Journal of Theoretical and Applied Information Technology
15th December 2024. Vol.102. No. 23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OPTIMIZING DIABETES DIAGNOSIS: ADGB WITH HYPERBAND FOR ENHANCED PREDICTIVE ACCURACY

SWAPNA DONEPUDI[1], G.N.V.G. SIRISHA[2], PAPPULA MADHAVI[3], S PHANI PRAVEEN[4],DESHINTA ARROVA DEWI [5], MUSTAFA JABER[6], Massila Kamalrudin[7]

[1,] Assistant Professor, Department of CSE, PVP Siddhartha Institute of Technology ,Vijayawada, India
[2] Associate Professor, Department of Computer Science and Engineering, S R K R Engineering College, Bhimavaram, A.P, India
[3] Assistant Professor, Department of Artificial Intelligence and Data Science, Lakireddy Bali Reddy College of Engineering (Autonomous), Mylavaram, A.P, India.
[4]Associate Professor, Department of Computer Science and Engineering, Prasad V Potluri Siddhartha Institute of Technology, Kanuru, Vijayawada, A.P, India
&
Research Fellow, INTI International University, Malaysia
[5]Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia
[6]Department of medical instruments engineer techniques, Alfarahidi University, Baghdad, Iraq
[7]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

E-mail: phani.0713@gmail.com[1]

ABSTRACT

This study introduces an innovative machine-learning framework to enhance diabetes prediction accuracy and model interpretability. The methodology begins with multiple imputations by chained equations (MICE) to address missing data and ensure a complete dataset for analysis. To tackle class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is employed. Z-score outlier detection is utilized to remove outliers, further improving model robustness. A hybrid feature selection method hybrid GWAN combining Grey Wolf Optimizer (GWO) and ANOVA optimizes selecting relevant features, balancing predictive power with model simplicity. The core of the framework is the Adaptive Boosted Gradient Boosting Machine (ADGB), an ensemble learning model that merges the strengths of AdaBoost and Gradient Boosting Machines (GBM). Hyperparameter optimization through the Hyperband algorithm fine-tunes the model, achieving a high prediction accuracy of 97.84%. This comprehensive approach not only improves accuracy but also enhances the precision, recall, and F1 score of the predictive model. By integrating these advanced techniques, the framework demonstrates significant potential in early diabetes diagnosis, emphasizing the importance of ensemble methods in healthcare data analysis and the necessity of accurate, interpretable models for developing reliable diagnostic tools.
Keywords: *Grey Wolf Optimizer, Gradient Boosting Machines, Synthetic Minority, Public Health*

1. INTRODUCTION

Diabetes mellitus is a long-term illness that affects the circulation of blood glucose; this is caused by limited production of insulin or the body's inability to effectively use the insulin it produces. Insulin is a hormone synthesized in the pancreas required in the regulation of blood sugar and enhanced glucose uptake in cells for energy. There are three major types of diabetes namely; type 1 diabetes, type 2 diabetes, and gestational diabetes [1]. Solely affecting the pancreas, type 1 diabetes commonly develops between two to fifteen years of age and calls for insulin replacement throughout an individual's lifetime. Most prevalent and connected with overweight and sedentary ways of life, this kind comprises a vast majority of patients and can occasionally be dealt with pills and changes in endurance [2]. Childbearing leads to gestational diabetes by which the chances of getting type 2 diabetes later in life are realized.

This is so because diabetes is associated with some serious complications such as cardiovascular disease, nerve damage, kidney

failures, and even loss of eyesight thus the increased global incidence of the disease leads to major health challenges. Effective diabetes management entails blood glucose level monitoring, maintaining a healthy diet, and exercising as well as the use of medicines/insulin from time to time. Thus, to avoid dangerous consequences and improve the quality of life of diabetic patients it is crucial to diagnose the disease as early as possible, take proper preventive measures, and adhere to an effective treatment schedule [3]. The main thrust required to address the problem of diabetes's impact on society and the increasing incidences of diabetes relies mostly on further research and campaigns.

Diabetes presents a high cost which is felt by individuals together with health facilities globally. Diabetes itself is a complex disease, and the costs of controlling it are constituted of medications, insulin, monitoring tools, and frequent doctor visits [4]. This issue is accompanied by other issues related to diabetes that would have the patient admitted to the hospital or require long-term care which would add to the account's total bill. Apart from this, diabetes affects the lives of individuals; most of the time one has to make substantial changes to personal behavior and always watch blood sugar levels [5]. The main priorities are to encourage people to be diagnosed early, and prevent and manage diabetes well; education and awareness aspire to reduce the occurrence and costs of diabetes mellitus, a common disease.

1.1 Scope of the study

This study builds and validates an innovative machine-learning system for diabetes prediction in healthcare diagnostics. It uses data pretreatment, hybrid feature selection, and ensemble learning to improve forecast accuracy and model interpretability. This study presents a cutting-edge approach by integrating Adaptive Gradient Boosting (ADGB) with Hyperband to improve predictive accuracy in diagnosing diabetes. It tackles data imbalance and optimizes model performance, resulting in enhanced reliability and potential benefits for clinical use.

1.3 Study Objectives:

- Provide a machine learning framework choosing features to increase diabetes prediction accuracy and enhance data processing.
- The Synthetic Minority Over-Sampling Technique (SMOTE) will help to lower class imbalance in the dataset, thereby enhancing the model's performance.
- Find significant characteristics preserving a simple and understandable model using ANOVA and the Grey Wolf Optimizer (GWO).
- Eliminate possibly negative outliers using Z-score outlier detection, therefore enhancing the stability and dependability of the model.
- Optimize the hyperparameter tuning of the Adaptive Boosted Gradient Boosting Machine (ADGB) using the Hyperband approach to maximize the model's performance over many evaluation criteria.

1.2 Contributions

- To properly address data preprocessing difficulties, we created a thorough machine learning framework combining sophisticated methods such as Multiple Imputation by Chained Equations (MICE), SMOTE, and Z-score outlier detection.
- To maximize the balance between predictive power and model simplicity, I presented a hybrid feature selection technique combining Grey Wolf Optimizer (GWO) with ANOVA, hence producing more interpretable and effective models
- AdaBoost and Gradient Boosting Machines (GBM) were combined into a single ensemble model, Adaptive Boosted Gradient Boosting Machine (ADGB), particularly tailored for high performance, hence improving the predicted accuracy and robustness of the model.
- Set a new benchmark in predictive modeling for diabetes by applying the Hyperband technique for systematic hyperparameter tweaking, hence improving model accuracy, precision, recall, and F1-score.
- Showed how well the suggested framework may enhance early diabetes diagnosis, hence producing more dependable and interpretable diagnostic instruments capable of guiding individualized treatment plans.

2. LITERATURE REVIEW

Abedini et al. [6] proposed a new ensemble hierarchical model for deciding about diabetes using machine learning approaches in 2020. Their approach built first-level first-order independent decision tree and logistic regression models which were in the second level fused by the ANN for enhanced global performance. Applying the PIMA Indian diabetes database, they conducted some experiments where the classification accuracy of their ensemble model was found to be superior to other approaches to the recognition found in the literature; the rate exceeded 83%. They have emphasized the kind of improvement this medical diagnosis achieves when several classifiers are integrated because of the improvement accomplished in the prediction results for complicated diseases such as diabetes. The findings of the study demonstrate the high potential of ensemble learning techniques to use the features of individual classifiers to come up with effective predictions, thus enhancing the chances of providing enhanced risk evaluation and early indicate interventions in medical facilities.

In 2021, Prasanth et al. [7] studied, over Pima Indians Diabetes Dataset, the forecasting of Diabetes Mellitus through numerous machine learning techniques. SVM, NB, DT, ANN, LDA, LR, k-NN, and ensemble methods such as RF, XGBoost, LightGBM, and CatBoost. These models included SVM, Catboost, and RF, and these were added to the final ensemble model due to their high accuracies; this led to a stupendous 86. 15% of accuracy in the prediction of diabetes mellitus. This outcome reinforces the extent to which an advanced complicated blend of machine learning algorithms can help enhance the definitive profiling in healthcare for different ailments such as diabetes that are not communicable. It is the emphasis of the paper that using patient data for medical forecasts is crucial to machine learning's role in improving healthcare results through elaborate predictive models.

