Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MO-AIWOA-XAI: A Multi-Objective Explainable Diabetes Risk Prediction integrating Feature Optimization and Cross-Dataset Validation

Sumalatha U
Mangalore Institute of Technology and Engineering
krishnaprakasha k
kkp.prakash@manipal.edu
Manipal Academy of Higher Education

Article

Keywords: Diabetes Risk Prediction, Explainable Artificial Intelligence, SHAP, LIME, DiCE, Adaptive Improved Whale Optimization Algorithm, Feature Selection, Ensemble Learning, Cross-Dataset Validation, Clinical Decision Support.

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10356229/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

MO-AIWOA-XAI: A Multi-Objective Explainable Diabetes Risk Prediction integrating Feature Optimization and Cross-Dataset Validation

SUMALATHA U [1], KRISHNA PRAKASHA K. [2] (SENIOR MEMBER, IEEE)

[1]Mangalore Institute of Technology and Engineering, Moodabidri, Karnataka 574225, India, (e-mail: suma0506@gmail.com)

[2]Manipal Institute of Technology (MIT), Manipal Academy of Higher Education (MAHE), Manipal, Karnataka 576104, India, (e-mail: kkp.prakash@manipal.edu)

Corresponding Author: Krishna Prakasha (kkp.prakash@manipal.edu)

ABSTRACT Clinical decision-making for diabetes management requires prediction models that are both accurate and interpretable. Although machine learning techniques have demonstrated promising performance in diabetes risk prediction, their adoption in healthcare is often constrained by limited transparency and feature redundancy. To address these challenges, the study proposes a Multi-Objective Adaptive Improved Whale Optimization Explainable Artificial Intelligence (MO-AIWOA-XAI) framework for diabetes risk prediction. The proposed framework integrates a multi-objective Adaptive Improved Whale Optimization Algorithm (AIWOA) for feature selection, an Optuna-optimized Extra Trees classifier, and explainable artificial intelligence techniques including SHAP, LIME, and DiCE. The feature selection strategy simultaneously maximizes predictive performance while minimizing feature dimensionality, enabling the identification of clinically relevant risk factors. The framework was evaluated using the PIMA Indians Diabetes Dataset and the Early Stage Diabetes Risk Prediction Dataset. Data preprocessing involved KNN imputation, feature engineering, Min-Max normalization, outlier treatment, and SMOTE-Tomek class balancing. Extensive experiments, including comparisons with PSO, GA, and WOA, cross-dataset validation, 10-fold cross-validation, calibration analysis, and decision curve analysis, were performed to assess model robustness. The proposed framework achieved an accuracy of 83.76%, AUC of 0.9004, F1-score of 0.84, MCC of 0.6757, and PR-AUC of 0.9068 on the PIMA dataset, while attaining 95.19% accuracy and 0.992 AUC on the Early Stage Diabetes dataset. In addition, the framework produced stable and clinically interpretable explanations through SHAP, LIME, and counterfactual analysis. The results demonstrate that the proposed MO-AIWOA-XAI framework effectively combines feature optimization, predictive modeling, and explainable artificial intelligence to provide a reliable and clinically interpretable solution for diabetes risk assessment and decision support.

INDEX TERMS Diabetes Risk Prediction, Explainable Artificial Intelligence, SHAP, LIME, DiCE, Adaptive Improved Whale Optimization Algorithm, Feature Selection, Ensemble Learning, Cross-Dataset Validation, Clinical Decision Support.

I. INTRODUCTION

Diabetes mellitus is among the most common chronic conditions worldwide and is considered an important public health problem owing to its rising prevalence, complications, and cost of care [1]. The International Diabetes Federation (IDF) reports that hundreds of millions of people are now suffering from diabetes, and the number is set to increase considerably in future decades. Diabetic patients suffer from serious complications, such as cardiovascular disease, renal failure, neuropathy, retinopathy, and early death. This means that it is important to identify people at risk of developing diabetes to implement effective preventive measures [2].

Due to the fast development of Artificial Intelligence and Machine Learning, data-oriented techniques have been gain-

ing prominence in the area of healthcare analytics and prediction of diseases. Machine learning can examine correlations between various clinical, demographic and lifestyle parameters and discover some patterns which cannot be detected using traditional statistical methods [3]. There are different supervised machine learning methods, such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, Gradient Boosting, and ensemble learning techniques, used for the prediction of diabetes [4], [5]. Still, the performance of all of them relies on how good features are selected for them. High-dimensional medical datasets tend to consist of irrelevant and noisy attributes which complicate the process of computations and may hinder the results of predictions.

In order to overcome these issues, feature selection has been seen as a vital step in the predictive models. In terms of various optimization techniques, there are better capabilities in finding the informative features through the efficient search of large search spaces using meta-heuristic algorithms. Some of the commonly used optimization algorithms include the Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Grey Wolf Optimization (GWO), and Whale Optimization Algorithm (WOA). The Whale Optimization Algorithm, based on the bubble-net foraging method of humpback whales, has achieved good outcomes for exploration and exploitation in the process of optimization. However, traditional WOA experiences premature convergence and poor diversity in optimization of complex medical problems [6].

