Original Article

# A Hybrid Whale Optimization and XGBoost Framework for Accurate Prediction of Type 2 Diabetes Mellitus

Prakash Arumugam [1], Abinayaa Sennanur Srinivasan [2], Divya Bhavani Mohan [3], Santosh Kumar [4], Miral Mehta [5], Mainul Haque [6,7,8,9,10]

Please Click on Photo

## ABSTRACT

### Introdaction

Type 2 Diabetes Mellitus (T2DM) has become a worldwide health issue that has to be taken care of. Thus, predictive models have to be developed that are accurate and efficient to help with the early diagnosis and preventive measures. In this work, a hybrid of the Whale Optimization Algorithm (WOA) and Extreme Gradient Boosting (XGBoost) is proposed to improve T2DM prediction.

### Materials and Methods

To optimize XGBoost hyperparameters for better generalization and fewer classification errors, the Whale Optimization Algorithm (WOA) was used. To assess the effectiveness and performance of the suggested approach, two benchmark datasets were evaluated: the PIMA Indian Diabetes dataset (PID) with 768 records and the Diabetes Risk Prediction dataset with 100,000 records.

### Results

The WOA-XGBoost model recorded an accuracy of 98.7%, precision, recall, and F1-score of 99% for the dataset, which consists of 768 records, and an accuracy of 99.84%, precision - 99.91%, recall - 99.89% and F1-Score - 99.9% for the dataset, which consists of 100000 records. It was observed that the proposed method performed better than the other state-of-the-art methods.

### Conclusion

The proposed WOA-XGBoost model demonstrates highly accurate and reliable prediction performance for T2DM across both small and large datasets. These results indicate that the hybrid optimization-based approach is practical for early diagnosis and can be valuable in real-world clinical decision-support systems.

## Keywords

Clinical Decision Support, Classification Model, Disease Risk Assessment, Early Diagnosis, Feature Importance, Healthcare Analysis, Hyperparameter Tuning, Machine Learning, Metaheuristic Algorithm, Performance Evaluation

## INTRODUCTION

Type 2 Diabetes Mellitus (T2DM) is a permanent metabolic syndrome, which is associated with insulin resistance and high levels of blood glucose in the body [1]. It represents around 90-95% of all the worldwide diagnosed cases of diabetes, and it remains an increasing concern as far as the health of the population is concerned

1. Department of Research & IQAC, Karnavati University, Gujarat, India.
2. Department of Electronics and Communication Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, Tamil Nadu, India.
3. Department of Computer Science and Engineering, Unitedworld Institute of Technology, Karnavati University, Gujarat, India.
4. Department of Periodontology and Implantology, Karnavati School of Dentistry, Karnavati University, India.
5. Department of Pediatric and Preventive Dentistry, Karnavati School of Dentistry, Karnavati University, India.
6. Independent Researcher. Former Professor, Department of Pharmacology and Therapeutics, National Defense University of Malaysia, Kuala Lumpur, Malaysia.
7. Department of Research, Karnavati School of Dentistry, Karnavati University, Gandhi Nagar, Gujarat, India.
8. Scientific Committee, Global Alliance for Infections in Surgery, Macerata, Italy.
9. Department of Pharmacology and Therapeutics, Eastern Medical College and Hospital, Cumilla, Bangladesh.
10. Public Health Foundation Bangladesh, Dhaka, Bangladesh.

## Correspondence

**1. Prakash Arumugam**

Department of Research & IQAC,Karnavati University, Gujarat, India. Email: prksh830@gmail.com

**2. Mainul Haque**

Independent Researcher. Block: C, Road: 10, House: 266, Khilgaon, Dhaka 1219, Bangladesh. Email: runurono@gmail.com. Cell Phone & WhatsApp: +8801703605918.

DOI: https://doi.org/10.3329/bjms.v25i1.86405

because of inactivity, poor diet, obesity, and inherited factors [2]. T2DM has become widespread in the world with disastrous consequences that include the occurrence of cardiovascular diseases, kidney diseases, nerve injuries, and vision loss. An early diagnosis and intervention are needed to avoid the progression of the disease and lessen the healthcare burden. Along with the widespread development of artificial intelligence (AI) and machine learning (ML), data-driven predictive models have proven helpful in healthcare practice, particularly in the early detection of chronic diseases, including T2DM [3]. These models will use historical and clinical data to identify patterns and factors that indicate the risk of disease onset [4]. Nevertheless, the performance of these models is usually limited by the issues of class imbalance, noise features, hyperparameter optimization difficulties, and the lack of data in some cases.

Extreme Gradient Boosting (XGBoost) is one of the leading ML models that has received significant attention, especially for classification tasks, due to its scalability and speed [5]. XGBoost is an ensemble learning technique with a gradient-boosting background that can handle missing values and overfitting through regularization [6]. The efficiency of the model can hardly be achieved unless the careful selection of hyperparameters is involved in its performance.

To overcome this weakness, hyperparameter optimization in ML became increasingly automated through nature-inspired metaheuristic techniques [7]. In these algorithms, simulated natural phenomena and biological behaviors are used to obtain solutions close to optimal values of high-dimensional search spaces [8]. This paper presents a recent swarm intelligence meta-heuristic technique, the Whale Optimization Algorithm (WOA), for the feeding of humpback whales [9]. WOA has been claimed to be easy and quick to converge and explore rather than exploit, which makes it a decent method for the optimization of XGBoost hyperparameters [10]. The suggested WOA-XGBoost model uses the computational strengths of WOA and the envisaging strength of XGBoost to enhance the prediction of T2DM. One of the unique aspects of this study is the evaluation of the model on two datasets: one with a small number of records and the other with a significantly large volume of data [11]. Evaluating the model with two datasets helps ensure its generalization and practical effectiveness, as real-world data scenarios can range from scarce to overwhelmingly large.

## Related Works

This section presents the various methodologies developed by researchers over the past years. A computational program with a Deep Neural Network (DNN) classifier and a feature importance model is proposed to determine and forecast the T2DM at an early stage precisely. This aims to enhance the prediction to identify and treat the disease on time [12]. A systematic review was conducted by the researchers to determine the best methods of Machine Learning (ML) and Deep Learning (DL) in the process of predicting T2DM overcoming heterogeneity, and interpretability issues of the current models. The authors attempted to inform the choice of the appropriate approaches to build a new predictive model for the prediction of T2DM [13]. The authors proposed an explainable AI method—a soft voting classifier—to predict diabetes mellitus, demonstrating its use in an accurate and interpretable manner. Its emphasis is on achieving optimality in prediction performance and on guaranteeing that medical practitioners can interpret the decisions of the model to use it on a clinical level [14]. An ensemble model based on ML, including Light Gradient Boosting Machine, k-NN, and Adaboost, is trained to anticipate the occurrence of T2DM with the ability to accurately predict 90.76% of such occurrences and also overcome the problem of imbalance in the classes [15]. A blended ensemble learning (EL) model comprising Bayesian Networks and Radial Basis Function (RBF) networks is proposed to improve accuracy in diagnosing and treating diabetes. The presented EL technique shows better quality with an excellent accuracy of 97.11% in the forecast of diabetic disease compared to other ML methods [16]. A new supervised ML asymmetric model is developed to precisely predict T2DM disease with respect to the metric information, with over 85% accuracy with the practice fusion data set. The purpose of the model is to offer the early warning system in medical assessment, eliminating the issue introduced by late diagnosis and low accuracy of the current ML perceptions [17]. This paper summarizes the effects of several factors during the prenatal, neonatal, and early childhood stages on the predisposition of individuals to T2DM and their development in adulthood, including nutrition, environmental exposures, and physiological factors. It is important to learn about these early-life determinants to develop prevention and delaying strategies for T2DM onset [18]. The author highlights the application of ML algorithms in the early diagnosis of