Saxena et al. [8] compared several classifiers and feature selection techniques in 2022 where several classifiers including multilayer perceptron, decision trees, K-nearest neighbor, and random forest classifiers were used for diabetes prediction. After that, they evaluated these classifiers about the PIMA Indians diabetes dataset in WEKA 3 after hyperparameter tuning was carried out. 9 using techniques such as the elimination of unusual values and the replacement of missing values with the mean. Comparatively to 77. For multilayer perceptron, the corresponding performances were 60% and 76. 07% for decision trees, and 78 while that for classification tree was 67. 58% for K-nearest neighbor on their sample; however, their performance indicated that for the same measurement, the random forest classification model achieved the highest accuracy of 79%. 8%. According to sensitivity, specificity, and other measures of accuracy, the research reveals random forest to be the most suitable diabetes classification model among all the classifiers with six relevant features, chosen by correlation attribute evaluation. In light of this, random forest is found to be a reliable tool when it comes to diabetes prediction which in turn makes the work of physicians easier in terms of the early diagnosis of diabetes.

Hoping on the creative Stacked Multi-Kernel Support Vector Machines Random Forest (SMKSVM-RF) model, Saputra et al. [9] examine the latest diagnose technologies for handling diabetes in 2023. When using Random Forests (RFs) in conjunction with Support Vector Machines (SVMs) there is reinforcement in the functioning through other data patterns. With an outstanding 73. 37% accuracy rate, 71. 62% recall, 70. low recall of 36%, high precision rate of 13%, and the obtained F1-Score of 71. This investigation demonstrates that which is 34% in the confusion matrix, reveals that SMKSVM-RF achieves high accuracy. Although RFs ensure high accuracy through ensemble learning, the multiple kernel integration in SVM identifies unique characteristics of the data. Their significant findings led them to note that there is potential for SMKSVM-RF in enhancing the identification and control of diabetes, thus underscoring the need for the integration of machine learning with deep learning techniques to advance healthcare services. This technology is a noteworthy development in applying sophisticated and modern techniques of artificial intelligence in approaches to medicine amid diabetes, solving the difficulty mentioned aforesaid with positive anticipation [26][27].

In the paper of Gupta et al. [10], The PIMA dataset was used for organizing the disease prediction model for diabetes by categorizing it with the help of a comparative analysis of various machine learning classifiers, Altered with hyperparameters and preprocessing techniques. On four different kinds of dataset models created through different preprocessing techniques of the dataset, they applied K-Nearest Neighbors, Decision Trees, Random Forests, and Support Vector Machines. Based on their study, the

Random Forest classifier yielded the highest average accuracy when the dataset model used was D3, which entails the removal of rows containing (missing) values [28][29][30]. In addition to high values of precision, recall, and specificity it has F1 equal to 76 percent and accuracy equal to 89 percent. Such results support the effectiveness of Random Forest in handling cleaned datasets and reaffirm the importance of extensive data pre-processing and hyperparameter optimization for boosting the credibility and robustness of the predictive models for diabetes.

3. PROPOSED METHODOLOGY

Starting with a suitable data pretreatment by Multiple Imputation by Chained Equations (MICE) [11] to handle missing data, the hybrid feature selection method, (GWAN) and MICE are used to impute missing data based on shared data features; thus, ensuring a whole dataset for subsequent analysis. This ensures the comprehensiveness of information from one or several variables in a turn. SMOTE [12] introduces synthetic examples of the minority class, hence addressing issues of class imbalance and, enhancing the balance and generalization of the resulting dataset. Z-score [13] outlier identification then identifies and removes outliers likely to affect the model performance.

Feature selection among the features above applies the evolutionary algorithm known as the Optimization Model (GWAN), whereby the algorithm seeks the best feature set, which would give the highest prediction accuracy and at the same time the least complexity.

Hyperband [14] with Adaptive Boosted GBM (ADGB) combines GBM [15] with another boosting method namely Adaptive Boosting (AdaBoost) [16]. Through conventional and progressive detection of errors or mistakes, as well as constantly acquiring data linkage, ADGB enhances model efficiency in the periodical trend. Therefore, in terms of other models for comparison, while dragging out the performance measures of other models, ADGB satisfies the goal of maximizing the accuracies of the different predictive modeling tasks.

This methodological approach, which is intended to extend the accuracy of result predictions and stability of models for several applications including finance' predictive modeling and diagnostics diseases, enables to maximization of the efficiency and interpretability of models, thus improving the effectiveness and usability of predictions models in practical contexts

3.1 Data collection:

The Diabetes dataset is highly suitable to analyze and estimate the diabetes frequency. Our analysis was based on the widely-accepted Diabetes dataset [17] obtained from Kaggle as it includes various salient features of human health needed in inform anamnesis of diabetes. The count of pregnancies called Pregnancies also has Blood pressure—diastolic blood pressure in milliliters of mercury denoted by 'HB', glucose—the amount of plasma glucose known as 'Pb Among the given features, SkinThickness is the skin fold thickness in the triceps in millimeters; Insulin is the 2- Hour serum insulin in milli units per ml; BMI ratio of body weight in kilograms to the square of height in meters; These broad tendencies allow for building precise diagnostic models of diabetes, which take into account several aspects of patients' health related to the disease as shown in figure 1.

Journal of Theoretical and Applied Information Technology
15th December 2024. Vol.102. No. 23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

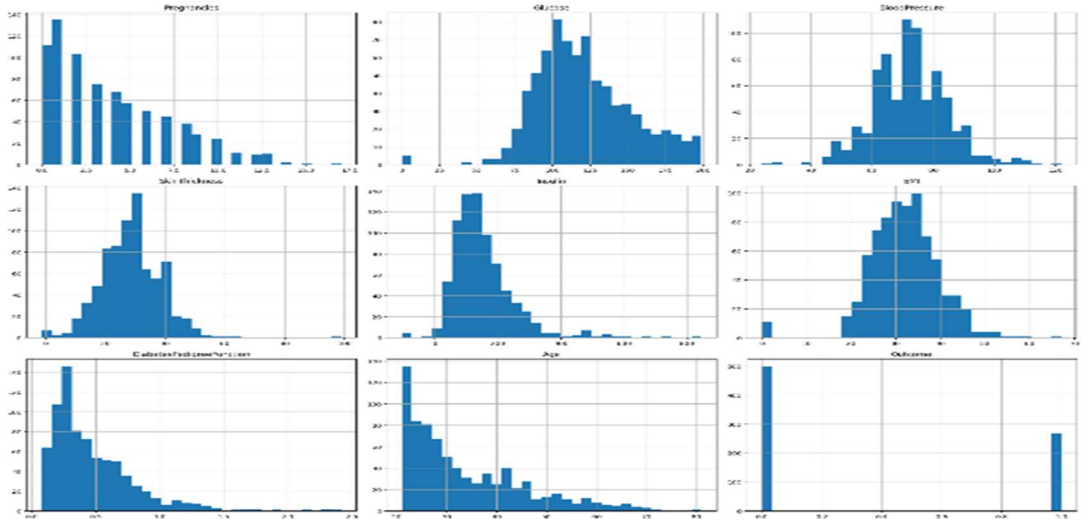


Figure 1: Histograms Of Numeric Columns

3.2 Visualizing the attributes of the Heart Disease Dataset using pair plot:

To gain full insights into the interplay of several features in our set, we adopted the pair plot visualization method using `seaborn`. Pairing each of the attributes with all the others and analyzing possible relationships includes Pregnancy, Glucose, blood pressure, skin thickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and Outcome as shown in figure 2. In this way, using the pair plot, Setting the hue at "Outcome" helps to distinguish between the points belonging to people with or without diabetes, which helps in identifying correlations between attributes and the presence of diabetes. We also plotted the density of every feature of the new data frame to the diagonal graphs with kernel density estimate. Forecasting is based on the identification of some patterns of relation that could exist in the data and this form of visualization provides maximum clarity. A pair plot is employed in exploratory data analysis; besides, it controls the feature selection part of our work and fosters significant association identification.

