_weight': 'balanced', 'classifier__max_depth': None, 'classifier__min_samples_split': 2, 'classifier__n_estim
ators': 50}
Tuning hyperparameters for ann...
Best parameters for ann: {'classifier__activation': 'relu', 'classifier__hidden_layer_sizes': (50, 50), 'classifier__learning_rate_init': 0.01}
Tuning hyperparameters for gb...
Best parameters for gb: {'classifier__learning_rate': 0.1, 'classifier__max_depth': 3, 'classifier__n_estimators': 100}
Tuning hyperparameters for dt...
Best parameters for dt: {'classifier__max_depth': None, 'classifier__min_samples_split': 5}
Tuning hyperparameters for nb...
Best parameters for nb: {}
Tuning hyperparameters for lr...
Best parameters for lr: {'classifier__C': 10, 'classifier__penalty': 'l2'}
Hybrid model with tuned hyperparameters saved successfully.
Model Accuracy: 0.7727272727272727
                    precision    recall  f1-score   support

Gestational Diabetes       0.51      0.95      0.67        21
         No Diabetes       0.99      0.74      0.85       132
              Type 1       0.06      1.00      0.12         1

            accuracy                           0.77       154
           macro avg       0.52      0.90      0.54       154
        weighted avg       0.92      0.77      0.82       154

[[20  1  0]
 [19 98 15]
 [ 0  0  1]]
AUC for Gestational Diabetes: 0.9838882921589689
AUC for No Diabetes: 0.9483471074380165
AUC for Type 1: 1.0
```

Fig. 5 Evidence of the Results of Evaluation Metrics and Hyperparameter Tuning

Table 3 Hyperparameter Tuning Summary

Model	**Key Parameters**
rf (Random Forest)	Balanced class weights, 50 trees, unrestricted depth
ann (Artificial Neural Network)	ReLU activation, two hidden layers of 50 neurons, learning rate = 0.01
gb (Gradient Boosting)	100 estimators, depth of 3, learning rate = 0.1
dt (Decision Tree)	Unlimited depth, minimum split = 5
nb (Naive Bayes)	No additional parameters needed
lr (Logistic Regression)	Penalty = L2, regularization strength C = 10

These were the best hyperparameters for each classifier through grid search optimization. These tuned parameters implies that the model was carefully adjusted for better performance.

Table 4 Classification Report Breakdown

Condition	Precision	Recall	F1-score	Support (cases)
Gestational Diabetes	0.51	0.95	0.67	21
No Diabetes	0.99	0.74	0.85	132
Type 1 Diabetes	0.06	1.00	0.12	1
Type 2 Diabetes	Nil	Nil	Nil	Nil

Precision values show the number of correct predicted cases and recall values indicate how many actual cases are identified. The harmonic mean of precision and recall is indicated by F1-Score. Nil implies that there were no sample of type 2 diabetes in the testing set. Accuracy of 0.77 means the model correctly classified 77% of the 154 total samples. Hybrid5 model saved “successfully” simply confirms that the trained model was stored properly and ready for reuse, deployment or further evaluation without retraining.

Table 5 Macro vs Weighted Averages

Average Type	Precision	Recall	F1-Score
Macro avg	0.52	0.90	0.54
Micro avg	0.92	0.77	0.82

Macro average treats all classes equally while weighted average considers class imbalance and heavily favors the class with large supports value, which is “No Diabetes” class. Since all classes are handled equally by Macro-average, recall value of 0.90 is used.

Table 6 Confusion Matrix

	Predicted		
	Gestational	No Diabetes	Type1
Actual Class Gestational	20	1	0
Actual Class No Diabetes	19	98	15
Actual Class Type1	0	0	1

The confusion matrix is a 3×3 matrix, each row represents the actual class, and each column represents the predicted class. Under gestational diabetes, true positives (TP) of 20 indicates the number of correct predictions as gestational class. False negatives (FN) of 1 predicted as no diabetes and 0 as class Type1. Thus, the accuracy for Gestational is very high, only one sample is misclassified. Under No diabetes class, true positives (TP) of 98 suggest the number of correct classifications. false negatives (FN) of 19 were misclassified as gestational and 15 as type1. Under Type1, true positives (TP) of 1 and no false negative implies that sample size for Type1 is small and may hinders evaluate performance. There is no prediction for Type2 and this can because of the absence of Type2 samples in the testing dataset. Hence there is need for further test with large volume of dataset.

VII. FURTHER RESULTS WITH SYNTHETIC DATASET

The model was also trained with 1.5 million synthetic datasets generated using python codes alongside with feature selection criteria. This was an act to test for the model’s generalization to unseen data. The results below were obtained

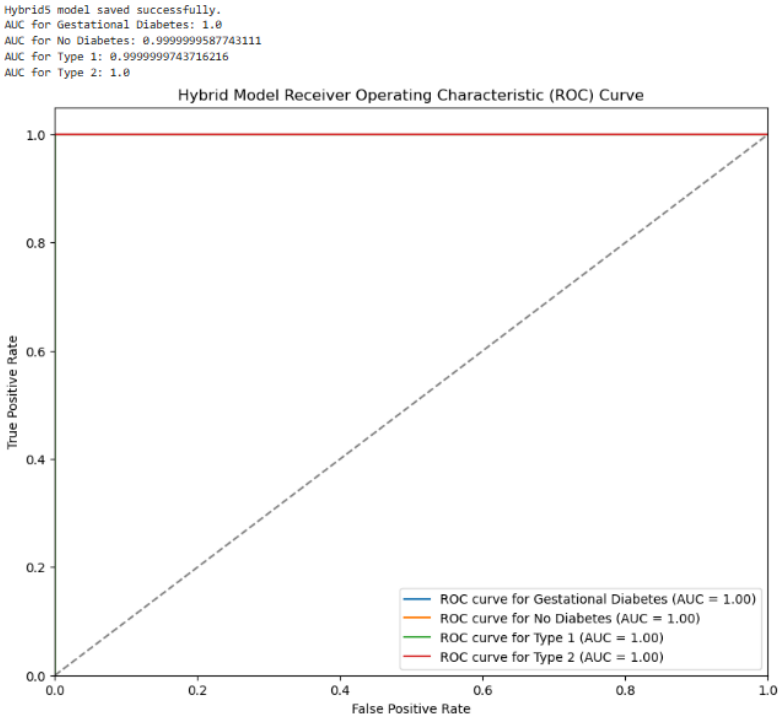


Fig. 6 Receiver Operating Characteristic (ROC) Curve with Synthetic Dataset

Table 7 Per-Class AUC Breakdown

Class	AUC	Interpretation
Gestational Diabetes	1.00	This particular class stood out crystal clear from the rest, with zero confusion.
No Diabetes	1.00	This flawless distinction means this class is perfectly separated with zero mix-ups.
Type 1 Diabetes	1.00	Full accuracy in identifying Type 1 cases
Type 2 Diabetes	1.00	Ideal performance in identifying Type 2 cases correctly.