In order to address these shortcomings, the current research work suggests an approach known as the Multi-Objective Adaptive Improved Whale Optimization Algorithm (MO-AIWOA) [7], which is used for feature selection in diabetes risk prediction. The main idea of the suggested approach is the simultaneous optimization of classification performance and the size of the selected feature subset, thus leading to the identification of a small number of clinical meaningful features with high classification performance [8].

However, despite the great potential of such algorithms and their predictive performance, their practical use is limited due to the fact that many of them have black-box nature. In order to increase transparency of these models and trust in their results, the field of Explainable Artificial Intelligence (XAI) has become one of the most relevant research topics nowadays [9].

Among the existing techniques for XAI, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) are popular techniques for explaining the behavior of a model. SHAP provides explanations that are global as well as local and are based on the principles of cooperative game theory to quantify the impact of each feature on the ultimate prediction. LIME constructs locally interpretable surrogate models to explain predictions. Recently, Diverse Counterfactual Explanations (DiCE) have been proposed as an action-able technique which finds minimal necessary changes in order to change the prediction of a model. Using SHAP, LIME and DiCE provides complimentary information about the model behavior enabling the clinicians to find the overall importance of features, patient-wise predictions and action-able intervention [10], [11], [12].

While many studies have been conducted in predicting diabetes using machine learning, most of these approaches are only concerned with the accuracy of prediction without considering the explainability of the model, its feature optimization and robustness analysis. In addition, there has been very less effort towards the use of optimization-based feature selection, explainable AI, cross-dataset validation, calibration analysis, decision curve analysis and counterfactual reasoning in one single study.

In order to fill these gaps, the research proposes MO-AIWOA-XAI, a framework for diabetes risk prediction that employs multi-objective feature optimization, ensemble learning, and advanced methods of explainability. In particular, the framework uses MO-AIWOA for the selection of optimal feature subsets, Optuna-optimized Extra Trees classifier for diabetes prediction, and SHAP, LIME, and DiCE for global, local, and counterfactual explanations, respectively. Additionally, the framework employs KNN-based missing value imputation, feature engineering, SMOTE-Tomek class balancing, calibration analysis, decision curve analysis, cross-validation, and cross-dataset validation.

The framework was validated on PIMA Indians Diabetes Dataset and Early Stage Diabetes Risk Prediction Dataset. As can be seen from the experimental results, the model proposed by the research achieves 83.76% accuracy, AUC = 0.9004, F1-score = 0.84, MCC = 0.6757, and PR-AUC = 0.9068 for the PIMA dataset and 95.19% accuracy and AUC = 0.992 for the Early Stage Diabetes dataset. Besides, the model shows high explanation stability, as demonstrated by its 1.0 SHAP stability score and provides clinically meaningful counterfactual explanations. The key contributions of the work are:

1) Proposal of MO-AIWOA (Multi-Objective Adaptive Improved Whale Optimization Algorithm) for efficient feature selection in case of diabetes diagnosis.
2) Integration of MO-AIWOA with the Extra Trees classifier optimized with Optuna for diabetes risk estimation.
3) Application of SHAP, LIME, and DiCE algorithms for generation of global, local, and counterfactual explanations.
4) Thorough validation of the framework with cross-dataset evaluation, 10-fold cross-validation, calibration analysis, decision curve analysis, bootstrap confidence interval analysis, MCC, Kappa, and PR-AUC.
5) Design of the clinically interpretable and computationally efficient framework suitable for diabetes risk screening tasks.

The structure of the paper is as follows. Related works are discussed in Section II. Section III details the MO-AIWOA-XAI framework, the datasets and experiments settings, Results and discussion. Conclusion and future research directions are provided in Section IV.

II. RELATED WORK

A. MACHINE LEARNING-BASED DIABETES RISK PREDICTION

Machine learning algorithms have been largely utilized in predicting diabetes cases owing to the capability of these algorithms to examine the complicated associations that exist between various demographic, clinical, and lifestyle factors. Classification methods like Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), and Naive Bayes have been employed widely for diagnosing diabetes. The application of these methods has yielded favorable predictions on benchmark databases such as the PIMA Indians Diabetes Database and Early Stage Diabetes Risk Prediction Database [13].

Xu et al. proposed a machine learning framework for Type 2 Diabetes Mellitus (T2DM) prediction using the NHANES 2021–2023 dataset comprising 6,337 participants. The study employed Monte Carlo multiple imputation, Principal Component Analysis (PCA), Random Forest (RF), Recursive Feature Elimination (RFE), and Adaptive Synthetic Sampling (ADASYN) for data preprocessing and feature selection while evaluating seven machine learning algorithms. The AdaBoost model achieved the best performance with an AUC of 0.85, an accuracy of 71%, and an F1-score of 0.71, identifying 24 significant multilevel risk factors. The framework demonstrated improved prediction by integrating multidimensional health ecological factors; however, its reliance on a single dataset and the absence of explainable AI and external validation may limit its generalizability and clinical applicability [14].