T2DM, revealing that the Random Forest algorithm attained the best accuracy of 98% in the identification of the disease. This emphasizes the potential of machine learning to improve precision in screening for T2DM and patient outcomes [19]. This research paper has studied the prevalence of type-2 diabetes mellitus using various methods, and the results showed the prevalence rate of 13.1% among the adults living in urban Meerut aged 30 and above. It has recognized age, socioeconomic status, educational status, marital status, family history, hypertension, alcohol use, and smoking as eminent risk factors of diabetes prevalence [20]. This paper addresses the intervention strategies, such as behavior correction or pharmacological agents, that have importance in the individuals with impaired glucose tolerance (IGT) and impaired fasting glycemia (IFG) in reducing the risk of developing T2DM. Strategies to reduce the risk in higher-risk populations and the effects of the disease, especially cardiovascular complications [21]. This paper talks about the pathophysiology, diagnostic parameters, and the methods of achieving remission of T2DM. Remission refers to the period of time below the level of glycated hemoglobin (HbA1c) that ignores pharmacological treatment, during which it has been at the level of at least half a year. It addresses the variety of triggers that can determine the development of T2DM and the necessity of follow-up observations after the remission. It mentions measures such as pharmacological interventions, nutritional adjustments, and metabolic surgery [22]. This paper provides a review of the two-way interdependence between pulmonary tuberculosis (TB) and T2DM. It shows how they increase the risks and severity of each other, resulting in a two-fold TB burden on global health. Both conditions need to be treated differently at times when they occur [23]. This review paper summarizes the present evidence and gives suggestions on how to diagnose and treat T2DM amongst patients with ischemic heart disease (IHD) and acute coronary syndromes (ACS). It emphasizes the significance of early identification and administration of certain glucose-reducing drugs that have established cardiovascular advantages to enhance success and free up the medical care [24]. In this paper, T2DM is described as a disorder characterized by an inadequate insulin response, leading to elevated blood glucose levels. It emphasizes that genetic and hereditary conditions may have a role in the issue of insulin resistance, as well as low expression of the insulin receptor gene and an increased level of fatty acids in the blood. It also notes that exercise has the potential to reduce insulin resistance [25]. This paper is a review of T2DM, a metabolic disease that has a critical prevalence rate around the globe, in terms of diagnosis, treatment options available, such as lifestyle adjustments and medication, and new alternatives. It emphasizes that, though new drugs are being prepared, there is still no cure for the disease [26]. In this article, T2DM, a chronic metabolic disease, is discussed, which is progressively becoming popular all over the globe concerning its diagnosis, existing mode of treatment encompassing lifestyle interventions and medication, and newly developing drugs. It indicates that new medications are under development, but there is still no cure against the disease [27]. In this paper, a literature review is undertaken to determine the current state of T2DM prediction, revealing that ML algorithms tend to be the best predictive models compared to conventional statistical regression models. It also points to the increased accuracy of predictions in the case of adding clinical data and biomarkers. Congenetic markers are associated with greater limitations to the dimensionality and heterogeneity [28]. An ML application, a form of artificial neural network, predicts T2DM occurrence based on measurements of relevant features. The model developed can generate positive results, with an accuracy of 86% and an ROC value of 0.934, demonstrating potential for both diagnosis and prevention [29]. This paper discussed the viability of the ML and DL approaches to diabetes prediction by evaluating their usability against the conventional ones and advanced ones, such as CNN and LST, and a combination of both: CNN and LSTM. According to the research, the mixed CNN+LSTM model is the best among the others, as it has 98% accuracy concerning diabetes prediction [30]. In this paper, the flexible ML approaches (DNN, XGBoost, RF) are assessed in predicting T2DM, also considering the issue of class imbalance in the data. It compares several methods, including switch thresholds, cost-sensitive learning strategies, and sampling methods, to improve the accuracy of the minority class [31]. This paper presents the framework of DL models to screen noninvasively for T2DM through the combination of the multimodal data, namely the chest X-ray (CXR) images and Electronic Health Records (EHRs). Combining these data sources and end-to-end DL architectures, e.g., the combined ResNet-LSTM architecture, leads to better efficiency and diagnostic value in T2DM detection than using only CXR and fewer training samples [32]. This paper presents a streamlined model (CNN-Bi-LSTM), for detecting and predicting T2DM in real time using the Indian diabetes dataset. With an accuracy of 98.85%, the proposed model comparatively outweighs other DL models, and the model is expected to equip the clinicians with complete information about the patients so that they can be proactive in patient care [33]. This paper constructs and tests a DL model to screen T2DM based on rusinghotographs and reveals more predictive ability, which, with the addition of conventional risk factors, can significantly enhance the invasive method, mainly representing the acquired characteristics of diabetes, where the study offers a practical application in risk stratification of general people [34].

Although diabetes prediction with diverse ML and optimization methods has already been researched, the majority of techniques do not demonstrate high accuracy rates across heterogeneous datasets with different magnitudes and do not include efficient mechanisms of hyperparameter tuning. To overcome the above gap, a hybrid WOA and XGBoost model is proposed, in which the exploration-exploitation characteristic of the Whale Optimization Algorithm (WOA) is utilized to undertake the optimum parameter optimization concerning XGBoost that is valid in small as well as large data sets.

### Objectives of The Study

This investigation aims to develop a new, highly accurate predictive model for T2DM using an optimized hybrid approach combining the XGBoost algorithm with the Whale Optimization Algorithm (WOA). By employing the WOA to find the optimal set of XGBoost hyperparameters, the proposed model will accurately classify and minimize classification error. This investigation also examines the scalability and robustness of the developed model by evaluating its performance across both large and small data volumes.

## MATERIALS AND METHODS

This section focuses on the specific architecture that will be utilized in the suggested approach to perform an effective classification and prediction of T2DM. This approach embraces the Synthetic Minority Oversampling Technique (SMOTE) to aid in the resolution of class discrepancy and the XGBoost classifier to enhance learning performance. This will include data preprocessing, SMOTE application, model training, and performance assessment.

### Datasets Used

Two publicly available datasets were used to estimate the performance and stability of the suggested Whale Optimization Algorithm-tuned XGBoost (WOA-XGBoost) model. These data sets were selected to represent the two different real-world situations, one involving a small number of samples and the other involving a large amount of data. This test illustrates the capability of generalization of the data, presenting various sizes in the proposed model.

### PIMA Indians Diabetes dataset (PID)

The widely used first dataset is the Pima Indians Diabetes (PID) dataset, which contains medical diagnostic data from female Pima Indian patients aged 21 years or older. It gives a total sample of 768 and 8 numerical inputs, including glucose level, BMI, age, and insulin level, with an output that depicts whether the patient has diabetes or not in the form of a binary. The dataset is known to be a low-sample dataset and is frequently used in the healthcare sector as a benchmark for the classification issue.