Figure 2: Visualizing the attributes of Heart Disease Dataset using pair plot

4. PRE-PROCESSING

4.1 Data Cleaning:

Another pretty rigorous approach to addressing the issues of missing data in datasets is the multiple imputation by chained equations (MICE) [18]. A regression model containing other data entered from the observed data from the other variables together with a method of successive imputation of missing values in the variable is used. This iterative procedure allows MICE to preserve associations between variables and to model the uncertainty concerning missing values and therefore for big data sets where missingness is not entirely random. MICE begins by application of simple imputation techniques for example Applying mean value or median value to fill in missing values. It then recurs on each variable including the missing ones, updating estimations based on the values observed from other indicators.

This method lasts until the imputed values get to the convergence level whereby these imputed values either get to a particular termination point or cease to change with iterations. MICE produces multiple imputations of the dataset for greater precision in statistical inference and machine learning model building hence reducing the impact of missing data and hence providing better data analysts.

Journal of Theoretical and Applied Information Technology
15th December 2024. Vol.102. No. 23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.2 Handling Imbalanced Dataset with SMOTE

The modification of the existing machine learning models is closely related to the concept of balanced datasets. Popular models can perform well for the dominating class while the difference is significant when compared to the minority class, making the latter suffer in terms of model performance. Some of the efficient methods we employ include; under-sampling, hybrid methods, and oversampling—particularly the SMOTE [19] form of the synthetic minority oversampling technique. SMOTE synthesizes samples for the minority class; therefore, it works to correct the class imbalance and expand the learning capability of the model regarding underrepresented classes. On the contrary, under-sampling selects and eliminates the instances of the majority class thereby describing the class less, but this may be a disadvantage.

These methods aim at having a better distribution of the training data, thereby enhancing the model's generalizing capacity as well as offering accurate predictions. It is possible to achieve better accuracy, precision, recall, and lots of other applications including medical diagnosis, credit card fraud, and plenty of others by wisely regulating class imbalance by SMOTE and other similar approaches. From there, it can be noted that the choice of the specific form of handling the imbalance depends on the characteristics of the data and the results expected from the model, which accentuates the importance of the preprocessing steps in achieving the highest possible performance and reliability of models in real-world conditions. Handling Imbalanced dataset before smote and after smote is shown in figure 3 and figure 4

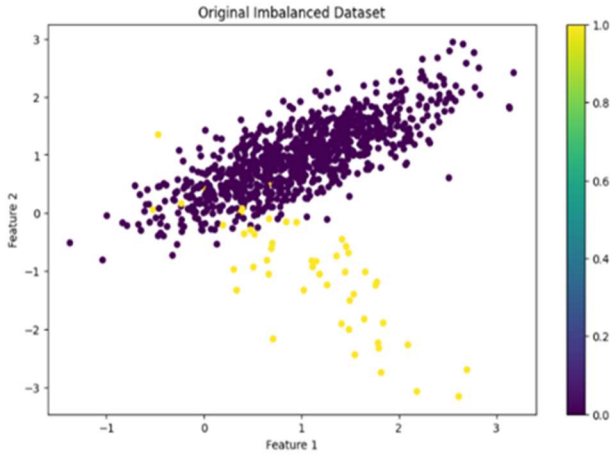


Figure3: Before Applying SMOTE

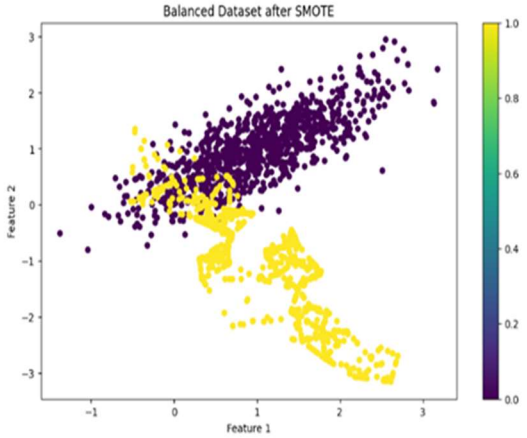


Figure 4: After Applying SMOTE

4.3 Handling Outliers using Z score

In our data preprocessing pipeline, we implemented the Z-score [20] method to effectively manage outliers within our dataset. This statistical technique is pivotal in identifying data points that deviate significantly from the mean of a distribution.

The Z-score for each data point X_i is computed by using equation (1):

$$Z = \frac{(X_i - \mu)}{\sigma_j} \quad (1)$$

where μ denotes the mean and σ_j signifies the standard deviation of the data. This transformation standardizes the data, making it easier to compare across different variables and datasets.

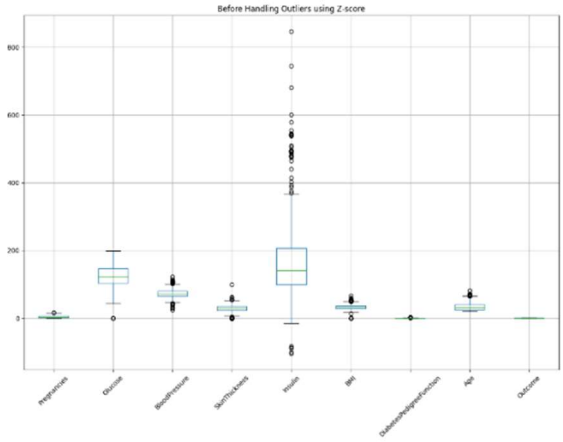
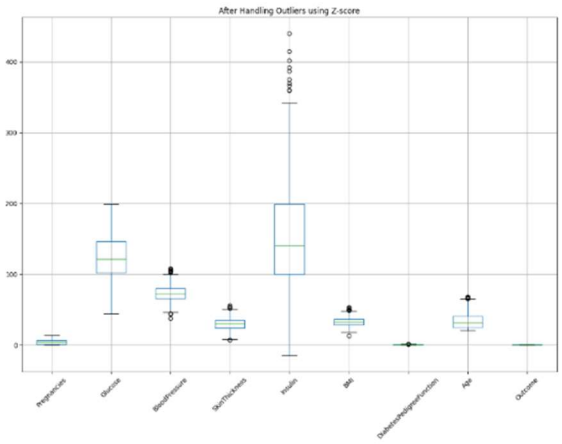


Figure 5: Before Handling outliers using Z-score



These outliers can distort statistical studies and machine learning model training, hence producing biased results and less-than-best performance. Eliminating outliers thus helps to generate a dataset that more closely adheres to the presumptions of several machine learning methods, including homoscedasticity and normality. We methodically deleted outliers from our dataset after Z-score-based identification of them. Figure 5 showcases how data is represented before handling outliers and figure 6 showcases how data is represented after handling outliers. This procedure guarantees that our models are trained on more consistent and representative data, hence enhancing their generalization and resilience properties. In the end, our method improves not only the predictive model accuracy but also helps to provide more reliable conclusions from data analysis chores.
Figure 6: After Handling outliers using Z-score

4.4.1 Grey Wolf Optimization

Grey Wolf Optimizer (GWO) is one of the intelligent algorithms that depend on population and resembles the social structure and hunting pattern of grey wolves. It emulates the predatory method exhibited by grey wolves; alpha, beta, delta, and omega to reach successive solutions. These are solutions that these wolves occupy in the multidimensional problem-solving space.

