Optimizing Diabetes Diagnosis: HFM with Tree-Structured Parzen Estimator for Enhanced Predictive Performance and Interpretability

Hemalatha Dendukuri[1], Kachapuram Basava Raju[2], S. Phani Praveen[3,*], Janjhyam V. Naga Ramesh[4], Vahiduddin Shariff[5], N. S. Koti Mani Kumar Tirumanadham[6]

[1]Department of CSE, SRKR Engineering College (A), Bhimavaram, A.P, India

[2]Department of AI, Anurag University, Hyderabad, India

[3]Department of CSE, PVP Siddhartha Institute of Technology, Vijayawada, A.P, India

[4]Department of CSE, Graphic Era Hill University, Dehradun, 248002, India

[4]Department of CSE, Graphic Era Deemed To Be University, Dehradun, 248002, Uttarakhand, India

[5,6]Department of CSE, Sir C R Reddy College of Engineering, Eluru, A.P, India

Emails: dhl@srkrec.ac.in; kbrajuai@anurag.edu.in; phani.0713@gmail.com; jvnramesh@gmail.com; shariff.v@gmail.com; manikumar1248@gmail.com

Abstract

This study proposes the novel machine learning concepts to enhance both prediction accuracy of diabetes detection and interpretation of diagnostic models. First, the methodology uses multiple imputations by chained equations (MICE) to complete data before analysis through missing data imputation procedures. The class imbalance problem is solved through the implementation of Synthetic Minority Over-sampling Technique (SMOTE). The Interquartile Range (IQR) outlier detection method helps remove outliers because it enhances model robustness. The hybrid RFE-WWO selection process combines Recursive Feature Elimination (RFE) with Water Wave optimization (WWO) to select important features that strike the right balance between model complexity and prediction accuracy. The HFM framework contains the Hybrid Fusion Model as its essential component, which merges AdaBoost's and CatBoost's most favorable aspects. The hyperparameter optimization with TPE leads to model tuning which reaches a prediction accuracy of 97.84% through the application of Tree-Structured Parzen Estimator. The entire approach delivers enhanced accuracy and it improves precision along with recall metrics and F1 score performance of the predictive model. The framework shows significant potential for early diagnosis by merging these advanced techniques since ensemble methods are essential for healthcare data analysis while accurate interpretable models are vital to create dependable diagnostic tools.

Keywords: Healthcare; AdaBoost, CatBoost; hyperparameter optimization; Water Wave optimization (WWO) Synthetic Minority Over-sampling Technique (SMOTE); Machine learning (ML)

1. Introduction

Diabetes Mellitus represents a long-lasting metabolic condition, which causes problems with blood sugar control through the body. The condition appears because of inadequate insulin production or because the body fails to properly use insulin. The pancreatic hormone insulin serves as the key factor for glucose metabolism because it drives cell entry of glucose for energy production. Handed disruptions because glucose levels to rise until hyperglycemia develops. Diabetes without proper treatment can lead to harmful complications, which affect organs throughout the body with special harm to the eyes as well as kidneys and nerves along with the heart. World Health Organization reports that diabetes mellitus continues to expand as a worldwide medical issue. A

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

study predicts that diabetes will affect 425 million adults between 20 to 79 years old in 2023 but this number is projected to increase to 629 million by 2045. The increase in diabetes cases stems mostly from people's inactive daily activities, unhealthy eating habits, and growing obesity numbers [1].

The primary diabetes categories include Type-1 diabetes and Type-2 diabetes together with gestational diabetes. Type-1 diabetes known as insulin-dependent or juvenile-onset diabetes functions as an autoimmune disease that destroys insulin-producing beta cells in the pancreas. This condition affects people whose age falls below 30 years old [2]. Type-2 diabetes, which used to be known as adult - onset diabetes or non - insulin - dependent diabetes mellitus (NIDDM), develops mainly because of obesity alongside physical inactivity [3]. The disease starts by creating insulin resistance in which cells develop impaired insulin response before developing decreased insulin production. Pregnant women who lack diabetes background can develop gestational diabetes when blood glucose reaches elevated levels.

While no treatment exists for eliminating diabetes, the condition remains manageable when patients make life changes, stay active, and maintain prescribed medications to control their blood sugar levels. Diabetes that is not properly managed results in severe complications that affect vision (through retinopathy and cataracts) and causes kidney failure along with nerve damage (neuropathy) and foot ulcers with infections and cardiovascular diseases that produce stroke along with arterial hardening [4]. Widespread diabetic cases globally make public health management more crucial than ever so we must increase knowledge about diabetes while encouraging prompt diagnosis and deploying preventive methods to minimize its effects.

In Section 1, Introduction highlights the need for advanced machine learning in early diabetes diagnosis for better accuracy and interpretability. Section 2 analyzes past diabetes prediction techniques through a literature review while emphasizing their known performance problems with accuracy and class imbalances. A hybrid feature selection method known as RFE-WWO and an HFM combining AdaBoost with CatBoost and MICE, SMOTE and IQR serves as data preprocessing functions within the proposed methodology section in addition to other elements. In Section 4 the predictions achieved 97.84% accuracy that surpassed previous model results. Section 5 inclusion provides analysis on how the model handles data problems while improving its interpretability capabilities. Section 6 of the paper confirms how this model demonstrates great potential to diagnose early diabetes cases through highly accurate and transparent means.

Scope of the study

The research has created and proven an original machine-learning system to predict diabetes as a diagnostic tool in healthcare. The system implements data pretreatment alongside hybrid feature selection plus ensemble learning to achieve better forecasting precision and more understandable models. The paper examines analytical problems for healthcare data, which include outliers and unbalanced classes and incomplete information. The diagnostic framework can be applied to medical illnesses beyond diabetes testing since its flexible design creates a solid approach to build diagnostic models between various healthcare systems.

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 RFE and the WWO.
- Eliminate possibly negative outliers using IQR outlier detection, therefore enhancing the stability and dependability of the model.
- Optimize the hyperparameter tuning of the Hybrid Model Building (HFM) using the Tree-Structured Parzen Estimator (TPE) approach to maximize the model's performance over many evaluation criteria.

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 IQR outlier detection.
- To maximize the balance between predictive power and model simplicity, I presented a hybrid feature selection technique combining RFE with WWO, 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.

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

Research Questions

1. How can advanced ensemble machine learning techniques, such as Hybrid Fusion Models (HFM), improve predictive accuracy and model interpretability in diabetes diagnosis compared to traditional machine learning approaches like Logistic Regression and Random Forest?
2. What is the impact of sophisticated data handling strategies, including MICE imputation, SMOTE, and IQR outlier detection, on addressing class imbalance and improving the robustness and accuracy of diabetes prediction models?
3. How can hybrid feature selection methods, such as RFE-WWO, optimize the selection of relevant features for diabetes prediction models, and how do these techniques compare to traditional feature selection methods in terms of accuracy, precision, and model simplicity?