```
Micro-average AUC: 1.00
Macro-average AUC: 1.00
Hybrid Model Accuracy: 0.9999966666666666
Hybrid Model AUC for Gestational Diabetes: 1.0
Hybrid Model AUC for No Diabetes: 0.9999999587743111
Hybrid Model AUC for Type 1: 0.9999999743716216
Hybrid Model AUC for Type 2: 1.0
                    precision    recall  f1-score   support

Gestational Diabetes       1.00      1.00      1.00     62448
         No Diabetes       1.00      1.00      1.00    115673
              Type 1       1.00      1.00      1.00     60424
              Type 2       1.00      1.00      1.00     61455

            accuracy                           1.00    300000
           macro avg       1.00      1.00      1.00    300000
        weighted avg       1.00      1.00      1.00    300000

[[ 62448      0      0      0]
 [     0 115673      0      0]
 [     0      1  60423      0]
 [     0      0      0  61455]]
```

Fig. 7 Evidence of the Results of Evaluation Metrics and Hyperparameter Tuning With Synthetic Dataset

Table 8 Overall Performance

Metrics	Value	Interpretation
Micro-average AUC	1.00	Area Under the Curve (AUC) was calculated by looking at all the categories as a whole. The perfect score means the model can clearly and accurately tell the difference between every classes.
Macro-average AUC	1.00	This implies the average AUC was calculated by giving equal weight to each class, showing the model is performing strongly and reliably across every type of diabetes.
Model Accuracy	1.00	This means every prediction made was spot-on, perfectly matching the actual labels with no misses and no errors.

Table 9 Classification Report Breakdown

Condition	Precision	Recall	F1-score	Support (cases)
Gestational Diabetes	1.00	1.00	1.00	62,448
No Diabetes	1.00	1.00	1.00	115,673
Type 1 Diabetes	1.00	1.00	1.00	60,424
Type 2 Diabetes	1.00	1.00	1.00	61,455

Under recall, all actual cases were identified without any omissions and precision values shows that every single prediction hit the mark with no false results. F1-Score also shows that there is optimal balance between precision and recall.

Table 10 Confusion Matrix

	Predicted			
	Gestational	No Diabetes	Type1	Type2
Actual Class Gestational	62448	0	0	0
Actual Class No Diabetes	0	115673	0	0
Actual Class Type1	0	0	60424	0
Actual Class Type2	0	0	0	61455

Each row in the matrix corresponds to the true labels of the data, while each column represents the predicted labels. Achieving diagonal dominance indicates that all predictions made are perfect, reflecting an ideal scenario where each predicted outcome aligns flawlessly with the actual results. There are absolutely no misclassifications between the different classes, meaning every instance has been accurately identified without any errors in categorization.

This hybrid model flawlessly predicted diabetes types in a sample size of 300,000 testing set. Such perfection suggests impeccable data quality and extremely effective feature selection. Since the model can identify the three classes of diabetes accurately without misclassifications using synthetic dataset of size 1.5 million, thus, it is recommended to incorporate real dataset of the size in further research in this domain.

VIII. CONCLUSIONS

Finally, the improved hybrid model for predicting diabetes is an improvement as it uses a voting system to blend various ML approaches. It has a recall of 0.90, making it extremely predictive and establishing a new standard for predicting diabetes. As an example of the predictive power added by intelligent pairing, the voting method is a nice illustration of the strengths of various algorithms combined, showing the value of ensemble modelling. The good performance of the model proves the scalability and adaptability of hybrid methods, paving the way for their generalization in the field of healthcare.