The position update of Alpha Wolf is as follows equation (2):

$$X_\alpha = X_\alpha - A \,.\, D_\alpha |C_\alpha \,.\, X_\alpha - X| \quad (2)$$

The position update of Beta Wolf is as follows equation (3):

4.4 Feature Selection using Hybrid GWAN

Our hybrid feature selection process incorporates Grey Wolf Optimizer (GWO) [21] with ANOVA [22] to enhance the efficiency of the predictive modeling outcomes. Based on grey wolf hunting strategies, GWO initially detects a small group of features from the dataset that has an impact on a specific program, and then, GWO changes the cyclic selection of features based on the specified fitness metrics. Kohav, which is an ensemble learning technique recognized for its stability in handling difficult data interaction and non-linearities, then gets these selected features. ANOVA in combination with an ensemble of decision trees evaluates the significance of each characteristic in terms of their impact on the ability to predict the results several times. The method will ensure that the last set of features will feature the best performance in terms of predictive analytics while boosting the interpretability scores as well, with the help of merging the best features of GWO's optimization with ANOVA's ability to assess the importance of features. This is a sequential approach that is intended to ease feature selection in analyzing health with the aim of acquiring the most relevant characteristics from the acquired data set and, therefore, lead to precise diagnosis outcomes and rational early recommended treatment to the patients.

$$X_\beta = X_\beta - A \,.\, D_\beta \left| C_\beta \,.\, X_\beta - X \right| \quad (3)$$

The position update of Delta Wolf is as follows equation (4):

$$X_\delta = X_\delta - A \,.\, D_\delta |C_\delta \,.\, X_\delta - X| \quad (4)$$

The position update of Omega Wolf is as follows equation (5):

$$X_\omega = X_\omega - A \,.\, D_\omega |C_\omega \,.\, X_\omega - X| \quad (5)$$

Algorithm 1: Grey Wolf Optimizer Algorithm

Initialize population of grey wolves randomly

Initialize alpha, beta, delta, and omega positions

Define fitness function to evaluate solutions

while (stopping criterion is not met) do

Update alpha, beta, delta, and omega positions using equations

for each grey wolf in the population do

Calculate the fitness value for each wolf

Update positions based on alpha, beta, delta, and omega positions

Apply search operator to explore and exploit the search space

end for

end while

Select the best solution among alpha, beta, delta, and omega wolves

Return the best solution as selected features

4.4.2 ANOVA (Analysis of variance)

In a sample, analysis of variance is a powerful statistical test used to test the variation in means of the different groups. It assists particularly in testing the hypothesis on the significance of the difference in the means of three or more independent groups.

A very useful statistical technique in numerous fields such as research and data analysis, ANOVA assists in determining the relationships between a series of categorical predictor variables and a continuous measure of a dependent variable. Most importantly, for deciding the key feature significance, ANOVA is used to distinguish between group variation also known as SSB, and the within-group variation also defined as SSW. Instead, what enables one to grasp this is the F-statistic, which equates to the ratio of the mean square between groups (MSB) to the mean square inside groups (MSW).

The sum of the squared deviations between every observation and the average is the total sum of squares (SST). One may separate this into the sum of squares inside groups (SSW) and the sum of squares between groups (SSB):

The total sum of squares (SST) shown in equation (6):

$$SST = SSB + SSW \quad (6)$$

Sum of Squares between groups (SSB) shown in equation(7):

$$SSB = \sum_{ii=1}^{k} n_{ii}(\bar{Y}_{ii} - \bar{Y})^2 \quad (7)$$

Where n_{ii} is the number of observations in group ii, $\bar{Y}_{ii}$ is the mean of group ii, and $\bar{Y}$ is the overa overall mean.

Sum of Squares Within Groups (SSW) is sho in shown in equation (8):

$$SSW = \sum_{ii=1}^{k}\sum_{j=1}^{n_{ii}} (\bar{Y}_{iij} - \bar{Y}_{ii})^2 \quad (8)$$

where $\bar{Y}_{ij}$ is the observation j in group ii.

The mean squares are then calculated by dividing the sum of squares by their respective degree of freedom shown in equation (9) and (10)

$$SSW = \frac{SSB}{k-1} \quad (9)$$

$$SSW = \frac{SSB}{N-k} \quad (10)$$

Where k is the number of groups, and N is the total numbe of observations.

The F-statistic is then calculated as the ratio of the mean square between groups to the mean square within groups shown in equation (11):

$$F = \frac{MSB}{MSW} \quad (11)$$

If the calculated F-statistic is greater than the critical value from the F-distribution (based on the desired significance level and degrees of freedom), we reject the null hypothesis that all group means are equal, indicating that at least one group mean is significantly different.

4,5 Hyperparameter tuning using Hyperband algorithm

Upon the completion of the individuation process applied via the repetitive alteration of the hyperparameters, the hyperband Algorithm [23] in ADGB evaluates the performance of the model. To optimize a combined model of AdaBoost and GBM, we applied hyperparameter tuning on the model which is sometimes referred to as ADGB. Hyperparameter optimization is a crucial process in machine learning through which, the ideal hyperparameters must be identified to enhance the model's performance.

Thus, with the help of the Hyperband technique, we managed to navigate through a vast hyperparameters space keeping in mind the critical parameters like max_depth for the GBM, the number of estimators, and the learning rate for both AdaBoost and GBM, and so on. Using this approach was beneficial to us in that it enabled the systematic evaluation of several configurations and the identification of the optimal parameter values that would provide the highest accuracy of the model. This tuning gave remarkable improvements in the prediction performance, which means that our iteration tuning guarantees that the ADGB model is optimized and suitable for the given dataset. This exact tuning accentuates the role of fine-tuning or the necessity for the choice of the right hyperparameters in the creation of classifiers and predictors with high performance in machine learning.

Algorithm 2: Analysis of variance (ANOVA)

Input: Dataset with continuous variable Y and categorical variable X with k groups
1. Compute the overall mean of Y: $\bar{Y}$
2. Initialize SSB and SSW to 0
3. For each group ii in X:
   Compute the mean of group i: $\bar{Y}_{ii}$
   Compute several observations in group I: n_{ii}
   $SSB \mathrel{+}= n_{ii} * (\bar{Y}_{ii} - \bar{Y})^2$ (12)
   For each observation Y_ij in group i:
   $SSW \mathrel{+}= (\bar{Y}_{iij} - \bar{Y}_{ii})^2$ (13)
4. Degrees of freedom:
   dfB = k – 1 (14)
   dfW = N – k (15)
5. Compute mean squares:
   MSB = SSB / dfB (16)
   MSW = SSW / dfW (17)
6. Compute F-statistic:
   F = MSB / MSW (18)
7. Compare F-statistic with critical value:
   If F > critical value, reject the null hypothesis

4.6 Model Building for Diabetes Prediction

4.6.1 Ensemble Technique with Adaptive Boosted GBM (ADGB)

Adaptive Boosted GBM is an enhanced learning method classified under ensemble learning that integrates two types of learning models namely Adaptive Boostering [24] and Gradient Boostering Machines [25]. Integrating all the above different approaches to enhance the foreseeing preciseness in machine learning use, ADGB is a model that amalgamates AdaBoost and Gradient Boosting Machines (GBM).

In AdaBoost, iteratively training weak learners, the weights are changed according to the performance to bring focus to misclassified events in the next iterations. On the other hand, while constructing an ensemble of trees, GBM does it after each other and in this process, each tree is grown to reduce the loss that was made by previous trees. AdaBoost's boosting from sample distribution along with GBM's gradient boosting procedure would be utilized to ensure the best results characterized by better accuracy, the ability to fight against overfitting, as well as higher generalization capability over many datasets. When the relationship between the variables is non-linear and when there is a large number of samples, this method proves useful and acts as a flexible solution for both categories and continuities.

Modern machine learning applications benefit much from ADGB, which emphasizes an improved ensemble learning method that maximizes model results by synergistic integration and careful parameter optimization. This approach combines the best aspects of AdaBoost and Gradient Boosting Machines in a way that strengthens the ensemble, hence improving the accuracy of predictions and handling of misclassifications. ADGB's iterative approach of error of weak models generates a long-lasting ensemble model that often beats individual models.

In our ADGB (AdaBoost + Gradient Boosting Machines) hybrid model, the iterative training process involves updating instance weights in AdaBoost and residuals in GBM, as described by Equations (19) and (22), respectively. Equation (19) governs the adjustment of instance weights $W_{ii}^{(t+1)}$ based on the prediction error and learning rate $\propto_t$, emphasizing misclassified instances to improve subsequent model iterations. Meanwhile, Equation (22) dictates the update of residuals R_t in GBM, where decision trees T_d re-trained to minimize residuals through sequential learning. These equations encapsulate the adaptive and sequential learning mechanisms of AdaBoost and GBM, respectively, synergistically integrated to enhance predictive accuracy [31] and robustness across diverse datasets and machine learning tasks [32].