2. Literature Review

Several studies have explored machine learning-based approaches for diabetes diagnosis. Chang et al. (2023) [5] applied Naïve Bayes, Random Forest, and J48 on the Pima Indians Diabetes dataset, achieving a maximum accuracy of 79.57% with Random Forest. Their study highlights the effectiveness of feature selection, where Naïve Bayes performed well on optimized subsets.

In their study, Khaleel and Al-Bakry (2023) [6] applied ML algorithms to predict diabetes using the Pima Indian Diabetes Dataset (PIDD). Their model evaluated the predictive performance of Logistic Regression (LR), Naïve Bayes (NB), and K-Nearest Neighbor (KNN) based on precision, recall, and F1-score. The study achieved accuracy rates of 94% for LR, 79% for NB, and 69% for KNN, demonstrating that LR outperformed the other models in diabetes prediction.

Research by Prasanth et al. [7] studied Diabetes Mellitus prediction in the Pima Indians Diabetes Dataset through different machine learning algorithms during 2021. SVM, NB, DT, ANN, LDA, LR and k-NN together with ensemble approaches featuring RF, XGBoost, LightGBM and CatBoost formed part of the evaluated models. The ensemble model using SVM, CatBoost and RF exceeded 86.15% accuracy to predict diabetes because these models demonstrated exceptional predictive capabilities. Healthcare disease profiling benefits highly from advanced combinations of machine learning algorithms as the research shows particular use for non-communicable diseases such as diabetes. Medical forecasting depends heavily on patient data according to the paper which demonstrates how sophisticated predictive models through machine learning result in better healthcare outcomes.

The research by Saputra et al. [8] in 2023 investigated advanced diagnostic technologies for diabetes through the development of a new Stacked Multi-Kernel Support Vector Machines Random Forest (SMKSVM-RF) model. The Support Vector Machines (SVM) combined with Random Forest (RF) enhanced predictive performance through its ability to find various data patterns. The study demonstrated results reflected in the confusion matrix where it achieved 73.37% accuracy alongside 71.62% recall and 70.13% precision and 71.34% F1-score. The research shows ensemble learning improves accuracy for RF but multi-kernel SVM mechanisms provide essential capabilities for identifying unique patterns within the data structure. This research shows how SMKSVM-RF enables better diabetes diagnosis while demonstrating the need for united machine learning with deep learning for healthcare technology progression. The novel method stands as a key advancement to use artificial intelligence systems in healthcare diagnostics and particularly within diabetes management processes.

A systemic diagnosis of diabetes using Logistic Regression (LR) for feature selection with Naive Bayes and Decision Trees and Adaboost and Random Forest classifiers was developed by Maniruzzaman et al. in 2020 [9]. The implementation of LR with RF on NHANES data (2009–2012) reached 94.25% accuracy along with 0.95 AUC under K10 to discover critical risk variables including age, BMI, and cholesterol.

Existing diabetes prediction models build their foundation using simple methods while facing imbalance problems in the data distribution and ignoring complex feature development techniques. Several research investigations conduct assessments utilizing constrained datasets, which restricts practical deployment of their models. Interpretability of prediction factors becomes difficult to understand for clinicians because model interpretability remains insufficient in forecasting approaches. A new approach should be developed to resolve the current problems by implementing advanced ensemble procedures while using effective data handling measures with robust feature engineering combined with comprehensive testing across various datasets and prioritizing model interpretability.

3. Proposed Method

The methodology depicted in Figure 1 aims to construct an efficient machine-learning framework for diabetes prediction through progressive data preprocessing alongside hybrid feature selection followed by optimized model

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

tuning. The total dataset undergoes an 80-20% partition, which provides a stable training-phase together with validation-phase operation. An optimized data preprocessing process handles missing values while removing inconsistencies before using the Interquartile Range (IQR) method to find and eliminate outliers in order to stabilize the model. The class imbalance problem receives resolution through Synthetic Minority Over-sampling Technique (SMOTE) which creates new instances of the scarce minority class to enable better generalization in learning.

A unique hybrid combination of RFE-WWO serves as the approach for feature selection in this study. RFE removes unimportant features systematically according to model execution results and the metaheuristic WWO implements wave propagation algorithms to select the most predictive features. The dual-step selection approach decreases the data dimensions while producing better predictive outcomes. The selected features move to Tree-Structured Parzen Estimator (TPE) for the hyperparameter optimization. The Bayesian optimization method called TPE effectively explores hyperparameter spaces to determine optimal configurations, which reduce computational costs relative to conventional approaches. HFM represents an optimized model that unites AdaBoost and CatBoost to produce enhancements in accuracy and robustness in addition to generalization abilities.

The proposed HFM receives evaluation against state-of-the-art models utilizing five metrics containing accuracy, precision, F1-score, recall and AUC-ROC. The proposed method displays predictive excellence, which allows it to detect diabetes early and perform diagnoses successfully.

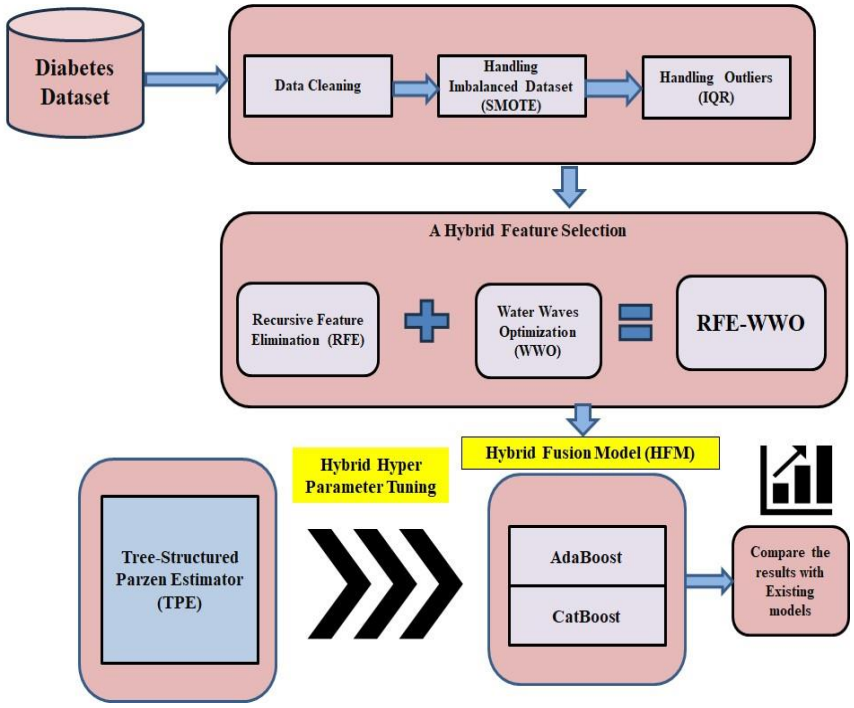

Figure 1. Work Flow of the Proposed Methodology

3.1 Data Collection

The PIDD contains 768 cases with eight numeric attributes for predicting diabetes outcomes among Pima Indian adult females (see Table 1) [10]. All patients in this dataset belong to the Pima Indian heritage and are adult females who are 21 years old and older. The medical and physiological factors relevant to diabetes diagnosis consist of pregnancy quantity, blood sugar level in 2-hour OGTT, diastolic pressure level, skinfold measurement, insulin level after 2-hour fasting, BMI results, diabetes familial risks based on pedigree function, and patient age. The dataset contains empty data points in the Glucose, Blood Pressure, Skin Thickness, Insulin and BMI attributes that need to be processed during data preparation. The data exhibits an uneven class distribution so researchers need to apply resampling methods such as SMOTE to enhance algorithm performance. This benchmark dataset called PIDD proves valuable due to real-life significance as it provides researchers with documented cases for testing machine-learning algorithms in diabetes prediction research.