### Diabetes Risk Prediction Dataset (DRP)

The second dataset is a high-volume diabetes risk prediction dataset with 100,000 records and six clinical and demographic variables, including age, blood pressure, gender, family history, cholesterol level, and activity level. The outcome measure is a dichotomous variable indicating the risk of diabetes (yes/no). This set of data reflects practical, large-scale, real-life situations where the scalability of the model and its efficiency are verified. The summary of the selected datasets is presented in Table 1.

**Table 1**: Comparative Summary of Datasets.

| Feature | PIMA Indians Diabetes | Diabetes Risk Prediction |
|---|---|---|
| Total Samples | 768 | 100,000 |
| Number of Features | 8 | 6 |
| Output Variable | Binary (Diabetes: Yes/No) | Binary (Risk: Yes/No) |
| Dataset Size Category | Low | High |
| Domain | Healthcare | Healthcare |
| Source | UCI Repository | Public Kaggle Dataset |
| Preprocessing Applied | Normalization, SMOTE | Normalization, SMOTE |

### Data Pre-processing

Data preprocessing is critical to guarantee that the datasets are proper enough to be used in the training and the evaluation of the proposed WOA-XGBoost model. The same procedures were used on each of the datasets:

### Data Cleaning

An in-depth review has shown that neither dataset has missing values. Therefore, there was no need to impute or drop records. Z-scores and the interquartile range (IQR) were among the statistical measures used to assess outliers. Columns such as glucose and BMI were checked to ensure that the extreme values in them were not kept out, since they are real-life situations.

### Feature Scaling

To address differences in feature scales, standardization was applied using the formula:

$$X' = \frac{X-\mu}{\sigma} \qquad (1)$$

Where $\mu$ is the mean, and $\sigma$ denotes the standard deviation.

### Class Imbalance Handling

Both the datasets were non-balanced, especially the PIMA dataset, which contained fewer positive diabetes cases. The Smote technique, i.e., the Synthetic Minority Oversampling Technique (SMOTE), was used to create synthetic minority samples:

$$x_{new} = x_i + \lambda(x_{\dot{z}} - x_i) \qquad (2)$$

Where '$x_i$' is a minority sample, $x_{\dot{z}}$ is its nearest neighbor, and $\lambda \in [0,1]$ is a random value.

### Data Partitioning

Following the preprocessing, each of the datasets was divided into training and test in a ratio of 80/20 to have a balanced evaluation:

$$(X_{train,}X_{test,}y_{train,}y_{test})=train\_test\_split(X,y,test\_size=0.2,random\_state=42) \qquad (3)$$

### Model Construction using XGBoost

XGBoost trains a series of decision trees like an ensemble. The new trees are added at every iteration to minimize the loss because the loss is kept as low as possible at each stage.

Let the objective function of XGBoost be defined as:

$$\mathrm{L}(\theta)=\sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)} + \sum_{k=1}^{t} \Omega(f_k) \qquad (4)$$

Where $\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i)$ and $f_k \in \mathcal{F}$, the space of regression trees; 'l' is the differentiable convex loss function; $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2$ and λ is the regularization term penalizing complexity.

### Whale Optimization Algorithm (WOA) for Hyperparameter parHyperparameter

WOA is a nature-based optimization algorithm that mimics the use of the bubble-net hunting technique used by the humpback whale, and it is used for tuning the hyperparameters of XGBoost.

WOA maintains a population of whales $X = \{X_1, X_2, ...., X_N\}$ where each whale $X_i$ represents a candidate solution (i.e., a set of XGBoost hyperparameters).

Encircling Prey (Exploitation Phase)

$$\vec{D} = |\vec{C}.\vec{X}^* - \vec{X}(t)| \qquad (5)$$

$$\vec{X}(t+1) = \vec{X}^* - \vec{A}.\vec{D} \qquad (6)$$

Where the current best solution is represented by $X$; $\vec{A} = 2\vec{a}.\vec{r} - \vec{a}$; $\vec{C} = 2.\vec{r}$; $a$ decreases linearly from 2 to 0 over the iterations; $\vec{r} \in [0,1]$ is a random vector.

Spiral Updating (Exploration Phase)

$$\vec{X}(t+1) = \vec{D}'.e^{bl}.\cos(2\pi l) + \vec{X}^* \qquad (7)$$

Where $\vec{D}' = |\vec{X}^* - \vec{X}(t)|$; b is a constant for spiral shape, $l \in [-1,1]$ and is a random number.

Position update (Search for Prey)

$$\vec{X}(t+1) = \vec{X}_{ramd} - \vec{A}.|\vec{C}.\vec{X}_{rand} - \vec{X}(t)| \qquad (8)$$

$\vec{X}_{ramd}$ denotes a randomly selected whale.

The algorithm iterates until a stopping criterion is met (maximum iterations or convergence), and returns the optimal hyperparameter set $\theta^*$.

### Final Model Training

Once the hyperparameters are optimal in WOA, they are used to fit an XGBoost classifier on the SMOTE-balanced dataset. The high learning mechanism and the global search implemented in the proposed model improve the classification accuracy and generalization ability of XGBoost and WOA, respectively. This optimized model is much more accurate than the conventional machine learning classifiers, because this model minimizes bias and variance, which is appropriate in high-stakes classification problems, such as medical diagnostics and anomaly finding problems in imbalanced datasets. The flow diagram of the proposed model is shown in Figure 1.