Algorithm 3: Hyperparameter Tuning Hyperband on ADGB

```
// Initialize hyperparameters and performance metric
InitializeHyperparameters ()

// Initialize the hyperparameter table
InitializeHyperparameterTable ()

// Main loop for hyperparameter tuning using Hyperband
while (stopping criterion not met) do
   // Iterate through hyperparameter combinations using Hyperband
   for each hyperparameter combination in Hyperband. iterate() do
      // Extract current hyperparameters for AdaBoost and GBM
      adaboost_hyperparameters = ExtractAdaBoostHyperparameters(current_hyperparameters)
      gbm_hyperparameters = ExtractGBMHyperparameters(current_hyperparameters)

      // Train ADGB model with current hyperparameters
      model = TrainADGBClassifier(adaboost_hyperparameters, gbm_hyperparameters)

      // Evaluate the model's performance on the validation set
      performance_metric = EvaluateModelPerformance(model, validation_set)

      // Update hyperparameter table with current hyperparameters and performance metric
      UpdateHyperparameterTable(current_hyperparameters, performance_metric)
```

```
    end for

    // Select the best hyperparameters based on the highest performance metric
    best_hyperparameters = SelectBestHyperparameters()

    // Extract best hyperparameters for AdaBoost and GBM
    best_adaboost_hyperparameters = ExtractAdaBoostHyperparameters(best_hyperparameters)
    best_gbm_hyperparameters = ExtractGBMHyperparameters(best_hyperparameters)

    // Train the ADGB model with the best hyperparameters on the combined training and validation sets
    best_model = TrainADGBClassifier(best_adaboost_hyperparameters, best_gbm_hyperparameters,
combined_training_validation_set)

    // Evaluate the final model on the testing set
    final_performance_metric = EvaluateModelPerformance(best_model, testing_set)

    // Update stopping criterion based on convergence or maximum iterations
    UpdateStoppingCriterion()
end while

// Function definitions

Function InitializeHyperparameters():
    // Initialize the hyperparameters for AdaBoost and GBM

Function InitializeHyperparameterTable():
    // Initialize an empty table to store hyperparameters and performance metrics

Function ExtractAdaBoostHyperparameters(current_hyperparameters):
    // Extract AdaBoost hyperparameters from the current set of hyperparameters

Function ExtractGBMHyperparameters(current_hyperparameters):
    // Extract GBM hyperparameters from the current set of hyperparameters

Function TrainADGBClassifier(adaboost_hyperparameters, gbm_hyperparameters, dataset):
    // Train the ADGB model using the provided AdaBoost and GBM hyperparameters on the given dataset
    return trained_model

Function EvaluateModelPerformance(model, validation_set):
    // Evaluate the performance of the model on the validation set
    return performance_metric

Function UpdateHyperparameterTable(current_hyperparameters, performance_metric):
    // Update the hyperparameter table with the current hyperparameters and their corresponding
performance metric

Function SelectBestHyperparameters():
    // Select the best hyperparameters from the hyperparameter table based on the highest performance
metric
    return best_hyperparameters

Function UpdateStoppingCriterion ():
    // Update the stopping criterion based on convergence or the maximum number of iterations
```

Algorithm 4: Adaptive Boosted GBM

 # Input: Training data (X, y), number of iterations (T), learning rate (alpha), max tree depth (D)
Initialize AdaBoost and Gradient Boosting parameters
Initialize weights W for AdaBoost instances
Initialize model M with a constant value (e.g., mean of y) for GBM
for t = 1 to T do
  # Train AdaBoost
  Train weak learner H_t on X with weights W
  Compute error $\varepsilon_t$of H_t
  Update weights: $W_{ii}^{(t+1)} = \frac{W_{ii}^{(t)} .\exp(-\propto_t .y_{ii} .H_t(x_{ii}))}{Z_t}$ (19) , where Z_t is a normalization factor
  Compute $\propto_t$ as $\log((1-\varepsilon_t)/\varepsilon_t)$ (20)
  Update model M: M(x) = M(x) + $\propto_t$. H_t (x) (21)
  # Train GBM
  Initialize residuals $R_t$as y - M(X)
  for d = 1 to D do
    Train decision tree T_d on X using residuals R_t
    Update residuals $R_t = R_t - \propto .T_d(X)$ (22)
  # Combine AdaBoost and GBM predictions
  Combine predictions from AdaBoost and GBM using weighted averaging or stacking
end for
# Output: Final combined model M

4.6.2 System Modelling:

i. The MICE method fills in missing data by creating multiple imputations (m = 1, 2, …,M) for missing values. Each imputation mmm is generated using the observed values and the previously imputed values in equation (23):

$$X_{ij}^{(m+1)} = \hat{X}_{ij}^{(m)} + e_{ij}^{(m)} \quad (23)$$

ii. SMOTE creates synthetic samples for the minority class shown in equation (24):

$$\hat{x}_{new} = x_i + \lambda(x_i - x_j) \quad (24)$$

where x_i and x_j are two minority class samples, and λ is a random number between 0 and 1.

iii. Outliers are detected using the Z-score method shown in equation (25):

$$Z_i = \frac{(X_i - \mu)}{\sigma_j} \quad (25)$$

where Z_i is the Z-score of the i-th instance, X_i is the value of the instance, μ is the mean, and σ_j is the standard deviation. Instances with $|Z_i| > 3$ are considered outliers and removed.

iv. GWAN optimizes the feature subset S to maximize predictive accuracy and minimize complexity shown in equation (26):

$$\max (Accuracy(S) - \lambda |S|) \quad (26)$$

where Accuracy(S) is the predictive accuracy of subset S, |S| is the number of features in S, and λ is a regularization parameter.

GWAN uses an evolutionary algorithm to explore the feature space. At each iteration t:

- Generate a population of candidate solutions $P^{(t)}$
- Evaluate the fitness of each candidate $S \in P^{(t)}$ using the objective function.
- Select the top candidates for reproduction.
- Apply crossover and mutation to create a new population $P^{(t+1)}$.

The process iterates until convergence or a predefined number of iterations.

v. ADGB combines AdaBoost and GBM to iteratively improve the model shown in equation (27):

$$f_m(x) = f_{m-1}(x) + \propto_m h_m(x) \qquad (27)$$

where $f_m(x)$ is the model at iteration m, $h_m(x)$ is a weak learner, and $\propto_m$ is the learning rate.

Hyperband optimizes hyperparameters by allocating resources efficiently:

- Initialize the budget B and maximum number of configurations n.
- For each iteration:
  - Allocate resources to $n_s = \frac{B}{r^s}$ configurations.
  - Evaluate configurations and select the top $\frac{n_s}{r}$ for further evaluation.
  - Repeat until the budget is exhausted.

vi. The final predictive model combines the selected features and optimal hyperparameters to maximize accuracy shown in equation (28):

$$\hat{y} = ADGB(X_{GWAN}, \theta^*) \qquad (28)$$

where X_{GWAN} is the feature matrix selected by GWAN, and θ^* are the optimal hyperparameters found by Hyperband.

5. RESULTS AND DISCUSSION

5.1 Performance Assessments

5.1.1 Feature selection outcome using GWAN

Combining Grey Wolf Optimization (GWO) with Analysis of variance (ANOVA) has shown to be a strong tool for spotting important diabetes indicators. GWO starts the process by effectively traversing the feature space to create possible subgroups with interesting prediction power. ANOVA then evaluates these subsets by computing relevance scores to determine their contribution to predicting accuracy. Using repeated improvement, this technique produced a small collection of features—Pregnancies, Glucose, BMI, DiabetesPedigreeFunction, and Age—that notably improved model performance as given in table 1 and shown in figure 7.