Aich et al. proposed a supervised feature selection approach based on a Gumbel copula upper-tail concordance score for diabetes risk prediction. The method was evaluated on the CDC ($n \approx 253,000$) and PIMA ($n = 768$) diabetes datasets and compared with Mutual Information, mRMR, ReliefF, and L1/Elastic-Net feature selection techniques using multiple classifiers. Experimental results showed that the proposed approach reduced the feature space by 52%, achieved competitive ROC–AUC performance, and consistently identified clinically relevant predictors. The method offers an efficient and interpretable feature screening mechanism; however, it primarily focuses on feature ranking and does not incorporate explainable AI techniques or external clinical validation [15].

Yang et al. developed a machine learning framework to predict diabetes risk among individuals with Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD) using a dataset of 6,310 participants collected between 2007 and 2018. The study employed Random Forest (RF), Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine Recursive Feature Elimination (SVM-RFE) for feature selection and evaluated nine machine learning models. The Light Gradient Boosting Machine (LightGBM) achieved the best predictive performance, while SHAP analysis identified urinary albumin, age, total cholesterol, and lipid accumulation product as the most influential predictors. The proposed framework enhanced diabetes risk prediction and model interpretability; however, its applicability is limited to the MASLD population and lacks external validation to assess its generalizability [16].

Srinithi et al. proposed a machine learning-based diabetes risk prediction and personalized diet recommendation system to facilitate early diabetes detection and promote healthy lifestyle modifications. The study employed the PIMA Diabetes dataset and utilized the Random Forest algorithm to predict diabetes risk using clinical parameters such as glucose level, body mass index (BMI), age, blood pressure, and insulin level. Based on the prediction results, the system generated personalized region-specific diet recommendations and healthy food alternatives. The proposed framework supported early diabetes prevention by integrating risk prediction with dietary guidance; however, it relied on a single machine learning algorithm and did not incorporate explainable AI techniques or extensive external validation [17].

Mohankumar et al. proposed a hybrid deep learning framework for early-stage diabetes risk prediction by integrating Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The study utilized the Early-Stage Diabetes Risk Prediction Dataset from Kaggle, comprising 520 patient records with clinical and demographic attributes. The CNN extracted representative features, while the LSTM captured temporal dependencies, achieving an accuracy of 99.42% and demonstrating strong generalization with minimal overfitting. The proposed framework significantly improved prediction performance; however, it lacked explainable AI mechanisms and validation on larger, diverse clinical datasets, which may limit its interpretability and generalizability [18].

Zhao et al. developed a deep learning-based prognostic framework to predict chronic kidney disease (CKD) progression in patients with Type 2 Diabetes Mellitus (T2DM) using a 17-year longitudinal electronic health record dataset comprising 569,680 individuals from 165 healthcare facilities in Hong Kong. The model integrated clinical, biochemical, and prescription history data, while SHAP was employed to interpret key predictive factors. The proposed framework achieved AUC values of 87.1%, 85.3%, and 84.7% for 2-, 5-, and 10-year predictions, respectively, and demonstrated good generalizability through external validation on the UK Biobank and CHARLS cohorts. The study provided an interpretable and scalable approach for personalized CKD risk prediction; however, it focused specifically on CKD progression in T2DM patients and may require further validation across broader and more diverse populations [19].

Bernardini et al. proposed a machine learning-based clinical decision support system for predicting hepatic fibrosis risk in patients with non-alcoholic fatty liver disease (NAFLD) using routine laboratory tests and electronic health records collected over 15 years. The framework was developed using two cohorts comprising 12,960 and 25,478 patients and incorporated established clinical indices, including

the Hepatic Steatosis Index (HSI) and Fibrosis-4 (FIB-4) index, with a minimal set of predictors. The proposed model achieved the best performance with an AUC of 92.97%, PRAUC of 75.44%, and sensitivity of 79.63%, demonstrating effective fibrosis risk screening. The framework provides a cost-effective and practical solution for clinical decision support; however, it is specifically designed for NAFLD-related fibrosis prediction and lacks explainable AI techniques and external validation across diverse populations [20].

Recent studies have demonstrated that machine learning techniques significantly improve diabetes risk prediction by effectively modeling complex relationships among clinical and demographic risk factors. Ensemble learning algorithms, including Random Forest, Gradient Boosting, XGBoost, LightGBM, AdaBoost, and Random Survival Forest, have consistently shown superior predictive performance across diverse diabetes datasets. Several studies have also incorporated feature selection methods and explainable artificial intelligence techniques to identify clinically relevant predictors and improve model transparency. However, most existing approaches primarily emphasize prediction accuracy while paying limited attention to simultaneous feature optimization, comprehensive interpretability, cross-dataset validation, and clinical applicability. These limitations highlight the need for an integrated machine learning framework that combines efficient feature selection, robust predictive modeling, explainable artificial intelligence, and extensive validation for reliable diabetes risk prediction.