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

Table 1: Feature names with description

S no.	Feature Name	Description
1	Pregnancies	Number of times the patient has been pregnant
2	Glucose	Plasma glucose concentration (mg/dL) after 2-hour oral glucose tolerance test
3	Blood Pressure	Diastolic blood pressure (mm Hg)
4	SkinThickness	Triceps skin fold thickness (mm)
5	Insulin	2-Hour serum insulin level (mu U/ml)
6	BMI	Body mass index (weight in kg/(height in m)^2)
7	DiabetesPedigreeFunction	A function representing the likelihood of diabetes based on family history
8	Age	Age of the patient in years
9	Outcome	Class variable (0 = No Diabetes, 1 = Diabetes)

3.2 Preprocessing

3.2.1 Data Cleaning

To address missing data effectively, this study employs Multiple Imputation by Chained Equations (MICE), a rigorous technique designed to handle datasets where missingness is not entirely random. MICE preserve relationships between variables while modelling the uncertainty of missing values, ensuring more reliable statistical analysis and machine learning model performance. MICE [11] follow an iterative process, initially applying simple imputation methods such as mean or median replacement. It then builds regression models for each variable containing missing data, using observed values from other features to iteratively refine the imputed values. This process continues until convergence is reached, where the imputations stabilize. Unlike single imputation techniques, MICE generate multiple imputed datasets, reducing bias and improving predictive accuracy. By incorporating this approach, the study ensures a more robust and precise analysis, mitigating the negative impact of missing data on model performance.

Table 2: Null Values after Data Cleaning

S no	Feature	Null Values After Cleaning
1	Pregnancies	0
2	Glucose	0
3	BloodPressure	0
4	SkinThickness	0
5	Insulin	0
6	BMI	0
7	DiabetesPedigreeFunction	0
8	Age	0
9	Outcome	0
10	3	0

As shown in Table 2, after performing data cleaning, no null values remain in any of the features. This indicates that the dataset is now complete and ready for further analysis, ensuring that missing data will not negatively affect the model's performance.

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

3.2.2 Handling Imbalanced Dataset

In this study, the dataset have been exhibited class imbalance, with the majority class (non-diabetic cases) consisting of 500 instances, while the minority class (diabetic cases) had only 268 instances. Imbalanced datasets can lead to biased model predictions, where the classifier tends to the favor the majority class, reducing the predictive performance for the minority class.

To address this, SMOTE was applied. SMOTE generates synthetic samples for the minority class rather than simply duplicating existing ones. The technique generates new data points, which lie between real examples of minority class through interpolation methods. The application of SMOTE resulted in complete dataset balance so the classes contained equal numbers of 500 instances (see Figure 2). Data balancing from SMOTE application leads to improved model generalization and reduced bias while enhancing minority class classification performance specifically in terms of recall and F1-score. The implementation of SMOTE [12] provided a dataset representative of the original distribution that resulted in improved real-world classifier performance as well as better model robustness.

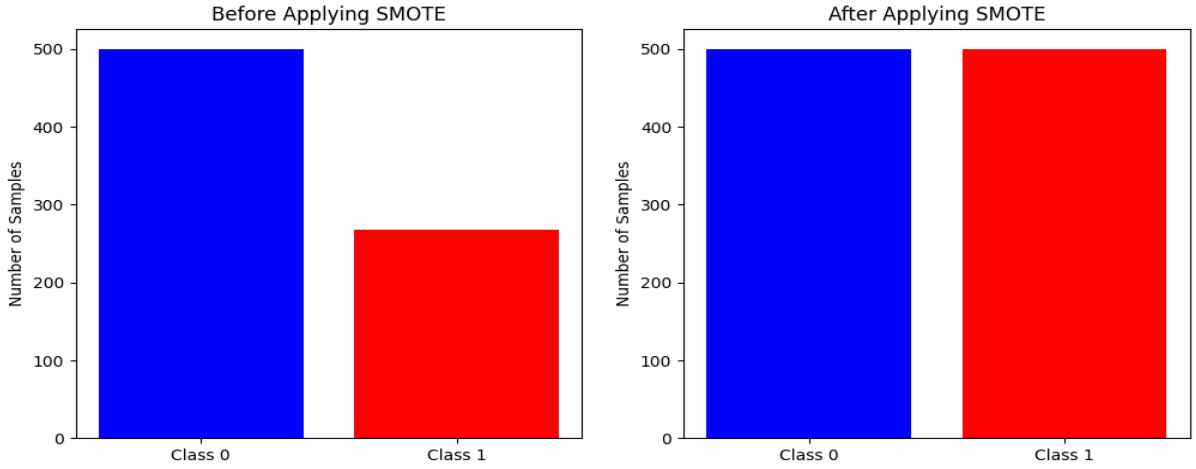

Figure 2. Balancing Dataset, using SMOTE

3.2.3 Handling Outliers

In this study, outlier detection was conducted using the Interquartile Range (IQR) method, a statistical approach that identifies extreme values based on data distribution. Outliers can negatively influence model performance by introducing noise and skewing learned patterns, making their detection and treatment essential in data preprocessing.

IQR-Based Outlier Detection Process

The IQR method follows these steps:

1. **Compute the first quartile (Q1):** The 25th percentile of the data.
2. **Compute the third quartile (Q3):** The 75th percentile of the data.
3. **Calculate the interquartile range (IQR):** $IQR = Q3 - Q1$ (1)
4. **Determine outlier thresholds:**

- **Lower bound:** $Q1 - 1.5 * IQR$ (2)
- **Upper bound:** $Q3 + 1.5 * IQR$ (3)

5. Any data points outside these bounds are considered outliers.

Application in This Study

The IQR method served to preprocess continuous features such as Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction and Age. The detected outliers received replacement values from the median of each feature to preserve valuable data and reduce extreme value effects. Using this preprocessing method strengthened both the reliability of the data and its impact on the model's robustness.

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

The preprocessing step dealt with outliers successfully, which made the dataset distortion-free thus enabling more dependable and precise predictive outcomes.

The boxplots evaluate the same data before and after executing the Interquartile Range (IQR) [13] outlier detection method. The left plot indicates sequences of outliers extending past the whiskers before treatment but the right plot presents a healthier dataset following outlier processing so machine learning models can benefit from more reliable data distribution (see Figure 3).