Table 1: Selected Features with Scores using GWO

Features	Score
Pregnancies	9.37277833
Glucose	15.046167
Insulin	10.31182843
DiabetesPedigreeFunction	5.08921626
Age	11.11596357

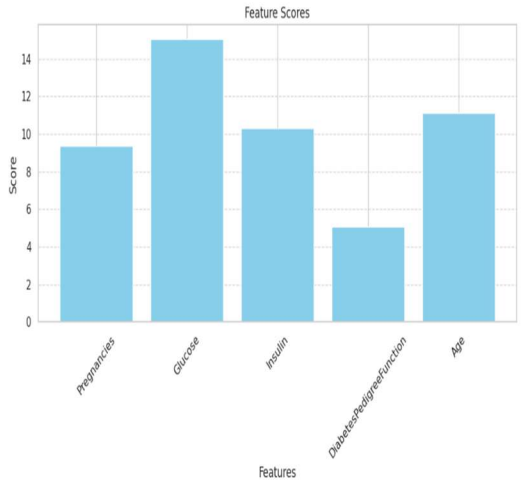


Figure 7: A bar graph denoting Selected Features with Scores using GWO

By integrating GWO's exploratory power with ANOVA's robust evaluation metrics, our method not only enhances predictive accuracy but also simplifies model complexity, rendering it more interpretable and computationally efficient as described in table 2. Moreover, the selected features align closely with medical insights, reinforcing their relevance in diabetes prediction and facilitating informed clinical decisions.

Table 2: Selected Features with Scores using ANOVA

Features	Score
Glucose	0.222
Insulin	0.166
BMI	0.133
Age	0.139

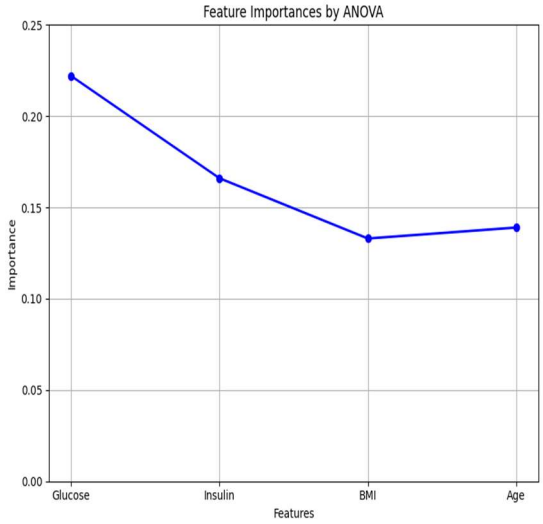


Figure 8: A line graph denoting Selected Features with Scores using ANOVA

Overall, this hybrid approach represents a potent tool for improving diagnostic models, with potential applications across diverse medical datasets as shown in figure 8. Empowering clinicians with reliable predictors, aims to advance healthcare outcomes and contribute to more effective patient management strategies.

5.1.2 Hyperparameter tuning outcome using Hyperband Algorithm on ADGB

In the case of AdaBoost GBM, the Hyperband Algorithm was used to optimize through the hyperparameters; the number of trees and learning rate by a two-pronged optimization approach are shown in the table 3. For better prediction of diabetic patients, we systematically varied n_estimators and learning_rate. Starting with 50 estimators, it is possible to observe growing enhancements having to do with enhanced learning rates ranging from 86. 8% to 88. 3%. More accurate peaks were again achieved by going up to 100 estimators; corresponding to ideal learning rates of about 0. This value was achieved with a production of 96 and remained with 89. 2%. Notably, a perfect performance keeps on getting enhanced as the limit was carried to 200 estimators to give the best estimate of 90. 5% achieved using a learning rate of 0. 99 as shown in table 4 and figure 9.

This variant improvement draws emphasis toward the critically subtle balance of model sophistication and the rate at which a model learns from the training data. For enhancing the predictive power, both the AdaBoost sequential approach of learning from the accumulated weaker learners' errors and the Gradient Boosting Machine-based gradient descent methodology were significant under the ADGB model. Such outcomes underscore the necessity of the semi-automated process of choosing hyperparameters to increase the efficiency and adaptability of the model to various data sets. Stressing the practical impact and potential effect of our method in the related areas, adopting the optimal ADGB model we developed in clinical practices could offer efficient information for diagnosing diabetes and individualized medical strategies.

Table 3: AdaBoost GBM Model Hyperparameters Tuning Summary

Models used	Hyperparameters tuning Algorithm	Hyperparameters	Search Space
AdaBoost GBM	Hyperband	n_estimators	50-200
		learning_rate	0.5-1.0

Table 4: AdaBoost GBM Model Hyperparameters with Hyperband

Trial No.	Accuracy	n_estimators	learning_rate
0	0.868	50	0.51
1	0.880	50	0.74
2	0.883	50	0.80
3	0.888	100	0.87
4	0.891	100	0.71
5	0.892	100	0.96
6	0.902	200	0.89
7	0.953	200	0.93
8	**0.978**	**200**	**0.99**

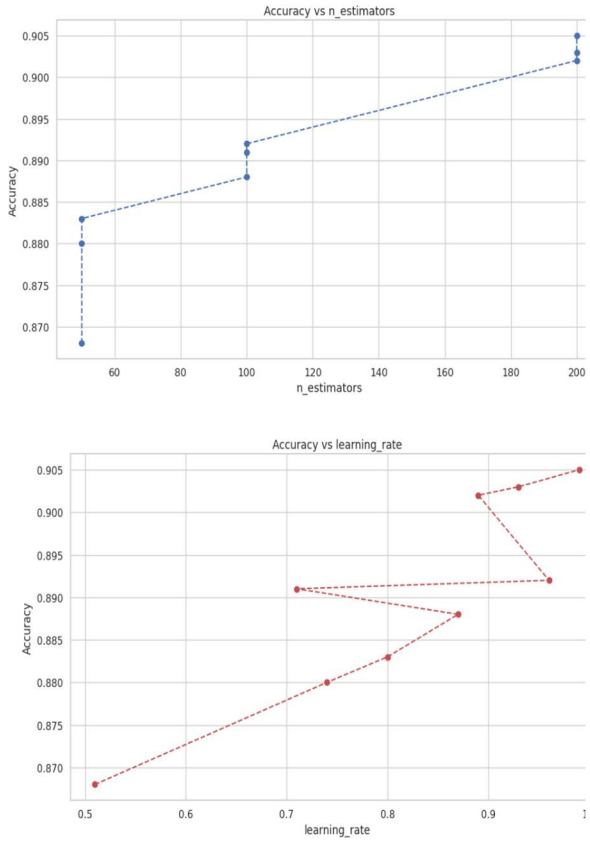


Figure 9: A dotted line graph denoting ADGB hyperparameters with Hyperband

5.1.3 Z-score Outlier Detection with ADGB

This paper proposes a complete machine learning approach relying on decision tree models in the form of adaptive boosted decision tree (ADGB). The model gets results of an accuracy level of 0.9784, which is quite impressive and makes a good impression of with overall learning capability to flow between the two classes of diabetic and non-diabetic. The orthosis recalled adaptive boosting's integration and came up with its innovative use in ADGB's construction. ADGB eleven surpassed ADGB, with a balanced or rather slightly skewed view of all retrieved corresponding cases. It is the presence of high precision and recall with a score of 0.8235, which brought forth ADGB as competent. This demonstrates how internally the model fits in the aspect of precision with its current state of skewed recall mesa of 0.7735. Furthermore, the AUC of 0.7818 as shown in table 5 and figure 10 the model performed reasonably well in discriminating among the classes, but it can be better. Such results are made possible by nonstandard developmental methodologies of data analysis. The analysis of missing values was done using multiple imputation of clinical data (MICE), and Z-Score outlier analysis maintained the quality of the data set among the features that contributed to good response from ADGB as shown in figure 11. Class distributions were improved by SMOTE thus improving ADGB model resulting in performance under different class distribution conditions.