B. METAHEURISTIC OPTIMIZATION FOR FEATURE SELECTION

Selection of features is an important step in enhancing the performance of predictions and minimizing the complexity of models in health care. There have been various metaheuristic optimization techniques developed that aid in determining the optimal subsets of features.

Some of the optimization techniques employed are Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Grey Wolf Optimization (GWO), Firefly Algorithm (FA), and Whale Optimization Algorithm (WOA). Of these, WOA has been widely used due to its efficient combination of exploration and exploitation based on the bubble-net approach of humpback whales.

There have been several attempts at incorporating optimization techniques along with the machine learning classification technique for better prediction of diabetes. However, the traditional approaches for feature selection usually focus on achieving higher classification accuracy while not considering the aspect of feature reduction and interpretation. Moreover, most of the optimization techniques exhibit an issue of premature convergence with lack of diversity in later iterations.

Özkan et al. proposed an optimization-driven hybrid machine learning framework for automatic brain tumor classification using MRI images. The study utilized a dataset of 834 MRI images and employed YOLOv11 for tumor localization, entropy-based feature extraction, and the Superb Fairy-Wren Optimization Algorithm (SFOA) for feature selection, with comparisons against PSO, HHO, and PO. The proposed SFOA-kNN model reduced the feature set from 39 to 5 features while achieving 99.20% classification accuracy, outperforming the comparative optimization methods and SVM classifier. The framework improved diagnostic accuracy and computational efficiency; however, it was evaluated on a single MRI dataset and did not incorporate explainable artificial intelligence techniques to enhance clinical interpretability [21].

Francis et al. presented a systematic literature review on metaheuristic-based cyberattack detection techniques for the Industrial Internet of Things (IIoT), analyzing 36 recent studies. The review examined machine learning and deep learning frameworks integrated with metaheuristic feature selection (MFS) algorithms and categorized them into Swarm Intelligence (SI), Evolutionary Algorithms (EA), Physics-Based (PHY), and Human-Behavior-Inspired (HBI) approaches. The analysis revealed that SI-based MFS was the most widely adopted approach, appearing in 25 of the 36 studies, while hybrid MFS frameworks achieved improved detection accuracy on benchmark datasets. The review provides a comprehensive overview of existing methodologies and identifies research gaps; however, it is limited to qualitative analysis and does not experimentally validate the reported techniques [22].

Mudassar et al. proposed a hybrid machine learning framework for short-term wind power forecasting by integrating Grey Wolf Optimization (GWO) for feature selection with a Double Attention Gated Recurrent Unit (DAGRU) model. The study evaluated the proposed framework using data from seven European wind farms and the ELIA dataset, where GWO combined with XGBoost identified the most informative features before time-series prediction. Experimental results demonstrated that the GWO-DAGRU model achieved superior forecasting accuracy and robustness compared with existing methods, with statistical significance validated using the Wilcoxon signed-rank test. The framework effectively improved feature selection and prediction performance; however, it was developed specifically for wind power forecasting and its applicability to other domains requires further investigation [23].

Ziane et al. proposed a hybrid metaheuristic feature selection framework for diabetes prediction by combining Particle Swarm Optimization (PSO) and Grey Wolf Optimizer (GWO) with an adaptive fitness function. The study evaluated the proposed approach on the Pima Indians, Frankfurt Hospital, and Iraq diabetes datasets using Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (kNN), and Logistic Regression classifiers with stratified five-fold cross-validation. The optimized Random Forest classifier achieved a mean accuracy of 84.74%, while the hybrid feature selector reduced the feature set by approximately 26% without compromising predictive performance. The proposed framework improved classification accuracy

and model interpretability through effective feature selection; however, it did not incorporate explainable AI techniques or external validation to assess clinical generalizability [24].

Sarker et al. proposed a metaheuristic-based machine learning framework for early-stage diabetes prediction by integrating the Fox Optimizer (FOX), Honey Badger Algorithm (HBA), and Tuna Swarm Optimization (TSO) with a Random Forest classifier. The study utilized the Early Stage Diabetes Risk Prediction dataset comprising 520 subjects with 16 predictors and employed SHAP to interpret feature importance and individual predictions. Among the proposed models, TSO_RF achieved the best performance with 100% accuracy without cross-validation and a mean 10-fold cross-validation accuracy of 98.14%, while reducing the feature set to 14 attributes. The proposed framework improved prediction accuracy, feature reduction, and model interpretability; however, it was evaluated on a single dataset and lacked external validation to demonstrate its generalizability in diverse clinical settings [25].