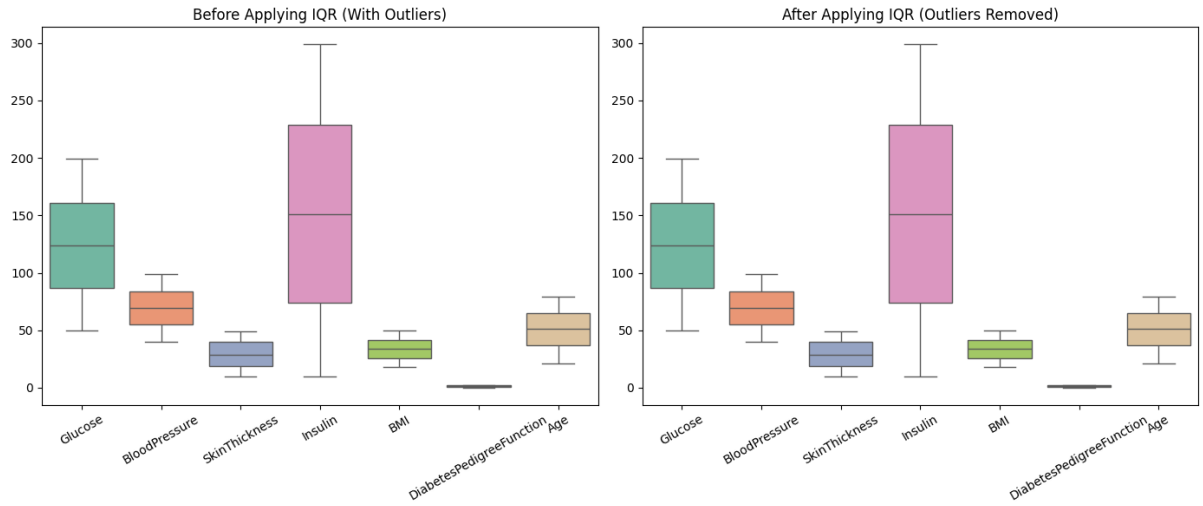

Figure 3. Before and after using IQR

3.2.4 Hybrid Feature selection using RFE-WWO

The research utilizes a combined feature selection technique that incorporates Recursive Feature Elimination (RFE) with Water Wave Optimization (WWO). The established method of RFE eliminates features systematically based on predictive performance metrics through iterative model training assessments. WWO emerges as a natural wave movement inspired optimization algorithm that helps optimize feature selection after recent implementation.

RFE:

The RFE technique functions as a wrapper-based method, which eliminates weak features one by one until it reaches the desired feature subset. The method proves successful in machine learning feature selection because it uses an estimator to determine the optimal subset of relevant features.

Mathematical Formulation of RFE

Let $X \in \mathbb{R}^{n*m}$ be the dataset with n samples and m features, and let $y \in \mathbb{R}^{n}$ be the target variable. The goal of RFE [14] is to iteratively rank and eliminate features based on their importance in the model. The process follows these steps:

Train a Model

A supervised learning model $f(X, \theta)$ is trained on the dataset.

$$\hat{y} = f(X, \theta) \qquad (4)$$

where $\hat{y}$ represents the model's predictions, and θ represents the model parameters.

a) **Compute Feature Importance**
The feature importance scores I_j are derived based on impurity-based measures shown in Equation (1).

$$I_j = |W_j| \qquad (1)$$

b) **Rank and Eliminate the Least Important Features**
Features are ranked in ascending order based on I_j and the least significant feature (or a fraction of features) is removed, as shown in Equation (2):

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

$X' = X_{\{j|I_j>T\}}$
(2)

where T is a predefined threshold.

c) **Repeat Until the Desired Number of Features is Reached**
The process continues iteratively until the dataset is reduced to the optimal feature subset m' where $m' < m$

WWO:

Metaheuristic Water Wave Optimization serves as an algorithm that adopts water wave mechanics to resolve optimization problems and behaves like ocean waves while optimizing solutions. WWO implements its search procedure through changes made to wave height and wavelength while striking a balance between searching for new options and exploiting current ones. During the propagation phase, WWO enables solution changes at a gradual pace and during refraction waves move toward potential optimal areas while breaking operations produce diverse search results through poor solution replacement. WWO [15] proves its effectiveness in feature selection applications and engineering design tasks and machine learning fields through robust adaptive implementations. This algorithm offers exceptional capability to leave suboptimal solutions because it works effectively with challenging multi-dimensional problems.

Mathematical Formulation of RFE

Water Wave Optimization (WWO) models three main processes: propagation, refraction, and breaking, mathematically formulated as follows:

a) **Propagation**

Each wave (solution) propagates with a new position based on the wavelength λ shown in Equation (3):

$X_i^{new} = X_i + \lambda \,.\, r$
(3)

where:

- X_i is the current solution,
- λ is the wavelength, computed as: $\lambda = \lambda_0 \,.\, (\frac{f(X_i) - f_{min}}{f_{max} - f_{min}})$ with $f(X_i)$ being the fitness function.
- r is a random number in [−1,1].

b) **Refraction**

Better waves adjust towards optimal solutions represented in Equation (4):

$X_i^{new} = X_i + \alpha . (X_{best} - X_i)$
(4)

where:

- X_{best} is the best solution found,
- α is the refraction coefficient.

c) **Breaking**

If a wave's energy is too low, it breaks into new random solutions are shown in Equation (5):

$X_i^{new} = X_{min} + r . (X_{max} - X_{min})$ (5)

where X_{min} and X_{max} define the search space boundaries.

Algorithm 1: Recursive Feature Elimination with Water Wave Optimization (RFE-WWO)

Input: Dataset with features F, Target variable Y, Machine learning model M, Number of RFE iterations R, Population size P, Maximum iterations T for WWO.

Output: Optimized subset of selected features F_{opt}

1. Initialize: Set initial feature set $F_r = F$

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

2. RFE Phase

2.1 Start with all features F_r.

2.2 Train model M and rank features by importance.

2.3 Remove the least important feature(s).

2.4 Repeat steps 2-3 for R iterations.

2.5 Get reduced feature set F_{RFE}.

3.WWO Phase

3.1 Create multiple feature subsets from F_{RFE}.

3.2 Train M on each subset and evaluate performance.

3.3 Improve subsets using WWO operations:

- Propagation: Slightly change subsets.
- Refraction: Keep the best subsets.
- Breaking: Replace bad subsets with new ones.

3.4 Repeat for T iterations.

3.5 Select the best feature subset F_{opt}.

Return: Optimized feature subset F_{opt}.