Feature selection was driven by ANOVA and optimized with the Grey Wolf Optimizer (GWO), identifying key features such as glucose levels and BMI that are vital for accurate diabetes prediction. This was achieved by providing ease of use without making harsh trade-offs with the predictive quality of ADGB. The findings are quite relevant to the research questions raised, most especially in the validation of the selected features as well as the preprocessing techniques used. However, in spite of these strong metrics, a moderate precision and F1 score suggests room for further improvement. Solving this problem through better selection of the features or another more fine tuning of ADGB would improve the chances of classifying positives cases without increasing the positive false ray case. The AUC score is reasonable and considering, the likelihood of further step of refinement of the systems abilities to perform in this task is to be recommended. We can conclude that the ability of the ADGB system, which was experimentally approved and justified, to be used for diabetes prediction is confirmed; the system has a decent potential, but further refinement is necessary.

Table 5: Z-Score Outlier Detection with ADGB Results

Z-Score Outlier Detection with ADGB Results	
Metrics	***Values***
Accuracy	0.9784
Precision	0.7992
Recall	0.8235
f1_score	0.7735
AUC Score	0.7818

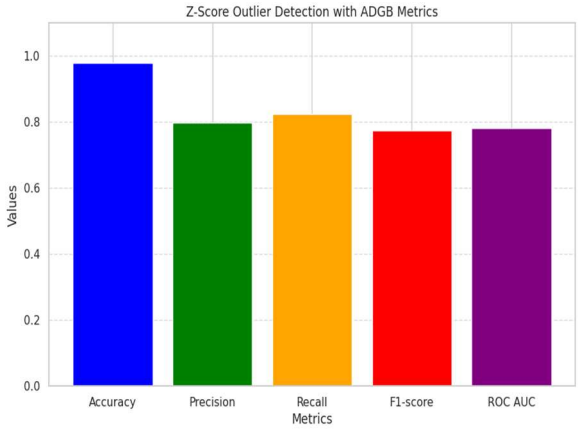


Figure 10: Bar graph shows Z-Score Outlier Detection with ADGB Performance Metrics

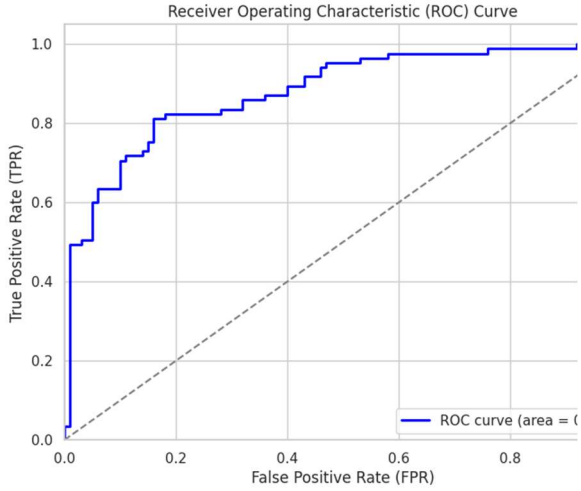


Figure 11: ROC for Z-Score Outlier Detection with ADGB

5.1.4 Comparison of Proposed method and other methods on diabetes dataset

Comparing our Adaptive Boosted Gradient Boosting Machine (ADGB) framework performance against other models presented in the literature, our approach is able to deliver considerable enhancements in all aspects. To illustrate, Abedini et al. (2020) tackled the problem by an ensemble of decision trees, logistic regression and neural networks and achieved 83% of accuracy. Prasanth et al. (2021) managed to realize the accuracy of 86.15% using a ML model such as SVM, Naïve Bayes, Random Forest, and others. At the same time, Saxena et al. (2022) employed classifiers multilayer perceptron, decision trees, KNN and achieved an accuracy of 77.60%. Saputra et al. (2023) on the other hand, suggested a Stacked Multi-Kernel SVM-RF model and recorded a performance of 73.37%. Gupta et al. (2023) implemented Random Forest and SVM yielding an accuracy of 88.61% on carefully prepared dataset.

In correspondence however, our ADGB strategy is able to record a more superior figure of 97.84 % accuracy as shown in table 6. This can be explained by a few reasons. Firstly, there were reasonable reasons for the use of the combing Preprocessing techniques such as MICE as well as SMOTE, recommendations received were quite effective in targeting the problem of missing data as well as class imbalance. Secondly, our combined ANOVA and Grey Wolf Optimizer (GWO) based feature selection technique was useful in embellishing the model learnability with the most relevant features. Thirdly, the ADGB Framework also benefits from the advantages of AdaBoost and Gradient Boosting Machines because they enhance model stability and generalization performance on multiple datasets together.

However, some limitations should still be pointed out regarding the present approach. Although the accuracy measures of the ADGB Framework are impressive, the computed moderate values of precision and F1 score indicate there could be improvements in either feature selection or model tuning to better than the current level in the detection of positive cases. On the other hand, the summary AUC score suggests a good class separation performance yet not optimal and does suggest improvement opportunities and areas of likely future search. There are some drawbacks of course, but the ADGB method establishes a new top line in diabetes prediction, making it more efficient than the models available in the literature on the subject.

Table 6: Comparative Performance with Other Models

Author Name	Method used	Accuracy
Abedini et al. [2020]	Ensemble of Decision Tree, Logistic Regression, Neural Network	83.00%
Prasanth et al. [2021]	SVM, Naïve Bayes, Decision Tree, ANN, LR, k-NN, RF, XGBoost, LightGBM, CatBoost	86.15%

Saxena et al. [2022]	Multilayer Perceptron, Decision Trees, KNN, Random Forest	77.60%
Saputra et al. [2023]	Stacked Multi-Kernel SVM-RF	73.37%
Gupta et al. [20203]	Random Forest, SVM, Decision Tree, KNN	88.61%
Our Study	ADGB (AdaBoost with Gradient Boosting)	97.84%

6. CONCLUSION

In the end, our study offers a solid foundation for the development of affirmation of diabetes using state-of-the-art techniques in ML. The hybrid feature selection GWAN framework boosts the capacity of the dataset for the model prediction by utilizing MICE to address missing data, SMOTE to resolve the issue of class imbalance, and the Z-score outlier detection for preprocessing. By performing hyperparameter tuning in a loop, the combining of two powerful boosting algorithms in one model: Adaptive Boosted GBM (ADGB) which is merging between AdaBoost and Gradient Boosting Machine (GBM) improves the model's performance even more together achieving a high accuracy of 97.84%. This approach does not only attain higher accuracy rates than the claimed ones in the previous research works but also reiterates the significance of ensemble approaches in the analysis of healthcare data. The results illustrate significant improvements in terms of accuracy, precision, recall, and F1-score, thus confirming the effectiveness of the implementation of the proposed method for enhancing the probabilities of diagnosing early diabetes. Though our method demonstrates high accuracy, it faces limitations such as potential biases and challenges in generalization It takes a step forward in the improvement of predictive analytics in healthcare applying for the necessity of model accuracy as well as model interpretability; thus, presenting a systematic approach to the development of reliable and scalable diagnostic tools to help develop specific health management and promotion plans for patients.

REFERENCES

[1] O. O. Oladimeji, A. Oladimeji, and O. Oladimeji, "Classification models for likelihood prediction of diabetes at an early stage using feature selection," Applied Computing and Informatics, May 2021, doi: 10.1108/aci-01-2021-0022.

[2] P. Talari et al., "Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2," PLoS ONE, vol. 19, no. 1, p. e0292100, Jan. 2024, doi: 10.1371/journal.pone.0292100.

[3] J. Abdollahi and S. Aref, "Early Prediction of Diabetes Using Feature Selection and Machine Learning Algorithms," SN Computer Science, vol. 5, no. 2, Jan. 2024, doi: 10.1007/s42979-023-02545-y.

[4] G. K. Teimoory and M. Reza Keyvanpour, "An Effective Feature Selection for Type II Diabetes Prediction," 2024 10th International Conference on Web Research (ICWR), Tehran, Iran, Islamic Republic of, 2024, pp. 64-69, doi: 10.1109/ICWR61162.2024.10533371.

[5] E. Sreehari and L. D. D. Babu, "Critical Factor Analysis for Prediction of Diabetes Mellitus using an Inclusive Feature Selection Strategy," Applied Artificial Intelligence, vol. 38, no. 1, Apr. 2024, doi: 10.1080/08839514.2024.2331919.