Adamu et al. proposed a hybrid metaheuristic feature selection framework, WaOA-MRFO, for cardiovascular disease prediction by combining the Walrus Optimization Algorithm (WaOA) and Manta Ray Foraging Optimization (MRFO) with adaptive diversity control. The study evaluated the proposed model on the Cleveland Heart Disease dataset using nested 5×5 cross-validation and further validated it on an independent three-site cohort. WaOA-MRFO achieved competitive predictive performance while reducing the feature set by approximately 67%, identifying a stable four-feature subset with strong cross-institutional transferability and enhanced process-level interpretability through EvoMapX analysis. The proposed framework improved feature selection transparency and reproducibility; however, it was developed specifically for cardiovascular disease prediction and its applicability to other medical domains requires further investigation [26].

Sokoot et al. proposed a hybrid machine learning framework for early diabetes diagnosis by integrating Borderline Synthetic Minority Oversampling Technique (BSMOTE), k-Nearest Neighbors (k-NN) imputation, the Giza Pyramids Construction (GPC) algorithm, and an ensemble classifier comprising k-NN, Support Vector Machine (SVM), Random Forest (RF), and Extreme Learning Machine (ELM). The proposed framework was evaluated on the PIMA Indians Diabetes Dataset (PIDD), where BSMOTE addressed class imbalance, k-NN handled missing values, and majority voting combined the predictions of multiple classifiers. Experimental results achieved prediction accuracies ranging from 91% to 93.9%, demonstrating improved performance over existing approaches. The framework effectively handled imbalanced and incomplete datasets while improving diagnostic accuracy; however, it relied on a single benchmark dataset and did not incorporate explainable artificial intelligence techniques or external validation to assess model generalizability [27].

Middha et al. proposed a hybrid optimization and deep learning framework for Type 2 Diabetes Mellitus (T2DM) prediction by integrating Yeo–Johnson transformation, Jaya-Dingo Optimization Algorithm (Jaya-DOA) for feature selection, and a hybrid Rider Neural Network (RideNN) and Deep Residual Network (DRN) classifier. The proposed framework was evaluated using the PIMA Indians Diabetes dataset and a gene expression dataset, where data augmentation and JCMRO-based classifier optimization were employed to improve predictive performance. Experimental results achieved a testing accuracy of 91% on both datasets, demonstrating the effectiveness of the hybrid optimization strategy. The proposed framework enhanced feature selection and classification performance; however, it involved a computationally complex architecture and lacked explainable artificial intelligence techniques and extensive external clinical validation [28].

Anam et al. proposed a diabetes prediction framework by integrating a Random Forest Classifier (RFC) with Binary Grey Wolf Optimization-based Feature Selection (BGWO-FS) and hyperparameter tuning. The proposed approach was evaluated using two independent diabetes datasets with different clinical characteristics, and SHAP was employed to interpret the selected clinical features. Experimental results achieved test accuracies of 78.2% and 78.8% with F1-scores of 75.3% and 70.0%, demonstrating improved predictive performance, computational efficiency, and model stability compared with GA-FS- and BPSO-FS-based approaches. The framework enhanced feature selection, interpretability, and generalizability; however, its predictive performance remained moderate, indicating the need for further optimization and validation on larger multicenter clinical datasets [29].

Recent studies have shown that machine learning combined with metaheuristic optimization significantly improves diabetes risk prediction by enhancing feature selection and classification performance. Algorithms such as PSO, GWO, BGWO, FOX, HBA, and TSO have been integrated with classifiers including Random Forest, LightGBM, SVM, and k-NN to identify informative features while reducing redundancy. Although some studies employ SHAP to improve interpretability, most focus primarily on predictive accuracy using a single dataset, with limited attention to multi-objective optimization, comprehensive explainability, cross-dataset validation, and clinical applicability.

C. EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR DIABETIS PREDICTION

Even though machine learning algorithms are capable of attaining very accurate predictions, their usage in the domain of healthcare is hindered because of their lack of interpretability. The decision-making process in clinical practice demands reliable and understandable predictions which should be validated by healthcare practitioners.

Explainable Artificial Intelligence (XAI) represents a promising area of research for solving these problems. SHapley Additive exPlanations (SHAP) help to obtain global and local explanations as they measure the effect that features

have on model predictions through cooperative game theory approaches. Local Interpretable Model-Agnostic Explanations (LIME) build surrogate models to interpret predictions locally. Lately, Diverse Counterfactual Explanations (DiCE) were developed to discover actionable explanations of what changes should be done to change the prediction results.

Several works have used SHAP and LIME in their models for explaining cardiovascular disease, diabetes, cancer, and chronic disease prediction models. However, most of the models proposed for the prediction of diabetes are limited to using SHAP and LIME as post-hoc techniques without considering the evaluation of the explanation consistency and stability. Moreover, SHAP, LIME, and counterfactual explanation have not been combined in diabetes prediction models.