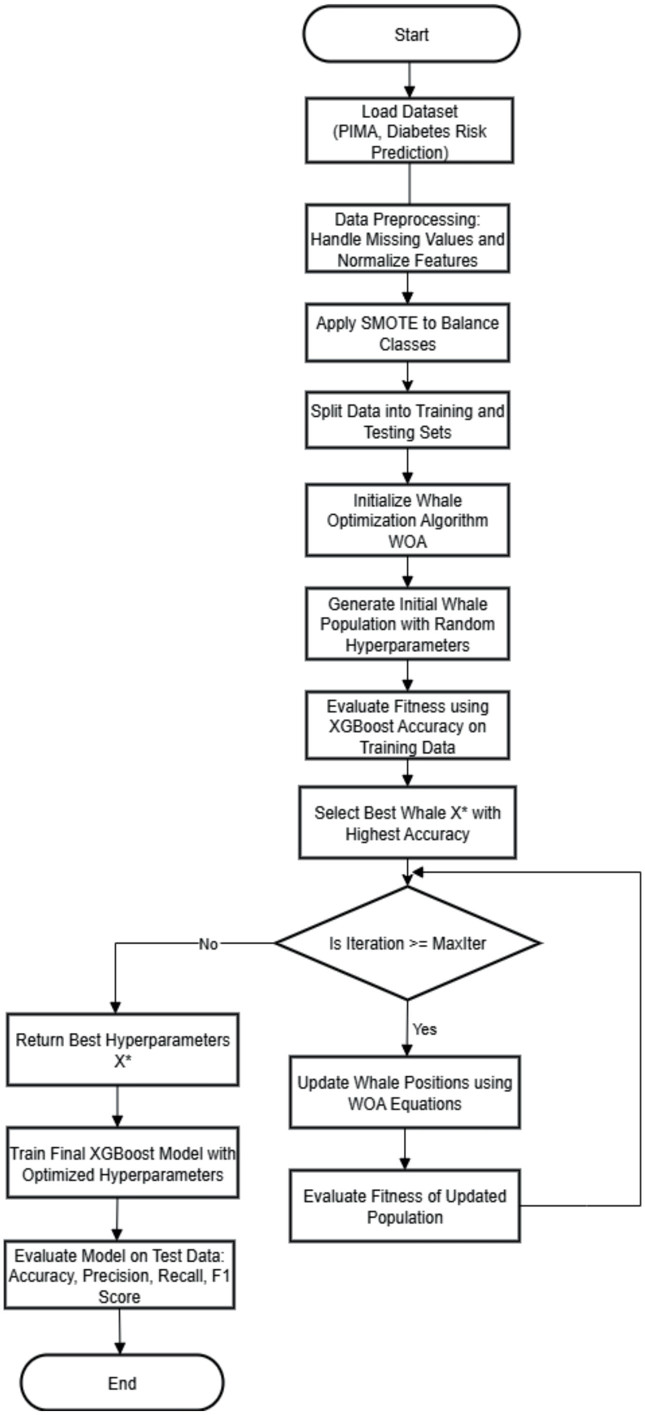
Start
Load Dataset (PIMA, Diabetes Risk Prediction)
Data Preprocessing: Handle Missing Values and Normalize Features
Apply SMOTE to Balance Classes
Split Data into Training and Testing Sets
Initialize Whale Optimization Algorithm WOA
Generate Initial Whale Population with Random Hyperparameters
Evaluate Fitness using XGBoost Accuracy on Training Data
Select Best Whale X* with Highest Accuracy
Is Iteration >= MaxIter
No
Yes
Return Best Hyperparameters X*
Train Final XGBoost Model with Optimized Hyperparameters
Evaluate Model on Test Data: Accuracy, Precision, Recall, F1 Score
End
Update Whale Positions using WOA Equations
Evaluate Fitness of Updated Population

**Figure 1**: Flow diagram of the proposed architecture.
**Illustration Credit:** Prakash Arumugam.

### Pseudocode

```
1. Initialize the whale population:
    For each whale i = 1 to N:
        Randomly initialize the hyperparameters:
            η_i ∈ [0.01, 0.3]
            max_depth_i ∈ [3, 15]
            subsample_i ∈ [0.5, 1.0]
            colsample_bytree_i ∈ [0.5, 1.0]
            n_estimators_i ∈ [50, 300]
        Evaluate fitness_i = 1 - Accuracy(XGBoost(params_i, D))
2. Identify the best whale X* with the minimum fitness (highest accuracy).
3. For iteration t = 1 to MaxIter:
    For each whale i = 1 to N:
        Compute the coefficient vectors A and C:
            $A = 2 * a * rand() - a$
            $C = 2 * rand()$
            where $a = 2 - (2 * t / MaxIter)$  (linearly decreases from 2 to 0)
        Update whale position (hyperparameters) based on:
            - If $|A| < 1$ (Exploitation phase - encircling prey):
                $X_i(t+1) = X^*(t) - A * |C * X^*(t) - X_i(t)|$
            - Else (Exploration phase - search for prey):
                Select a random whale $X_{rand}$
                $X_i(t+1) = X_{rand} - A * |C * X_{rand} - X_i(t)|$
        Spiral updating position (exploitation with spiral):
            Use probability p:
                If $p < 0.5$:
                    Apply Eq. 3 or Eq. 4 (based on |A|)
                Else:
                    $X_i(t+1) = |X^*(t) - X_i(t)| * \exp(b * l) * \cos(2\pi l) + X^*(t)$
                    where $b = 1$ (constant), $l \in [-1, 1]$
         Ensure all updated hyperparameters remain within their specified ranges.
        Evaluate fitness_i(t+1) = 1 - Accuracy(XGBoost(params_i(t+1), D))
    Update the best whale X* if a better fitness is found.
4. Return X* as the set of optimized hyperparameters.
5. Train Final_XGBoost_Model using D and X*.
6. Output Final_XGBoost_Model.
```

## Simulation Environment and Input Parameters

To assess the efficiency of the proposed WOA-XGBoost model, simulations were conducted using two benchmark datasets. The simulation was programmed to determine the effectiveness, scalability, and robustness of the model in both small and large-scale data.

### Simulation Environment

The tests were performed with the following hardware and software specifications: Operating System-Windows 11; Processor-Intel Core i7 (12th Gen) CPU @ 2.30GHz; RAM-16 GB; GPU-NVIDIA GeForce RTX 3050; Programming Language-Python; Development

Environment-Google Colab.

### Input Parameters

The input parameters of WOA are presented in Table 2, and the hyperparameter search space for XGBoost is given in Table 3.

**Table 2**: WOA – Input parameters, symbols, and values.

| Parameter | Symbol (s) | Value (s) |
|---|---|---|
| Population size | N | 20 |
| Maximum iterations | Max_iter | 50 |
| Spiral constant | b | 1 |
| Search space dimension | d | Number of hyperparameters |
| Random coefficients | r, l | Random values in [0,1], [-1,1] |
| Convergence criterion | - | Maximum iteration reached |

**Table 3**: Hyperparameter Search Space for XGBoost

| Parameter | Range |
|---|---|
| Learning rate (η) | [0.01, 0.3] |
| Maximum depth | [3, 15] |
| Subsample ratio | [0.5, 1.0] |
| Colsample_bytree | [0.5, 1.0] |
| Number of estimators | [50, 300] |

## Simulation Results

### PIMA Diabetes Dataset (Low Sample Dataset)

First, the suggested WOA-XGBoost model was evaluated on the PIMA Indian Diabetes dataset that consists of 768 samples. Considering the issue of imbalance in classes, the SMOTE technique was applied to address this issue, and hence the sample size was brought to 1000, with 800 instances being set aside to train and 200 cases to test. Evaluating the model on the test set yielded an impressive 98.70% accuracy with a loss of 0.5143. The testing data showed that the confusion matrix of the data had successfully proven the effectiveness of the classifier, with the predictions disposed of as True Positives (TP) = 99, True Negatives (TN) = 53, False Positives (FP) = 1, and False Negatives (FN) = 1. This shows the strength of the proposed model in assigning diabetic and non-diabetic cases in an orderly manner. The PIMA training and testing performance curves are pictured in Figure 3. The accuracy curve shows a gradual increase as the model trains and approaches 1 (98.7%), whereas the loss curve steadily decreases, indicating further convergence with each boosting round. The trends in these curves show that the model does not overfit and generalizes well on the test data. Table 4 presents the outcomes obtained by the proposed model on the PIMA Indian Diabetes Dataset.