IX. REFERENCES

[1] WHO Diabetes Report, 14 November, 2024. [Online] Available https://www.who.int/news-room/fact-sheets/detail/diabetes/ (WHO 2023)

[2] Tabish, S. A. (2017). Lifestyle diseases: consequences, characteristics, causes and control. *J Cardiol Curr Res*, *9*(3), 00326.

[3] Casey, R., & Ballantyne, P. J. (2017). Diagnosed chronic health conditions among injured workers with permanent impairments and the general population. Journal of Occupational and Environmental Medicine, 59(5), 486-496.

[4] Pandeeswari, L., Rajeswari, K., & Phill, M. (2015). K-means clustering and Naïve Bayes Classifier for categorization of diabetes patients. *Eng Technol*, *2*(1), 179-185.

[5] Teju, V., Sowmya, K. V., Yuvanika, C., Saikumar, K., & Krishna, T. B. D. S. (2021, December). Detection of diabetes melittus, kidney disease with ML. In 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) (pp. 217-222). IEEE

[6] Butler, A. E., & Misselbrook, D. (2020). Distinguishing between type 1 and type 2 diabetes. *bmj*, *370*

[7] Balasubramanian, S., Kashyap, R., CVN, S. T., & Anuradha, M. (2020, December). Hybrid prediction model for type-2 diabetes with class imbalance. In *2020 IEEE international conference on machine learning and applied network technologies (ICMLANT)* (pp. 1-6). IEEE.

[8] Sai, M. J., Chettri, P., Panigrahi, R., Garg, A., Bhoi, A. K., & Barsocchi, P. (2023). An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes. *International Journal of Computational Intelligence Systems*, *16*(1), 14.

[9] Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, *10*(1-2), 1-10.

[10] Thakur D., Gera T., Bhardwaj V., AlZubi A. A., Ali F. & Singh A. (2023). An enhanced diabetes prediction amidst COVID-19 using ensemble models. *Frontiers in Public Health*

[11] Hasan, M. K., Alam, M. A., Roy, S., Dutta, A., Jawad, M. T., & Das, S. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked*, *27*, 100799.

[12] Jo, T.; Japkowicz, N. Class imbalances versus small disjuncts. ACM Sigkdd Explor. Newsl. 2004, 6, 40–49

[13] Wade, C., & Glynn, K. (2020). *Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python*. Packt Publishing Ltd.

[14] Aggarwal, C. C. (2018). *Neural networks and deep learning* (Vol. 10, No. 978, p. 3). Cham: springer.

[15] Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *arXiv preprint arXiv:1502.03774*.

[16] Kumar, A., Bawa, S., & Kumar, N. (2024). iDP: ML-driven diabetes prediction framework using deep-ensemble modelling. *Neural Computing and Applications*, *36*(5), 2525-2548.

[17] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

[18] Ojetunmibi, T., Asagba, P. O., & Okengwu, U. A. (2023). Pneumonia disease detection and classification system using naive Bayesian technique. *Scientia Africana*, *22*(1), 97-114.

[19] Rish, I. (2001, August). An empirical study of the Naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46).

[20] Denil, M., Matheson, D., & De Freitas, N. (2014, January). Narrowing the gap: Random Forests in theory and in practice. In *International conference on machine learning* (pp.665-673). PMLR.

[21] Yang, B., Di, X., & Han, T. (2014). Random forests classifier for machine

[22] Chung, D., Lee, C. G., & Yang, S. (2023). A hybrid machine learning model for demand forecasting: Combination of k-means elastic-net and gaussian process regression. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(6s), 325-336.