Ahmed et al. proposed an explainable machine learning framework for personalized post-prandial glucose excursion (PPGE) prediction using continuous glucose monitoring (CGM) data from the CGMacros cohort. The study employed a LightGBM regression model with participant-grouped cross-validation and SHAP-based interpretability, utilizing 23,519 meal–CGM pairs collected from eight participants. The proposed model achieved a mean absolute error of 22.44 mg/dL and an R^2 value of 0.549, while SHAP analysis identified pre-meal glucose slope, baseline glucose, and short-term variability as the most influential predictors. The framework improved personalized glycemic prediction and model interpretability; however, it was evaluated on a limited number of participants and lacked comprehensive nutritional information, which may affect its generalizability [30].

Mukherjee et al. proposed an Internet of Things (IoT)-enabled diabetes monitoring framework that integrates a TabTransformer model with Explainable Artificial Intelligence (XAI) techniques for real-time diabetes risk prediction. The framework was evaluated using the Pima Indians Diabetes and UCI Early Stage Diabetes Risk Prediction datasets, where SHAP and LIME were employed to interpret model predictions and IoT devices continuously monitored patient glucose levels. The proposed TabTransformer achieved an accuracy of 94.50%, outperforming Random Forest, Logistic Regression, and TabNet while demonstrating stable performance through cross-validation. The framework enhanced real-time diabetes monitoring, prediction accuracy, and interpretability; however, its practical deployment and scalability in real-world IoT healthcare environments require further clinical validation [31].

Rossi et al. proposed the Diabetes Risk via Explainable AI Modeling (D.R.E.A.M.) framework for Type 2 Diabetes Mellitus (T2DM) risk prediction using the PIMA Indians Diabetes and BRFSS 2015 datasets. The framework integrated clinically guided feature engineering, class balancing, ensemble classifiers including Random Forest, XGBoost, and LightGBM, precision–recall curve-based threshold optimization, and SHAP for model interpretability. All models achieved AUC values above 0.83 and F1-scores of 0.78, while the Random Forest model provided the best balance between sensitivity and interpretability for clinical decision support. The proposed framework enhanced calibrated probabilistic prediction and explainability; however, its real-world clinical deployment and validation across diverse healthcare settings require further investigation [32].

Wang et al. developed ensemble machine learning models to predict health-related quality of life (HRQoL) in patients with diabetes using social determinants of health (SDoH). The study utilized the 2017 Behavioral Risk Factor Surveillance System (BRFSS) dataset comprising 4,946 adults with diabetes and evaluated multiple machine learning algorithms using demographic, socioeconomic, behavioral, and clinical features. The AdaBoost model achieved the best performance with an AUROC of 0.70, an accuracy of 65%, a precision of 68%, a recall of 73%, and an F1-score of 0.70. The proposed framework demonstrated the potential of SDoH-based machine learning for early risk stratification; however, its moderate predictive performance and lack of external validation and calibration optimization limit its immediate clinical applicability [33].

Cichosz et al. developed explainable machine learning models to predict week-to-week risks of hyperglycemia, hypoglycemia, and glycemic variability using continuous glucose monitoring (CGM) data from patients with Type 1 diabetes. The study utilized 46,593 weeks of CGM data from 187 patients for model development and internally validated Logistic Regression, XGBoost, and TabNet models, followed by external validation on two independent cohorts comprising 223 additional patients. The XGBoost model achieved the best performance with ROC–AUC values of 0.90, 0.89, and 0.80 for predicting hyperglycemia, hypoglycemia, and glycemic variability, respectively, while maintaining good generalizability across external cohorts. The proposed framework demonstrated accurate and explainable prediction of glycemic instability; however, it was designed specifically for Type 1 diabetes using CGM data, limiting its applicability to broader diabetes populations without further validation [34].

Ghaderzadeh et al. developed an interpretable Extra Trees Classifier (ETC) framework for early Type 2 Diabetes Mellitus (T2DM) prediction using data from the PERSIAN Dena Cohort comprising 3,203 adults. The study employed feature selection, randomized hyperparameter optimization, isotonic regression for probability calibration, and SHAP with Morris sensitivity analysis to improve model interpretability, evaluating the framework through ten-fold cross-validation and an 80/20 hold-out validation. The optimized ETC achieved a cross-validated accuracy of 69% with an AUC of 0.69, while SHAP identified fasting blood sugar, fatty liver status, age, kidney stone history, and triglycerides as the most influential predictors. The proposed framework enhanced model transparency, calibration, and fairness; however, its moderate predictive performance and limited validation within a single cohort indicate the need for further optimization and external multicenter validation [35].

Kim et al. developed an explainable machine learning

framework for depression prediction using data from the 2021 National Mental Health Survey of Korea, comprising 5,511 adults. The study evaluated multiple feature selection methods, including ReliefF, Markov Blanket, and Information Gain, across 12 machine learning classifiers, while SHAP was employed to provide model interpretability. The Stacking classifier with ReliefF achieved the highest performance (F2-score = 0.9851), whereas Markov Blanket performed best with ExtraTrees and LightGBM, identifying loneliness, reluctance to seek professional help, quality of life, and physical comorbidities as key predictors. The proposed framework improved prediction accuracy and interpretability through a dual-layer feature selection and SHAP approach; however, it was validated only on a Korean population, limiting its generalizability to other demographic groups [36].