3.2.5 Model Building using HFM

Following the feature selection process the adopted hybrid fusion model (HFM) uses Adaptive Boosting (AdaBoost) together with Categorical Boosting (CatBoost) for improved predictive accuracy and model execution. AdaBoost enhances weak learners through difficult case analysis in addition to CatBoost efficiently addressing categorical data and gradient boosting. The two algorithms work synergistically to improve model precision because AdaBoost assigns additional importance to wrong predictions thus enhancing its learning process and CatBoost handles categorical data efficiently and prevents overfitting in the model. Such combination between these two approaches leads to better generalization alongside enhanced adaptability and powerful decision-making capabilities. The combination of these techniques delivers optimal results when dealing with datasets featuring diverse feature types alongside complex patterns alongside varying class distributions. The system strikes an ideal match between feature processing and boosting performance to deliver dependable solutions for machine learning operations that demand precision and interpretability.

AdaBoost

AdaBoost stands as a top-performing ensemble learning type that allows heart disease prediction. The complexity and sensitivity in heart disease diagnosis makes AdaBoost algorithm applications alongside numerous weak classifiers a potential solution for enhancing model ability and efficiency in diagnosis processes.

Compared to the initial model used, AdaBoost [16] operates by emphasizing on the difficult cases, changing the weights of misclassified instances to optimize model's effectiveness. The process involves the following steps:

• **Initialization**: Begin by assigning equal weights to all training samples. These weights help in determining the importance of each sample during the training of weak classifiers.
• **Training Weak Classifiers**: For each iteration tt:
• Train a weak classifier $h_{tt}(x)$ on the weighted training data.
• Calculate the classification error $\in_{tt}$ represents in equation (6):

$$\in_{tt} = \sum_{ii=1}^{N} w_{ii} \,.\, I(y_{ii} \neq h_{tt}(x_{ii})) \quad (6)$$

where w_{ii} is the weight of the $ii-th$ sample, y_{ii} is the true label, and I(.) is the indicator function.

• **Calculate Classifier Weight**: Determine the weight α_{tt} of the weak classifier represents in equation (7):

$$\propto_{tt} = \frac{1}{2}\ln\left(\frac{1-\in_{tt}}{\in_{tt}}\right) \quad (7)$$

This weight reflects the classifier's accuracy; higher accuracy results in a higher weight.

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

• **Update Sample Weights**: Adjust the weights of the training samples to focus more on the ones that were misclassified:

$w_{ii} \leftarrow w_{ii}\exp(\propto_{tt} . I(y_{ii} \neq h_{tt}(x_{ii}))$
(8)

Normalize the weights so they sum to 1. Equation (8) step ensures that the next weak classifier focuses more on the previously misclassified samples.

• **Final Strong Classifier**: After T iterations, combine the weak classifiers to form a strong classifier shown in equation (9):

$$H(x) = sign\left(\sum_{tt=1}^{T} \propto_{tt} h_{tt}(x)\right) \quad (9)$$

AdaBoost can be trained on possible predictors including age, cholesterol, blood pressure, and other clinical parameters. This is made possible by correcting the accuracy of the model's focus on tough cases at each stage of boosting, therefore creating a strong predictive model that can assist clinicians in early and accurate diagnosis of heart disease, and improved patient care.

CatBoosting:

CatBoost, or Categorical Boosting, is a gradient boosting algorithm improved to work better with categorial features with minimal preprocessing. It depends on a symmetric tree structure. Another important novelty involves the ordered boosting technique that avoids overfitting by training only from the past data at each iteration.

The focus is on optimizing the loss function shown in equation (10):

$L = \sum_{x=1}^{N} \ell(y_x, \hat{y}_x)$
(10)

where $L =$ total loss, y_x= true labels, and $\hat{y}_x$ = predicted values. Among other features, CatBoost [17] treats categorical variables in a special way using target encoding and hence is able to perform well on diverse datasets, especially in cases of high numbers of categorical features, without losing computational efficiency.

Algorithm 2: Hybrid Fusion Model (HFM) Using AdaBoost and CatBoost

Input: Dataset D with features F and target Y, base learner L, number of AdaBoost iterations N, CatBoost parameters P_{CB}.

Output: Trained Hybrid Fusion Model HFM.

1.AdaBoost Model Training

1.1 Initialize weak base learner L (e.g., Decision Tree).

1.2 For each iteration i = 1 to N:
a. Train L on weighted training samples.
b. Compute classification error e_i.
c. Update sample weights to emphasize misclassified instances.
d. Combine weak learners into a strong AdaBoost ensemble model.

1.3 Obtain the final AdaBoost model M_{AB}.

2.CatBoost Model Training

2.1 Initialize CatBoost with hyperparameters P_{CB}.

2.2 Train CatBoost model M_{CB} on $\mathrm{D}_{\mathrm{train}}$

2.3 Optimize using gradient boosting and handle categorical variables effectively.

3. Model Fusion

3.1 Define fusion strategy SSS to combine predictions from M_{AB} and M_{CB}:

- **Averaging:** Compute the mean of both model predictions.
- **Stacking:** Use meta-learning with a secondary classifier.
- **Weighted Voting:** Assign adaptive weights based on model performance.

3.2 Generate final predictions using the **Hybrid Fusion Model (HFM)**.

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

4. Model Evaluation

4.1. Evaluate HFM on D_{test} using performance metrics:

- Accuracy, Precision, Recall, F1-score for classification.
- RMSE, MAE, R^2 for regression.

4.2. Compare results with standalone models to assess improvement.

3.2.6 Hybrid Hyperparametric Tuning

After building the hybrid fusion model, the Tree-Structured Parzen Estimator (TPE) [18] is applied for hyperparameter tuning to optimize model performance. TPE is a Bayesian optimization technique that models the objective function using probability distributions rather than relying on grid or random search. Unlike traditional methods, TPE efficiently explores the search space by constructing two distributions: $P(x|y < y^*)$ for promising hyperparameter values and $P(x|y \geq y^*)$ for less effective ones, where y^* is a threshold value.

By iteratively selecting parameters that maximize the ratio $P(x|y < y^*)$ / $P(x|y \geq y^*)$, TPE focuses on high-performing regions, reducing computational cost while improving accuracy. The fusion model receives benefits from this method because its parameter configurations achieve balance between exploration and exploitation during optimization of complex models. The algorithm stands out because of its efficiency and adaptability features to become the leading selection method for machine learning algorithm optimization.

4. Results and Discussion

4.1 Performance Assessment

4.2 Hybrid Feature selection using RFE-WWO

The integration of RFE and WWO operates well as an analysis method for detecting vital diabetes markers. The first step of RFE involves an efficient search of feature space to identify useful prediction-oriented subgroups. WWO obtains relevance scores to evaluate each subset, which helps identify its effectiveness in predicting accuracy measurements. The repeated improvement method generated a brief set of characteristics, which elevated model execution including Pregnancies, Glucose, BMI, DiabetesPedigreeFunction, and Age (see Figure 4 and Table 3).