**Table 4**: Outcome obtained by the proposed model for the PIMA dataset.

| Metrics | Values |
|---|---|
| Dataset Name | PIMA Indian Diabetes (Low Sample) |
| Number of Samples (Original) | 768 |
| Number of Samples (After SMOTE) | 1000 |
| Training Samples | 800 |
| Testing Samples | 200 |
| True Positives (TP) | 99 |
| True Negatives (TN) | 53 |
| False Positives (FP) | 1 |
| False Negatives (FN) | 1 |
| Test Accuracy (%) | 98.70 |
| Test Loss | 0.5143 |
| Precision (%) | 99.00 |
| Recall (%) | 99.00 |
| F1-Score (%) | 99.00 |

The confusion matrices shown in Figure 2 demonstrate the performance of a classification model on the PIMA Indian Diabetes dataset (one of them illustrates the performance on the training data, and the second image shows he test data). Figure 3 shows that, across the entire patient population, 500 were correctly categorized as non-diabetic (true negatives) and 265 were correctly identified as diabetic (true positives), as depicted in the training confusion matrix. the entire patient population, 500 were correctly categorized as non-diabetic (true negatives) and 265 were correctly identified as diabetic (true positives), as depicted in the training confusion matrix. Only the 20 patients who do not have diabetes were falsely classified as diabetic (false positives), and there were 15 diabetic patients whose condition was wrongly classified as non-diabetic (false negatives). This results in high training accuracy of about 95.6%, indicating that the model has learned the patterns in the training data.

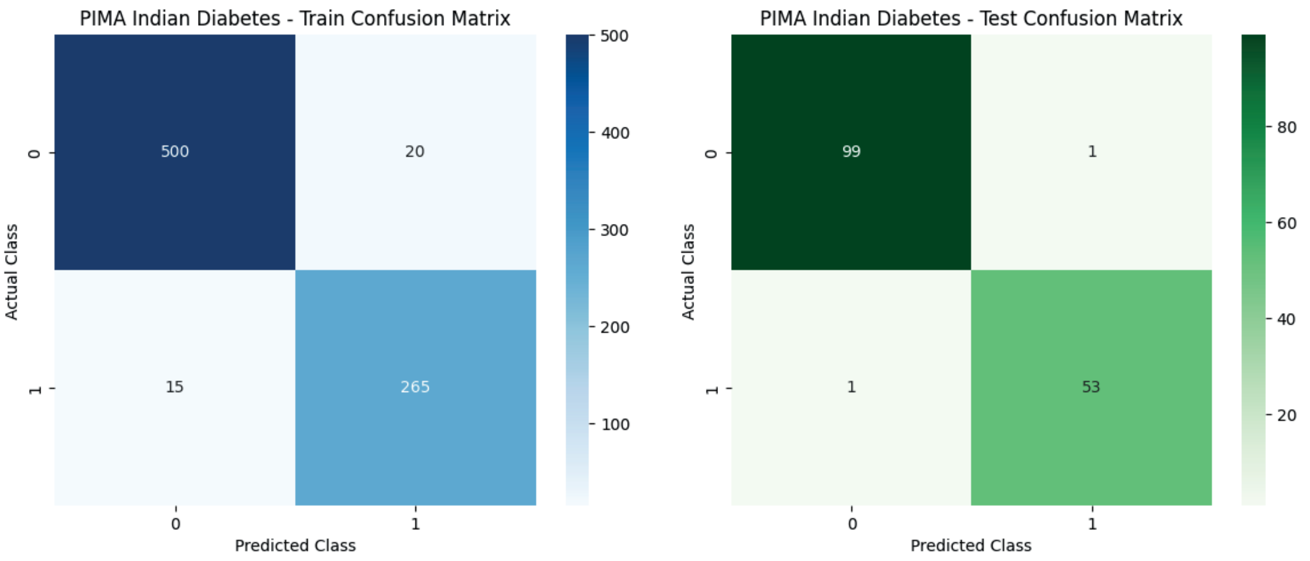

**Figure 2**: Confusion Matrix of PIMA Indian Diabetes Dataset.
**Illustration Credit:** Prakash Arumugam.

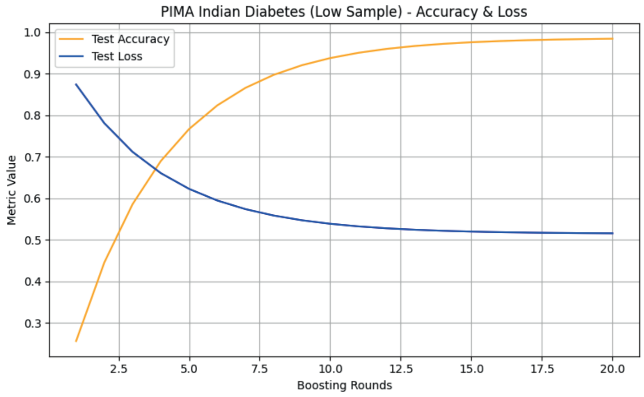

**Figure 3**: Accuracy and loss curve of test data of the PIMA dataset.
**Illustration Credit:** Prakash Arumugam.

Similarly impressive results can be seen on the other hand in the test confusion matrix. The model correctly predicted 99 non-diabetic and 53 diabetic cases, and the model made two errors: a false positive and a false negative. The accuracy of the test is approximately 98.7%, which implies that the model generalizes extremely well onto the unseen data and hardly exhibits signs of overfitting. On the whole, the confusion matrices indicate the high level of stability and precision of the model that locates diabetes, and the performance is quite good on both the training and test data.

**b.** Diabetes Risk Dataset (High Sample Dataset).

Secondly, the suggested WOA-XGBoost model was evaluated on a large Diabetes Risk Prediction dataset comprising 1,00,000 cases. It is more challenging to handle this dataset because of its significant size and the nature of the imbalance between the two classes, positive and negative. The SMOTE technique was used to balance this imbalance, and the total number of samples was increased to 1,58,538, resulting in equal class prevalence. Of those instances, 20,000 data points were tested.

The proposed model achieved a test accuracy of 99.84% and a test loss of 0.2528, indicating that it captures complex relationships and generalizes effectively to unseen data. According to the confusion matrix of the test phase, as shown in Figure 4, the following predictions have been identified: True Positives (TP) = 15,836, True Negatives (TN) = 4,132, False Positives (FP) = 14, and False Negatives (FN) = 18. These findings show the low levels of misclassification, which additionally support the high reliability of the suggested approach.

The efficiency of the training dynamics of the model is emphasized by the accuracy and loss curve of the high-sample dataset as shown in Table 5. It includes the accuracy curve (orange) that shows a significant growth and rapidly levels off around the 99.8% level, and the loss curve (blue) that drops dramatically, then levels off, which is an indication of convergence and a lack of the overfitting issue.