Recent studies have shown that machine learning combined with metaheuristic optimization and explainable AI improves disease prediction by enhancing feature selection, predictive performance, and model transparency. Ensemble classifiers integrated with optimization algorithms have demonstrated promising results across various clinical datasets, while SHAP and LIME have improved interpretability. However, most existing approaches emphasize predictive accuracy using single datasets, with limited attention to multi-objective optimization, comprehensive explainability, cross-dataset validation, calibration, and clinical applicability.

D. ENSEMBLE LEARNING AND HYPERPARAMETER OPTIMIZATION

The increasing use of ensemble learning algorithms in healthcare data analytics is attributed to their reliability and enhanced performance capabilities. The commonest ensemble learners used in medical structured data include Random Forest, Extra Trees, Gradient Boosting, XGBoost, LightGBM, and CatBoost.

The current literature has shown that hyperparameter tuning helps to enhance the performance of ensemble learning algorithms. Hyperparameter tuning strategies like Grid Search, Random Search, Bayesian Optimization, and Optuna have been applied in the process of finding the best configuration for parameters. Optuna has received significant interest due to its efficiency and reduced computation cost.

Kundu et al. developed a multiclass machine learning framework to classify diabetes, prediabetes, and non-diabetes using data from the Bangladesh Demographic and Health Survey (BDHS) 2017–18. The study employed Recursive Feature Elimination with Cross-Validation (RFECV), hyperparameter tuning using Grid Search, and six machine learning algorithms, including Gradient Boosting Machine (GBM), Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Naïve Bayes (NB), and Multinomial Logistic Regression (MLR). The optimized GBM model achieved the best performance with an accuracy of 74.4%, precision of 66.5%, F1-score of 65.8%, and an AUC of 66.7%. The proposed framework enabled multiclass diabetes risk stratification for targeted screening; however, its moderate predictive performance and lack of explainable AI techniques and external validation may limit its clinical applicability [37].

Ulum et al. proposed a diabetes prediction framework by integrating the Extreme Gradient Boosting (XGBoost) algorithm with the Synthetic Minority Over-sampling Technique (SMOTE) and Bayesian Optimization for hyperparameter tuning. The study utilized the Pima Indians Diabetes Dataset comprising 768 patient records and employed SMOTE to address class imbalance, followed by Stratified 5-Fold Cross-Validation to optimize model performance. The optimized XGBoost model achieved an accuracy of 88%, precision of 79%, recall of 91%, F1-score of 84%, and ROC–AUC of 0.955. The proposed framework effectively improved diabetes classification by handling imbalanced data and optimizing model parameters; however, it was evaluated on a single benchmark dataset and lacked explainable AI techniques and external validation to demonstrate clinical generalizability [38].

Dhanka et al. conducted a systematic literature review on hybrid optimization frameworks that integrate hyperparameter tuning and feature selection to improve machine learning-based disease classification. The review analyzed studies on cardiovascular disease, diabetes, cancer, oncological, and metabolic disorders, focusing on Bayesian optimization, grid search, metaheuristic algorithms, and feature selection techniques such as filter, wrapper, and embedded methods. The findings showed that hybrid optimization strategies improved diagnostic accuracy by 12–15% compared with sequential optimization approaches while enhancing computational efficiency and model interpretability. The review provides valuable guidelines for designing clinically applicable machine learning models; however, it is based on comparative analysis of existing studies and does not include experimental validation of the proposed optimization frameworks [39].

Resma et al. developed an ensemble machine learning framework to predict diabetes mellitus and identify its associated risk factors among adults in northern Bangladesh using data from 1,408 participants. The study employed Recursive Feature Elimination (RFE), Random Forest feature importance, and Best First Search for feature selection, followed by six machine learning models and a stacking ensemble model, with SHAP analysis and multivariable logistic regression used for model interpretation and validation. The stacking ensemble achieved the best performance with an accuracy of 91.67% and an AUC of 0.967, while SHAP identified age, family history of diabetes, BMI, dietary behaviors, family income, and gender as the most influential predictors. The proposed framework improved predictive performance and interpretability through ensemble learning and explainable AI; however, it was validated on a single regional cohort, limiting its generalizability to broader populations [40].

Alam et al. developed an optimized machine learning framework for early Type 2 Diabetes Mellitus (T2DM) prediction using a dataset of 485 patients from Bangladesh.

The study evaluated eight machine learning algorithms with GridSearchCV-based hyperparameter tuning and five class-balancing techniques, including SMOTE, ADASYN, Random Oversampling, Random Undersampling, and SMOTE-Tomek. The optimized Random Forest model combined with Random Oversampling achieved the best performance, obtaining 99.50% accuracy, 99.50% F1-score, and an AUC of 0.98. The proposed framework effectively addressed class imbalance and improved predictive performance; however, it was evaluated on a relatively small regional dataset and did not incorporate explainable artificial intelligence techniques or external validation to assess its clinical generalizability [41].