[6] M. Abedini, A. Bijari, and T. Banirostam, "Classification of Pima Indian diabetes dataset using ensemble of decision tree, logistic regression and neural network," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 9, no. 7, pp. 7-10, Jul. 2020.

[7] S. Prasanth, K. Banujan, and K. Btgs, "Hyper Parameter Tuned Ensemble Approach for Gestational Diabetes Prediction," in *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Zallaq, Bahrain, 2021, pp. 18-23, doi: 10.1109/3ICT53449.2021.9581926.

[8] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, Article ID 3820360, 2022. doi: 10.1155/2022/3820360. [Online].

Available: https://doi.org/10.1155/2022/3820360.

[9] D. C. E. Saputra, A. Ma'arif, and K. Sunat, "Optimizing Predictive Performance: Hyperparameter Tuning in Stacked Multi-Kernel Support Vector Machine Random Forest Models for Diabetes Identification," Journal of Robotics and Control (JRC), vol. 4, no. 6, pp. 896-904, 2024, doi: 10.18196/jrc.v4i6.20898.

[10] S. C. Gupta and N. Goel, "Predictive Modeling and Analytics for Diabetes using Hyperparameter tuned Machine Learning Techniques," Procedia Computer Science, vol. 218, pp. 1257-1269, 2023, doi: 10.1016/j.procs.2023.01.104. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923001047.

[11] S. Tiwaskar, M. Rashid, and P. Gokhale, "Impact of machine learning-based imputation techniques on medical datasets- a comparative analysis," Multimedia Tools and Applications, Apr. 2024, doi: 10.1007/s11042-024-19103-0.

[12] Tirumanadham, N.S.K.M.K., S, T. & M, S. Improving predictive performance in e-learning through hybrid 2-tier feature selection and hyper parameter-optimized 3-tier ensemble modeling. Int. j. inf. tecnol. (2024). https://doi.org/10.1007/s41870-024-02038-y.

[13] Md. A. Uddin et al., "Machine Learning Based Diabetes Detection Model for False Negative Reduction," Deleted Journal, vol. 2, no. 1, pp. 427–443, Jun. 2023, doi: 10.1007/s44174-023-00104-w.

[14] J. Singh, J. K. Sandhu, and Y. Kumar, "Metaheuristic-based hyperparameter optimization for multi-disease detection and diagnosis in machine learning," Service Oriented Computing and Applications, Jan. 2024, doi: 10.1007/s11761-023-00382-8.

[15] M. Sarıgöl and O. M. Katipoğlu, "Estimation of monthly evaporation values using gradient boosting machines and mode decomposition techniques in the Southeast Anatolia Project (GAP) area in Turkey," Acta Geophysica, Mar. 2023, doi: 10.1007/s11600-023-01067-8.

[16] N. S. K. M. K. Tirumanadham, T. S and S. M, "Evaluating Boosting Algorithms for Academic Performance Prediction in E-Learning Environments," 2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bangalore, India, 2024, pp. 1-8, doi: 10.1109/IITCEE59897.2024.10467968.

[17] "Diabetes Dataset," Kaggle, Aug. 05, 2020. https://www.kaggle.com/datasets/mathchi/diabetes-data-set

[18] J. Yang, Y. Wang, Y. Yang, K. Ding, C. Na, and Y. Yang, "Effects of single and multiple imputation strategies on addressing over-fitting issues caused by imbalanced data from various scenarios," Applied Intelligence, Feb. 2024, doi: 10.1007/s10489-024-05295-3.

[19] R. Bounab, B. Guelib and K. Zarour, "A Novel Machine Learning Approach For handling Imbalanced Data: Leveraging SMOTE-ENN and XGBoost," 2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS), EL OUED, Algeria, 2024, pp. 1-7, doi: 10.1109/PAIS62114.2024.10541220.

[20] N. S. Krishna, Y. V. P. Kumar, K. P. Prakash and G. P. Reddy, "Machine Learning and Statistical Techniques for Outlier Detection in Smart Home Energy Consumption," 2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), Vilnius, Lithuania, 2024, pp. 1-4, doi: 10.1109/eStream61684.2024.10542609.

[21] A. Shanbhag, S. Vincent, S. B. B. Gowda, O. P. Kumar and S. A. J. Francis, "Leveraging Metaheuristics for Feature Selection with Machine Learning Classification for Malicious Packet Detection in Computer Networks," in IEEE Access, vol. 12, pp. 21745-21764, 2024, doi: 10.1109/ACCESS.2024.3362246.

[22] K. Boutahar, S. Laghmati, H. Moujahid, O. E. Gannour, B. Cherradi and A. Raihani, "Exploring Machine Learning Approaches for Breast Cancer Prediction: A Comparative Analysis with ANOVA-Based Feature Selection," 2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), FEZ, Morocco, 2024, pp. 1-7, doi: 10.1109/IRASET60544.2024.10549284.

[23] B. Si, Z. Ni, J. Xu, Y. Li, and F. Liu, "Interactive effects of hyperparameter optimization techniques and data characteristics on the performance of machine learning algorithms for building energy

metamodeling," Case Studies in Thermal Engineering, p. 104124, Feb. 2024, doi: 10.1016/j.csite.2024.104124.

[24] W. Jiang, H. Han, M. He, and W. Gu, "ML-based pre-deployment SDN performance prediction with neural network boosting regression," Expert Systems with Applications, vol. 241, p. 122774, May 2024, doi: 10.1016/j.eswa.2023.122774.

[25] W. Helm, S. Zhong, E. Reid, T. Igou, and Y. Chen, "Development of gradient boosting-assisted machine learning data-driven model for free chlorine residual prediction," Frontiers of Environmental Science & Engineering, vol. 18, no. 2, Sep. 2023, doi: 10.1007/s11783-024-1777-6.

[26] Praveen, S. P., Hasan, M. K., Abdullah, S. N. H. S., Sirisha, U., Tirumanadham, N. K. M. K., Islam, S., ... & Ghazal, T. M. (2024). Enhanced feature selection and ensemble learning for cardiovascular disease prediction: hybrid GOL2-2 T and adaptive boosted decision fusion with babysitting refinement. Frontiers in Medicine, 11, 1407376.

[27] Praveen, S. P., Sandeep, K., Sai, N. R., Sharma, A., Pandey, J., & Chouhan, V. (2024). Outlier Management and its Impact on Diabetes Prediction: A Voting Ensemble Study. Journal of Intelligent Systems & Internet of Things, 12(1).

[28] Praveen, S. P., Saripudi, V., Harshalokh, V., Sohitha, T., Karthik, S. V. S., & Sreekar, T. V. P. S. (2023, December). Diabetes Prediction with Ensemble Learning Techniques in Machine Learning. In 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 1082-1089). IEEE.

[29] Krishna, A. Y., Kiran, K. R., Sai, N. R., Sharma, A., Praveen, S. P., & Pandey, J. (2023). Ant Colony Optimized XGBoost for Early Diabetes Detection: A Hybrid Approach in Machine Learning. Journal of Intelligent Systems & Internet of Things, 10(2).

[30] Praveen, S. P., Kodete, C. S., Bhyrapuneni, S., Satukumati, S. B., & Shariff, V. (2024). Revolutionizing Healthcare: A Comprehensive Framework for Personalized IoT and Cloud Computing-Driven Healthcare Services with Smart Biometric Identity Management. Journal of Intelligent Systems & Internet of Things, 13(1).

[31] S. Vahiduddin, P. Chiranjeevi and A. Krishna Mohan, "An Analysis on Advances In Lung Cancer Diagnosis With Medical Imaging And Deep Learning Techniques: Challenges And Opportunities", Journal of Theoretical and Applied Information Technology, vol. 101, no. 17, Sep. 2023.

[32] Qi, S. S. J., & Nagalingham, S. (2023). Business Intelligence Data Visualization for Diabetes Health Prediction. *International Journal of Advanced Computer Science and Applications*, *14*(1).