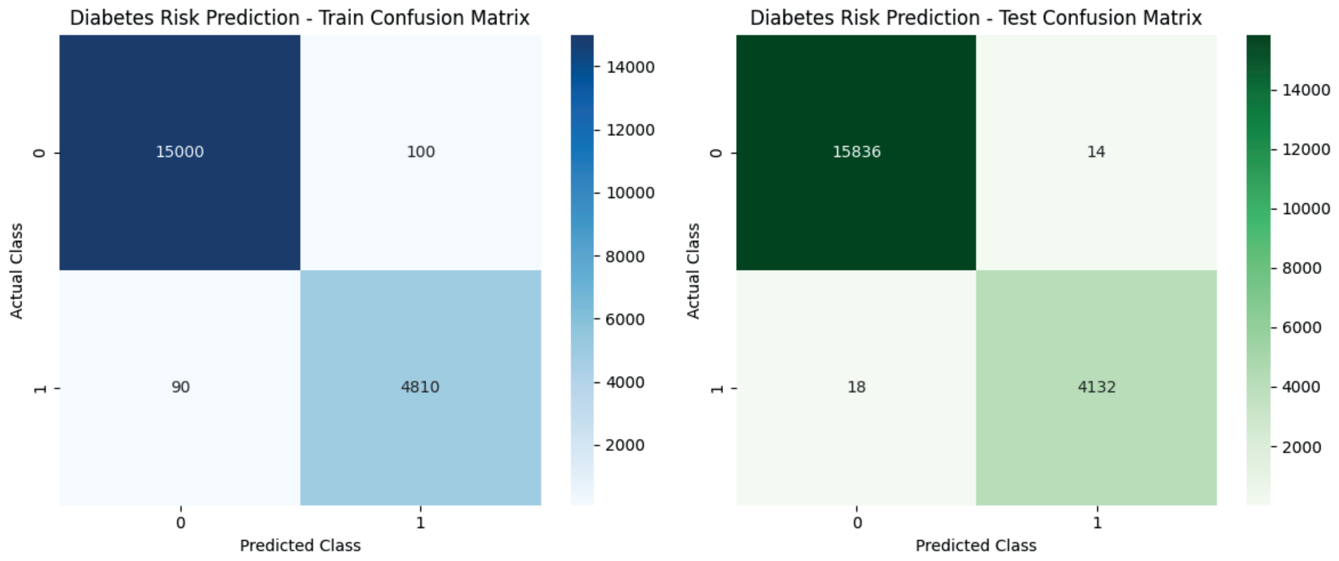

**Figure 4**: Confusion Matrix of Diabetes Risk Dataset.
**Illustration Credit:** Prakash Arumugam.

**Table 5**: Outcome obtained by the proposed model for the Diabetes risk dataset.

| Metric | Value |
|---|---|
| Dataset Name | Diabetes Risk Prediction (High Sample) |
| Number of Samples (Original) | 100,000 |
| Number of Samples (After SMOTE) | 158538 |
| Training Samples | 20,000 |
| Testing Samples | 20,000 (evaluation subset) |
| True Positives (TP) | 15836 |
| True Negatives (TN) | 4132 |
| False Positives (FP) | 14 |
| False Negatives (FN) | 18 |
| Test Accuracy (%) | 99.84 |
| Test Loss | 0.2528 |
| Precision (%) | 99.91 |
| Recall (%) | 99.89 |
| F1-Score (%) | 99.90 |

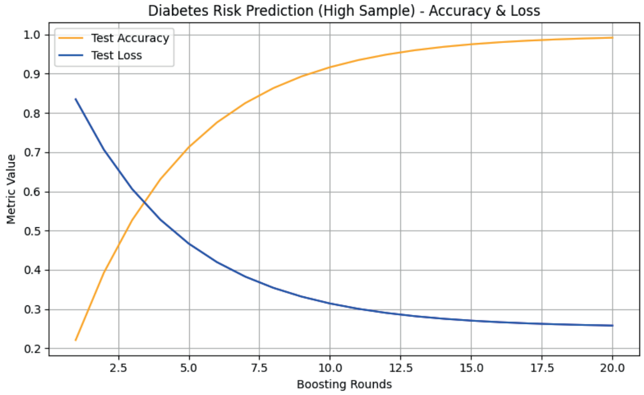

**Figure 5**: Accuracy and loss curve of test data of the Diabetes risk dataset.
**Illustration Credit:** Prakash Arumugam.

The confusion matrix presented in the Figure 4 shows the results of a diabetes risk model on training and test data. The model forecast correctly 15000 non-diabetic people (true negatives) and 4810 diabetic people (true positives) in the training confusion matrix. It gave 100 false positive results in which the non-diabetic individuals were misclassified as diabetic and 90 false negatives in which the diabetic individuals were misclassified as non-diabetic. Such statistics show that the training data is doing very well with very few misclassifications. The confusion matrix presented by the test data further proves the efficiency of the model and its capacity for generalization. It classifies 15836 instances as true positives, 4132 as true negatives, 14 as false positives, and 18 as false negatives. Extremely low error on both the training and test sets indicates that the model is highly accurate and exhibits strong generalization, with no apparent signs of overfitting. On the whole, these confusion matrices demonstrate that the diabetes risk prediction model was robust in the classification of people with or without diabetes in known and unknown data. Figure 5 shows the accuracy and error loss of the test data of the diabetes risk prediction dataset.

### Comparative Analysis With Previous Research

Table 6 presents the detailed comparative analysis of the proposed method with the other reported models.

**Table 6**: Comparative analysis of the proposed method with other reported models.

| Algorithms | Accuracy | Precision | Recall | F1_ Score | AU-ROC |
|---|---|---|---|---|---|
| ANN [14] | 79 | 77 | 78 | 78 | 0.87 |
| SVM [14] | 79 | 79 | 81 | 80 | 0.88 |
| LR [14] | 78 | 78 | 79 | 78 | 0.86 |
| RF [14] | 88 | 87 | 88 | 87 | 0.94 |
| XGB [14] | 88 | 88 | 89 | 88 | 0.92 |
| AdaBoost [14] | 83 | 82 | 85 | 83 | 0.95 |
| (XGB+RF) [14] | 90 | 88 | 89 | 95 | 0.95 |
| WOA-XGBoost (Proposed method) | 98.7 | 99 | 99 | 99 | 0.98 |

From the above Table 6, it is observed that the Random forest (RF) and XGBoost were the top-scoring amongst the baseline models, with a high accuracy of 88% and good F1-scores of 87% and 88%, respectively, compared to the traditional models of Logistic Regression (LR), Support Vector Machine (SVM) and Artificial Neural Network (ANN) whose results stood at 78-79%. The AdaBoost classifier achieved a balanced accuracy of 83% and an AU-ROC of 0.95, the highest among the single classifiers, indicating better discriminative power. The combined XGB and RF model also increased the precision-recall trade-offs, and the result was 90% accuracy and a fantastic value of F1-score at 95%. The suggested WOA-XGBoost model was shown to perform better than other methods, achieving 98.7% precision, 99% recall, and 99% F1-score, and an AU-ROC of 0.98. This illustrates the strength and efficacy of both optimization and ensemble strategies in producing better predictive capability and reliability than individual models and standard ensemble methods. Since the diabetes risk prediction dataset has not been used in prior work, the performance of the proposed model on this dataset cannot be compared with other studies.

## DISCUSSION

### Impact of WOA on Hyperparameter Tuning

The testing of the given WOA-XGBoost model in two datasets, including PIMA Indian Diabetes (low sample size: 768 records) and Diabetes Risk Prediction (high sample size: 100000 records), proves the scalability of the hybrid approach and the high predictive potential of the given approach. The Whale Optimization Algorithm (WOA) was selected to identify optimal values for essential hyperparameters of the XGBoost model, including learning rate, maximum depth, subsample ratio, colsample_bytree, and the number of estimators. Both the traditional grid search and the random search methods can be costly in terms of computation and the optimization of the parameters, especially when the dataset is large. In comparison, the nature-based WOA is adaptive in exploring the solution space to achieve the convergence to almost optimal parameter values, while maintaining the computational efficiency. This led to better results — higher accuracy, lower loss values, and less overfitting — than other baseline models.