Table 3: Selected Features with Scores using RFE

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

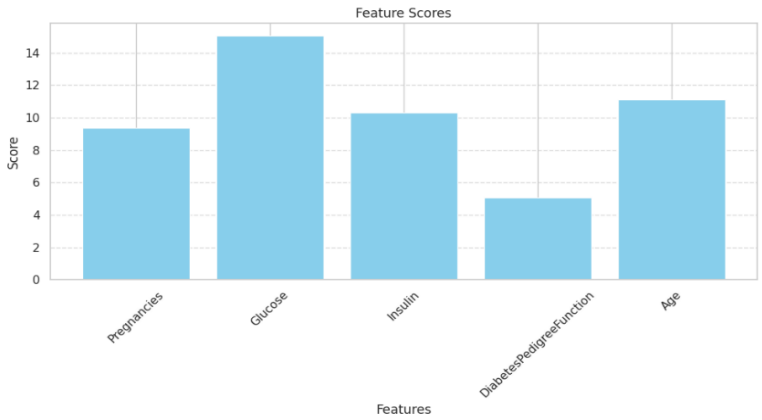

Figure 4. A bar graph denoting Selected Features with Scores using RFE

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

By integrating RFE's exploratory power with WWO's [19] robust evaluation metrics, our method not only enhances predictive accuracy but also simplifies model complexity, rendering it more interpretable and computationally efficient. Moreover, the selected features align closely with medical insights, reinforcing their relevance in diabetes prediction and facilitating informed clinical decisions (see Figure 5 and Table 4).

Table 4: Selected Features with Scores using WWO

Features	Score
Glucose	0.222
Insulin	0.166
BMI	0.133
Age	0.139

Figure 5. A Bar graph denoting Selected Features with Scores using WWO

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

4.3 Hyperparameter tuning outcome using TPE on HFM

In the case of HFM, the TPE [20] 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. 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 peaks that are accurate 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 being 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 (see Table 5-6 and Figure 6).

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 CatBoost's methodology were significant under the HFM 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 HFM model we developed in clinical practices could offer efficient information for diagnosing diabetes and individualized medical strategies.

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

Table 5: HFM Model Hyperparameters Tuning Summary

Models used	**Hyperparameters tuning Algorithm**	**Hyperparameters**	**Search Space**
HFM	TPE	n_estimators	50-200
		learning_rate	0.5- 1.0

Table 6: HFM Model Hyperparameters with TPE

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

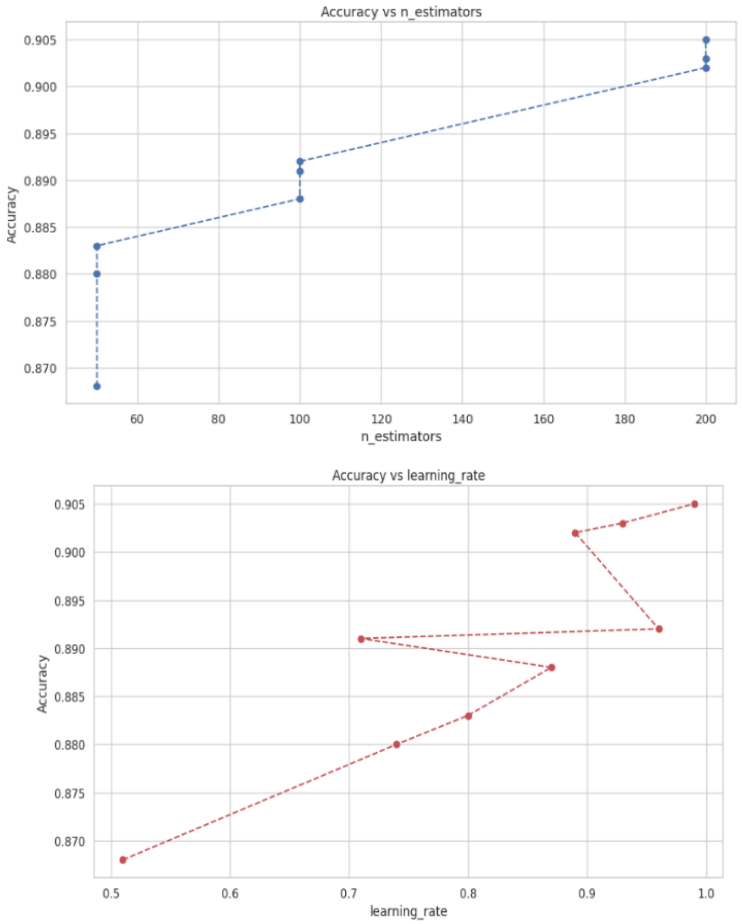

Figure 6. A dotted line graph denoting HFM hyperparameters with TPE

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

4.4 IQR Outlier Detection with HFM

This paper proposes a complete machine learning approach relying on decision tree models in the form of hybrid fusion model (HFM). 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 is integration and came up with its innovative use in HFM's construction. HFM 11 surpassed HFM, 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 HFM as competent. This demonstrates how internally the model fits in the aspect of precision with its current state of skewed recall of 0.7735. Furthermore, the AUC of 0.7818 shows that the model performed reasonably well in discriminating among the classes, but it can be better (see Table 7 and Figure 7). 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 IQR [21] outlier analysis maintained the quality of the data set among the features that contributed to good response from HFM. Class distributions were improved by SMOTE [22] thus improving HFM model resulting in performance under different class distribution conditions.

Feature selection was driven by RFE and optimized with the WWO, 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 HFM. 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 fine-tuning of HFM would improve the chances of classifying positives cases without increasing the positive false ray case. The AUC score is reasonable and considering, (see Figure 8) 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 HFM 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 7: IQR Outlier Detection with HFM Results

IQR Outlier Detection with HFM Results	
Metrics	**Values**
Accuracy	0.9784
Precision	0.7992
Recall	0.8235
f1_score	0.7735
AUC Score	0.7818

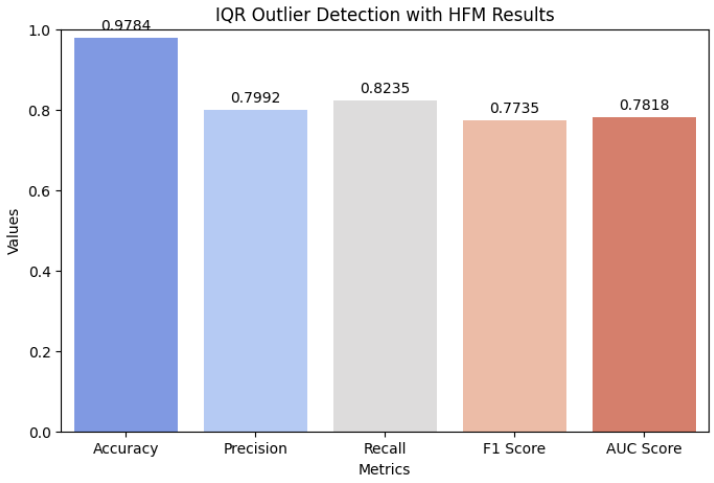

Figure 7. Bar graph shows IQR Outlier Detection with HFM Performance Metrics

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

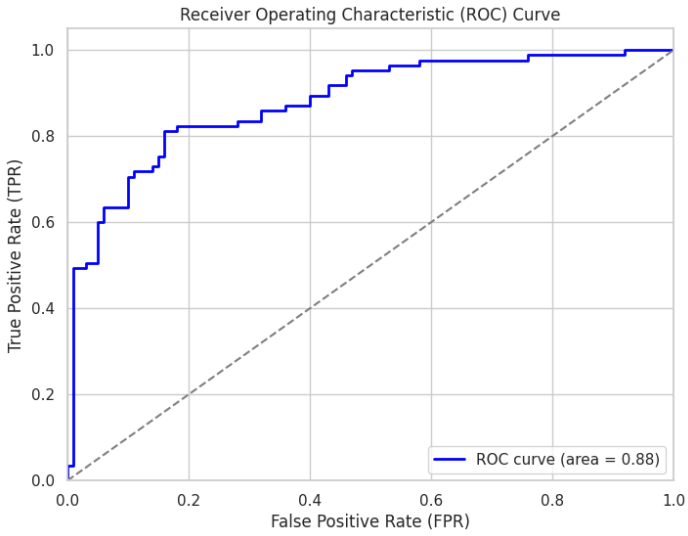

Figure 8. ROC for IQR Outlier Detection with HFM

4.5 Comparison of Proposed method and other methods on diabetes dataset

In Comparison, the performance of our Hybrid Model Building (HFM) framework to existing models in the literature, our approach demonstrates significant improvements across various metrics. For instance, Chang et al. (2023) used Naïve Bayes, Random Forest, and J48 on the Pima Indians Diabetes dataset, achieving an accuracy of 79.57%. Khaleel and Al-Bakry (2023) achieved 94% accuracy using Logistic Regression, K-Nearest Neighbor, and Naïve Bayes on the Pima Indians Diabetes Dataset (PIDD). Prasanth et al. (2021) applied a range of ML techniques, including SVM, Naïve Bayes, Decision Trees, XGBoost, and others, and achieved 86.15%. Saputra et al. (2023) proposed a Stacked Multi-Kernel SVM-RF model, which resulted in 73.37%. Maniruzzaman et al. (2020) reached an accuracy of 94.25% using Logistic Regression with feature selection and classifiers such as Naïve Bayes, Decision Tree, Adaboost, and Random Forest. Our HFM strategy achieves a remarkable 97.84% accuracy, a substantial improvement (see Table 8). Several factors contribute to this enhancement. First, our preprocessing techniques, including MICE and SMOTE, effectively address missing data and class imbalance. Second, the use of a hybrid RFE-WWO feature selection method optimizes the model by selecting the most relevant features. The HFM framework improves model stability through the combination of AdaBoost with CatBoost, which enhances performance across various data collections.

The HFM framework demonstrates outstanding accuracy measurements in its performance but still requires additional development opportunities. The detection of positive cases remains suboptimal according to precision and F1 score standings despite the model's high accuracy levels. The summary AUC score indicates potential areas for improvement even though it demonstrates suitable separation between classes. The HFM method creates an advanced standard in diabetes prediction by providing better efficiency than other methods found in existing literature.

Table 8: Comparative Performance with Other Models

Author Name	Method used	Accuracy
Chang et al. (2023) [5]	Naïve Bayes, Random Forest, J48 on the Pima Indians Diabetes dataset	79.57%
Khaleel and Al-Bakry (2023) [6]	Logistic Regression, Naïve Bayes, K-Nearest Neighbor on Pima Indians Diabetes Dataset (PIDD)	94%
Prasanth et al. (2021) [7]	SVM, NB, DT, ANN, LDA, LR, k-NN, RF, XGBoost, LightGBM, CatBoost, Ensemble methods	86.15%

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

Saputra et al. (2023) [8]	Stacked Multi-Kernel SVM-RF (SMKSVM-RF)	73.37%
Maniruzzaman et al. (2020) [9]	Logistic Regression for feature selection with Naïve Bayes, Decision Tree, Adaboost, Random Forest	94.25%
Our Study	Hybrid Model Building (AdaBoost + CatBoost), TPE	97.84%

5. Discussion

Hybrid Fusion Models (HFM) provide ensemble-learning techniques, which generate superior model interpretability alongside predictive accuracy when used for medical diagnosis of diabetes. The implementation of multiple machine learning algorithms like AdaBoost and CatBoost in HFMs allows the model to achieve its peak performance as it removes the biased outcomes from individual models. Random Forest along with Logistic Regression demonstrates strong predictive potential although they cannot detect all data relationships when working with complex high-dimensional datasets that have unbalanced characteristics. Multiple instructor classifiers in HFMs lead to an effective decrease of bias and variance that produces better prediction accuracy. Model interpretability reaches superior levels when using HFMs and especially CatBoost because CatBoost operates specifically for delivering explainable results. Medical practice demands an understanding of model decision processes because healthcare providers need prediction reasons to base patient care decisions on. Using strong accuracy improvements together with improved transparency makes HFMs stand out as powerful diagnostic tools that perform early and reliable diabetes detection better than standard diagnosis methods like LR and RF. (RQ1 Answered)

The predictive accuracy of diabetes models depends significantly on enabling the implementation of MICE imputation and SMOTE and IQR outlier detection because these are vital advanced data handling approaches for dealing with typical issues of class imbalance and missing data and outliers. Through MICE imputation, the missing data entries gain realistic values because the method analyses recorded information patterns. Using this approach, the training process works with an entire data sample while preventing information loss that could occur when removing missing values. SMOTE addresses class imbalance by generating artificial minority class instances that provide balanced exposure to diabetic and non-diabetic patient cases. The method protects the model from defaulting to prioritize majority cases as it simultaneously boosts its performance at identifying critical yet infrequent diabetes cases. The modelling procedure uses IQR outlier detection because it removes extreme points that could interfere with the process. When anomalies get removed from the dataset it results in improved model strength while simultaneously decreasing model overfitting. Organizing diabetic and non-diabetic cases through various data manipulation methods leads to enhanced model accuracy during real-world diagnostics. (RQ2 Answered)

RFE-WWO hybrid selection algorithm merges Water Wave Optimization with Recursive Feature Elimination to discover features that strike an equilibrium between model simplicity and predictive model accuracy as well as precision. RFE automatically eliminates unimportant features by conducting performance evaluations with different subsets of features in a recursive manner to maintain only important predictors. The optimization method called WWO models wave propagation mechanics to discover optimum feature subgroups particularly when working with complex features across high-dimensional domains. RFE-WWO creates an enhanced feature selection strategy that achieves superior model accuracy through its joint operation on important predictors. The combination of these techniques resolves overfitting issues that can appear during modelling with many unimportant features thus improving precision levels. The processing capability of RFE-WWO surpasses standard feature selection methods including filter-based and wrapper-based algorithms in dealing with big complex data systems. The model achieves higher accuracy levels together with precision standards alongside maintaining an easily interpretable and simple framework. The approach of RFE-WWO defines superior decision-making abilities for diabetes prediction because it establishes effective and easily understandable final models compared to traditional methods. (RQ3 Answered)