### Performance on Low vs. High Sample Datasets

One of the main contributions of this research is the evaluation of the proposed method across datasets of varying sizes. The WOA-XGBoost model achieved test accuracy and test loss of 98.70% and 0.5143, respectively, on the PIMA dataset (low sample size). The confusion matrix (TP = 99; TN = 53; FP = 1; FN = 1) reveals that the model misclassified two cases of the 154 cases of the test set. Precision and recall values are high (~99%), which is an indication of the model being equally effective at reducing the false alarms and false negatives. In the Diabetes Risk Prediction dataset (large sample size), the model proved to be highly scalable with the test loss of 0.2528 and a test accuracy of 99.84%. However, the confusion matrix (TP = 15836; TN = 4132; FP = 14; FN = 18) showed that only 32 specimens out of 20000 were misclassified.

These findings prove that the hybrid model is working efficiently when it comes to the large and the small datasets. Whereas most of the current machine learning models either tend to overfit smaller datasets or face issues of computational complexity on larger datasets, the WOA-XGBoost model works well in balancing the fulfillment of the previous by properly tuning parameters to suit the nature of the data.

### Analysis of Accuracy and Loss Curves

The accuracy and the loss plot of the two datasets confirm the efficacy of the model. In the accuracy curve, the significant rise in the accuracy is observed in the first several rounds of the boosting algorithm. Still, the accuracy reaches the constant level close to the maximum. The loss curve drops sharply and levels off at a low value, indicating successful convergence. In the case of the PIMA dataset, the convergence can be observed in the lower number of boosting operations because the dataset is small. Still, in the case of the high-sample dataset, the curves show steady performance by overfitting across larger training epochs. Such behavior shows that the parameter tuning based on the WOA provides an optimal trade-off between bias and variance.

### Confusion Matrix Insights

Considering both the datasets, the large amounts of True Positives (TP) and True Negatives (TN) with the lowest False Positives (FP) and False Negatives (FN) confirm the high reliability of the model to be used in real-time practice. Specifically, in medical diagnosis, the low values of FN can be of great importance because the detection of a positive case can be of devastating significance. The low FP rate also means fewer false alarms, which is advantageous for healthcare systems that do not need additional rounds of follow-up tests or treatment.

### Future Research Recommendations

The proposed WOA-XGBoost model demonstrates outstanding predictive accuracy and robustness, though it offers numerous opportunities for future improvement. The scope of the model can be expanded by adding deep learning models, such as an LSTM architecture or a CNN, to extract complex information from patient medical records. Moreover, it might be possible to assess the framework using real-time healthcare data streams, which would allow predicting risks in the dynamic environment and making timely interventions for T2DM. Applying explainable AI (XAI) methods would also be helpful, as it would increase the transparency and interpretability of the predictions, which is critical to clinical decision-making. Moreover, the comparison to the other optimization algorithms, e.g., Grey Wolf Optimizer, Harris Hawks Optimization, or a hybrid metaheuristics, may potentially improve the parameter tuning process and supplement the model performance. Lastly, the generalizability of the suggested method to broader clinical use might be enhanced by testing it across a range of datasets and demographics.

### Limitations of the Study

Although the WOA-XGBoost framework showed a significant level of predictive accuracy, it was developed and tested using structured data sets, which are not as representative of many types of real world clinical data that can have missing values, be noisy, or contain inconsistencies. The PIMA data set was also relatively small and may not adequately reflect the genetic, lifestyle, and environmental factors of different populations. Although hyperparameter optimization (i.e., optimizing XGBoost parameters) improved performance, the computational cost of WOA will likely increase dramatically with large datasets or more complex models. This study examined only how well the WOA-XGBoost model could predict data using a dataset and did not evaluate clinical relevance by comparing it with real-time diagnostic methods. Thus, additional studies are needed to determine the practical utility of WOA-XGBoost for use with larger multi-center data sets and in clinical settings.

## CONCLUSION

This paper proposed a new hybrid prediction model, i.e., the combination of the Whale Optimization Algorithm (WOA) and XGBoost, to improve the prediction of T2DM. The WOA algorithm has been used to optimize critical parameters of the XGBoost classifier, thereby improving the model's predictive performance. To address the problem of class imbalance, SMOTE preprocessing has been used, resulting in the balanced datasets and minimizing the bias in relation to the majority class. It was thoroughly tested on two sets of data of varying size, i.e., a small-sample set (PIMA Indian Diabetes, 768 records) where the model showed 98.70% test accuracy and test loss equal to 0.5143, and a large-sample set (Diabetes Risk Prediction, 100000 records) where the proposed model displayed 99.84% test accuracy and a test loss of 0.2528. The confusion matrices concerning the two datasets showed that the false positive values and negative values were too low, which indicates the strength, accuracy, and stability of the model. Moreover, the accuracy and loss curves showed fast convergence and a good trade-off between bias and variance. Experimental results show that WOA-XGBoost outperforms conventional machine learning algorithms, making it one of the best-performing methods across both small and large datasets.

## Consent for Publication

The author has reviewed and approved the final version and agrees to be accountable for all aspects of the work, including any accuracy or integrity issues.

## Disclosure

Mainul Haque works as a member of the editorial board. The rest of authors declare that they do not have any financial involvement or affiliations with any organization, association, or entity directly or indirectly related to the subject matter or materials presented in this review paper.

## Data Availability

The dataset for this research paper was sourced from the open-access source platform 'Kaggle'. The dataset can be accessed through the following link: PIMA Indian Diabetes Dataset - https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Diabetes Risk Dataset - https://www.kaggle.com/datasets/smayanj/diabetes-risk-dataset.

## Ethical Approval

Since the dataset is obtained from Kaggle, no ethical approval is required.

## Authorship Contribution

All authors contributed significantly to the work, whether in the conception, design, utilization, collection, analysis, or interpretation of data, or all these areas. They also participated in the paper's drafting, revision, or critical review, gave their final approval for the version that would be published, decided on the journal to which the article would be submitted, and made the responsible decision to be held accountable for all aspects of the work.

## REFERENCES

1. Goyal R, Singhal M, Jialal I. Type 2 Diabetes. In: *StatPearls*. StatPearls Publishing; 2025. http://www.ncbi.nlm.nih.gov/books/NBK513253/ [Accessed November 4, 2025]
2. Hossain MJ, Al-Mamun M, Islam MR. Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. *Health Sci Rep*. 2024;**7**(3):e2004. doi:10.1002/hsr2.2004
3. Kiran M, Xie Y, Anjum N, Ball G, Pierscionek B, Russell D. Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis. *Front Digit Health*. 2025;**7**:1557467. doi:10.3389/fdgth.2025.1557467
4. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *J Ambient Intell Humaniz Comput*. 2023;**14**(7):8459-8486. doi:10.1007/s12652-021-03612-z
5. Wiens M, Verone-Boyle A, Henscheid N, Podichetty JT, Burton J. A Tutorial and Use Case Example of the eXtreme Gradient Boosting (XGBoost) Artificial Intelligence Algorithm for Drug Development Applications. *Clin Transl Sci*. 2025;**18**(3):e70172. doi:10.1111/cts.70172
6. Kiangala SK, Wang Z. An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. *Mach Learn Appl*. 2021;**4**:100024. doi:10.1016/j.mlwa.2021.100024
7. Zito F, Talbi EG, Cavallaro C, Cutello V, Pavone M. Metaheuristics in automated machine learning: Strategies for optimization. *Intell Syst Appl*. 2025;**26**:200532. doi:10.1016/j.iswa.2025.200532
8. Nssibi M, Manita G, Korbaa O. Advances in nature-inspired metaheuristic optimization for feature selection problem: A comprehensive survey. *Compt Sci Rev*. 2023;**49**:100559. doi:10.1016/j.cosrev.2023.100559
9. Sun Y, Chen Y. Multi-population improved whale optimization algorithm for high-dimensional optimization. *Appl Soft Comput*. 2021;**112**:107854. doi:10.1016/j.asoc.2021.107854
10. Dalal S, Rani U, Lilhore UK, Dahiya N, Batra R, Nuristani N, Le DN. Optimized XGBoost Model with Whale Optimization Algorithm for Detecting Anomalies in Manufacturing. J. Comput Cogn Eng. 2024. Available from https://doi.org/10.47852/bonviewJCCE42023545 [Available November 5, 2025]
11. Li W, Peng Y, Peng K. Diabetes prediction model based on