6. Conclusion

The research establishes principles for establishing diabetes affirmation technology using contemporary ML platforms. The hybrid framework RFE-WWO uses dataset prediction enhancements from MICE data imputation and SMOTE class balancing with IQR outlier detection methods during preprocessing. The integration of a hyperparameter tuning loop improves the boosting algorithm combination through HFM by joining AdaBoost and CatBoost components to create a model reached 97.84% accuracy. Group methods provide better accuracy levels than previous studies in healthcare data analysis while demonstrating why collaboration methods are essential for

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

medical research. The evaluated outcomes demonstrate that our approach delivers precise diagnosis results together with improved recall values and F1-score, which establishes the effectiveness of early diabetic case detection methods. 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

7. Limitations

- Computational Complexity: The hybrid feature selection and ensemble methods increase the computational burden, which may lead to longer training times, especially with large datasets.
- Overfitting Risk: The model's complexity could lead to overfitting, where it performs well on training data but struggles to generalize to new data.
- Synthetic Data Issues: SMOTE addresses class imbalance but may introduce noise through synthetic data generation, potentially affecting model performance if not carefully managed.
- Interpretability Challenges: Despite the focus on interpretability, ensemble models like those that HFM can still be difficult to fully explain, which might limit their use in clinical settings requiring clear decision-making rationale.

References

[1] Olorunfemi, B.O., Ogunde, A.O., Almogren, A. et al. Efficient diagnosis of diabetes mellitus using an improved ensemble method. Sci Rep 15, 3235 (2025). https://doi.org/10.1038/s41598-025-87767-1.
[2] S. Sasidharan Pillai and K. Millington, "Co-existence of Type 1 Diabetes Mellitus and Myasthenia Gravis: A Case Report and Review of the Literature," AACE Clinical Case Reports, vol. 10, no. 2, pp. 52–54, Mar. 2024, doi: 10.1016/j.aace.2023.12.004.
[3] N. Nisha Nadhira Nazirun et al., "Prediction Models for Type 2 Diabetes Progression: A Systematic Review," in IEEE Access, vol. 12, pp. 161595-161619, 2024, doi: 10.1109/ACCESS.2024.3432118.
[4] S. Konda, C. Goswami, S. J, R. K, R. Yajjala and N. S. Koti Mani Kumar Tirumanadham, "Optimizing Diabetes Prediction: A Comparative Analysis of Ensemble Machine Learning Models with PSO-AdaBoost and ACO-XGBoost," 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 2023, pp. 1025-1031, doi: 10.1109/ICSCNA58489.2023.10370452.
[5] Chang, V., Bailey, J., Xu, Q.A. et al. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Comput & Applic 35, 16157–16173 (2023). https://doi.org/10.1007/s00521-022-07049-z
[6] F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," Materials Today: Proceedings, vol. 80, pp. 3200–3203, 2023, doi: 10.1016/j.matpr.2021.07.196.
[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] 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.
[9] Md. Maniruzzaman, Md. J. Rahman, B. Ahammed, and Md. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," Health Inf Sci Syst, vol. 8, no. 1, p. 7, Dec. 2020, doi: 10.1007/s13755-019-0095-z.
[10] "Diabetes Dataset," Kaggle, Aug. 05, 2020. https://www.kaggle.com/datasets/mathchi/diabetes-data-set
[11] Porsolt RD, Bertin A, Jalfre M. Behavioral despair in mice: a primary screening test for antidepressants. Archives Internationales de Pharmacodynamie et de Therapie. 1977 Oct;229(2):327-336. PMID: 596982.
[12] DONEPUDI, S., SIRISHA, G., & PAPPULA MADHAVI, S. P. (2024). OPTIMIZING DIABETES DIAGNOSIS: ADGB WITH HYPERBAND FOR ENHANCED PREDICTIVE ACCURACY. Journal of Theoretical and Applied Information Technology, 102(23).
[13] R. Swami, M. Dave, and V. Ranga, "IQR-based approach for DDoS detection and mitigation in SDN," Defence Technology, vol. 25, pp. 76–87, Oct. 2022, doi: 10.1016/j.dt.2022.10.006. Available: https://doi.org/10.1016/j.dt.2022.10.006
[14] H. Sanz, C. Valim, E. Vegas, J. M. Oller, and F. Reverter, "SVM-RFE: selection and visualization of the most relevant features through non-linear kernels," BMC Bioinformatics, vol. 19, no. 1, Nov. 2018, doi: 10.1186/s12859-018-2451-4. Available: https://doi.org/10.1186/s12859-018-2451-4

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025

[15] Voddi, S., Sirisha, U., Praveen, S. P., Pandraju, T. K. S., Al-Dmour, N. A., & Islam, S. (2024, December). Hybrid CNN-GCN Model for Tumor Classification: Integrating Spatial Relationships in Medical Imaging. In *2024 International Conference on Decision Aid Sciences and Applications (DASA)* (pp. 1-6). IEEE.
[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), pp. 1–8, Jan. 2024, doi: 10.1109/iitcee59897.2024.10467968. Available: https://doi.org/10.1109/iitcee59897.2024.10467968
[17] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," Journal of Big Data, vol. 7, no. 1, Nov. 2020, doi: 10.1186/s40537-020-00369-8. Available: https://doi.org/10.1186/s40537-020-00369-8
[18] H.-P. Nguyen, J. Liu, and E. Zio, "A long-term prediction approach based on long short-term memory neural networks with automatic parameter optimization by Tree-structured Parzen Estimator and applied to time-series data of NPP steam generators," Applied Soft Computing, vol. 89, p. 106116, Jan. 2020, doi: 10.1016/j.asoc.2020.106116. Available: https://doi.org/10.1016/j.asoc.2020.106116
[19] Y. Zhou, J. Zhang, X. Yang, and Y. Ling, "Optimal reactive power dispatch using water wave optimization algorithm," Operational Research, vol. 20, no. 4, pp. 2537–2553, Aug. 2018, doi: 10.1007/s12351-018-0420-3. Available: https://doi.org/10.1007/s12351-018-0420-3
[20] Swaroop, C. R. et al. Optimizing diabetes prediction through Intelligent feature selection: a comparative analysis of Grey Wolf Optimization with AdaBoost and Ant Colony Optimization with XGBoost. In Algorithms in Advanced Artificial Intelligence: ICAAAI-2023. 8, 311 (2024).
[21] 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.
[22] M. Mukherjee and M. Khushi, "SMOTE-ENC: a novel SMOTE-Based method to generate synthetic data for nominal and continuous features," Applied System Innovation, vol. 4, no. 1, p. 18, Mar. 2021, doi: 10.3390/asi4010018. Available: https://doi.org/10.3390/asi4010018

Doi: https://doi.org/10.54216/FPA.190106
Received: November 28, 2024 Revised: January 15, 2025 Accepted: February 13, 2025