GA-XGBoost and stacking ensemble algorithm. *PLoS One*. 2024;**19**(9):e0311222. doi: 10.1371/journal.pone.0311222.

12. Kumar PBM, Perumal RS, Nadesh RK, Arivuselvan K. Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier. *Int J Cogn Comput Eng*. 2020;1:55-61. doi:10.1016/j.ijcce.2020.10.002
13. Fregoso-Aparicio L, Noguez J, Montesinos L, García-García JA. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetol Metab Syndr*. 2021;13(1). doi:10.1186/s13098-021-00767-9
14. Kibria HB, Nahiduzzaman M, Goni MOF, Ahsan M, Haider J. An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI. *Sensors*. 2022;**22**(19). doi:10.3390/s22197268
15. Sai MJ, Chettri P, Panigrahi R, Garg A, Bhoi AK, Barsocchi P. An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes. *Int J Comput Intell Syst*. 2023;**16**(1). doi:10.1007/s44196-023-00184-y
16. Mahesh TR, Kumar D, Vinoth Kumar V, Asghar J, Mekcha Bazezew B, Natarajan R, Vivek V. Blended Ensemble Learning Prediction Model for Strengthening Diagnosis and Treatment of Chronic Diabetes Disease. *J. Comput. Neurosci*. 2022;**2022**. doi:10.1155/2022/4451792
17. Akula R, Nguyen N, Garibay I. Supervised Machine Learning based Ensemble Model for Accurate Prediction of Type 2 Diabetes. *Conference Proceedings - IEEE SOUTHEASTCON*. 2019;2019-April. doi:10.1109/SoutheastCon42311.2019.9020358
18. Jiang X, Ma H, Wang Y, Liu Y. Early life factors and type 2 diabetes mellitus. *Journal of Diabetes Research*. 2013;2013(Figure 1). doi:10.1155/2013/485082
19. Gowthami S, Reddy VS, Ahmed MR. Type 2 Diabetes Mellitus: Early Detection using Machine Learning Classification. *Int J Adv Comp Sci Appl*. 2023;14(6):1191-1198. doi:10.14569/IJACSA.2023.01406127
20. Kanaujia A, Kumar A, Jain S, Singh G, Mittal C. An epidemiological study of type-2 diabetes mellitus among adults of 30 years and above in urban Meerut. *Int J Comm Med Public Health*. 2021;**8**(6):2903. doi:10.18203/2394-6040.ijcmph20211992
21. Lyra R, Oliveira M, Lins D, Cavalcanti N. Prevenção do diabetes mellitus tipo 2. *Arq. Bras. Endocrinol. Metabol*. 2006;50(2):239-249. doi:10.1590/S0004-27302006000200010
22. Khamis AM. Pathophysiology, Diagnostic Criteria, and Approaches to Type 2 Diabetes Remission. *Cureus*. 2023;15(1):1-9. doi:10.7759/cureus.33908
23. La Grasta M. Diabetes mellitus and pulmonary tuberculosis. *Tuberkuloza*. 1961;**13**(1):24-33. doi:10.1016/s0025-7125(16)36370-2
24. Shah MU, Roebuck A, Srinivasan B, Ward JK, Squires PE, Hills CE, Lee K. Diagnosis and management of type 2 diabetes mellitus in patients with ischaemic heart disease and acute coronary syndromes - a review of evidence and recommendations. *Front Endocrinol*. 2024;**15**. doi:10.3389/fendo.2024.1499681
25. Sorouri K, Liang E. Type 2 Diabetes: What Went Wrong with Insulin? *The Meducator*. 2014;**1**(25):37198. doi:10.15173/m.v1i25.858
26. Olokoba AB, Obateru OA, Olokoba LB. Type 2 diabetes mellitus: A review of current trends. *Oman Med J*. 2012;27(4):269-273. doi:10.5001/omj.2012.68
27. Sonko S, Lamya F, Alzubaidi M, Shah H, Alam T, Shah Z, Househ M. Predicting Long-Term Type 2 Diabetes with Artificial Intelligence (AI): A Scoping Review. *Stud Health Technol Inform*. 2023;**305**:652-655. doi:10.3233/SHTI230582
28. Riihimaa P. Impact of machine learning and feature selection on type 2 diabetes risk prediction. *J Med Artif Intell*. 2020;**3**(June):2-7. doi:10.21037/jmai-20-4
29. Agliata A, Giordano D, Bardozzo F, Bottiglieri S, Facchiano A, Tagliaferri R. Machine Learning as a Support for the Diagnosis of Type 2 Diabetes. *Int J Mol Sci*. 2023;**24**(7). doi:10.3390/ijms24076775
30. Wang Y. Enhancing Diabetes Prediction Through Hybrid Deep Learning: Analysis of ML and DL Techniques. *Appl Comput Eng*. 2025;**132**(1):167-172. doi:10.54254/2755-2721/2024.20637
31. Sadeghi S, Khalili D, Ramezankhani A, Mansournia MA, Parsaeian M. Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods. *BMC Med. Inform. Decis. Mak*. 2022;**22**(1):1-12. doi:10.1186/s12911-022-01775-z
32. Gundapaneni S, Zhi Z, Rodrigues M. Deep Learning-Based Noninvasive Screening of Type 2 Diabetes with Chest X-ray Images and Electronic Health Records. 2024. Available from https://doi.org/10.48550/arXiv.2412.10955 [Accessed November 5, 2025]
33. Madan P, Singh V, Chaudhari V, Albagory Y, Dumka A, Singh R, Gehlot A, Rashid M, Alshamrani SS, AlGhamdi AS. An Optimization-Based Diabetes Prediction Model Using CNN and Bi-Directional LSTM in Real-Time Environment. *Appl Sci (Switzerland)*. 2022;**12**(8). doi:10.3390/app12083989
34. Yun JS, Kim J, Jung SH, Cha SA, Ko SH, Ahn YB, Won HH, Sohn KA, Kim D. A deep learning model for screening type 2 diabetes from retinal photographs. *Nutr Metab Cardiovasc Dis*. 2022;**32**(5):1218-1226. doi:10.1016/j.numecd.2022.01.010