Valilou et al. proposed a deep learning framework for diabetes prediction by integrating CatBoost, Convolutional Neural Networks (CNN), and Bidirectional Long Short-Term Memory (Bi-LSTM) networks with Mountain Gazelle Optimizer (MGO)-based hyperparameter tuning. The study utilized the Pima Indians Diabetes Dataset and employed Recursive Feature Elimination (RFE) to select the most relevant features for model development. The proposed framework achieved an accuracy of 95.5%, precision of 96%, recall of 95%, and F1-score of 95%, outperforming conventional machine learning and standalone deep learning models. The framework improved prediction accuracy and feature interpretability through optimized deep learning and feature selection; however, it was evaluated on a single benchmark dataset and lacked external clinical validation to demonstrate its generalizability [42].

Recent studies have shown that optimization-driven machine learning and deep learning models substantially improve diabetes prediction by enhancing feature selection, hyperparameter tuning, and classification accuracy. Explainable AI techniques have further increased model transparency by identifying clinically significant predictors. However, most existing approaches rely on single datasets and emphasize predictive performance, with limited focus on multi-objective optimization, comprehensive explainability, cross-dataset validation, calibration, and clinical applicability.

E. CROSS-DATASET VALIDATION AND TRUSTWORTHY CLINICAL AI

Most diabetes prediction algorithms are validated using a single dataset and classic training-testing splitting approaches. Even though the form of evaluation gives some insights into model quality, it does not give adequate information about model generalizability and robustness on diverse patients.

Recently, cross-dataset validation became widely recognized as an efficient method of evaluating machine learning algorithms in healthcare applications. By applying validation of algorithms on different datasets, one may analyze their robustness, transferability, and applicability in the real world. Furthermore, the methods of calibration analysis, decision curve analysis, bootstrapping, and statistics, such as Matthews Correlation Coefficient (MCC), Cohen's Kappa, and Precision-Recall Area Under Curve (PR-AUC), became important in evaluation of trustable AI systems in healthcare.

Muntaha et al. proposed a multi-disease machine learning framework for predicting diabetes while simultaneously analyzing associated kidney and cardiovascular disorders. The study utilized three publicly available datasets and applied a unified preprocessing pipeline, mutual information ranking, model-based feature selection, and seven supervised machine learning classifiers evaluated using stratified k-fold cross-validation. Among the evaluated models, Random Forest achieved the best performance with an accuracy of 76% and an ROC–AUC of 0.823, while glucose, BMI, and blood pressure were consistently identified as the most significant shared predictors across the three diseases. The proposed framework improved integrated chronic disease risk assessment by identifying common clinical risk factors; however, its predictive performance remained moderate and further validation on larger multicenter datasets is required [43].

Hudumula et al. proposed a stacking ensemble framework for diabetes prediction by integrating Random Forest, XGBoost, LightGBM, Support Vector Classifier, and a Deep Neural Network with Logistic Regression as the meta-classifier. The framework was evaluated on three publicly available datasets, namely the Pima Indians Diabetes Dataset, Frankfurt Hospital Diabetes Dataset, and CDC BRFSS Survey, while SHAP, LIME, and Kendall's τ were employed to assess model interpretability and explanation consistency. The proposed ensemble achieved AUC values of 0.901, 0.817, and 0.859 on the Pima, Frankfurt, and BRFSS datasets, respectively, with strong SHAP–LIME agreement observed on clinical datasets. The framework improved predictive performance, cross-dataset evaluation, and explanation reliability; however, the lower interpretability agreement on heterogeneous population-level data indicates the need for more robust explainability methods for diverse real-world datasets [44].

Ansari et al. proposed a leakage-safe benchmarking framework for multi-disease prediction by evaluating six conventional machine learning models, namely Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), XGBoost, Decision Tree (DT), and k-Nearest Neighbors (KNN). The framework was assessed on three benchmark disease datasets using nested cross-validation, while Local Interpretable Model-Agnostic Explanations (LIME) and Permutation Feature Importance (PFI) were employed to provide local and global model interpretability. Experimental results showed that Random Forest and XGBoost achieved the highest classification accuracy, whereas SVM and Logistic Regression obtained superior ROC–AUC performance under outer-fold validation. The proposed framework established a robust and interpretable benchmarking pipeline for disease prediction; however, it focused on comparative evaluation rather than developing a novel predictive model and requires validation on larger real-world clinical datasets [45].

Khan et al. proposed an interpretable hybrid machine

learning framework for Type 2 Diabetes Mellitus (T2DM) prediction using structured Electronic Health Record (EHR) data. The framework integrated systematic data preprocessing, Random Forest-based feature selection, SHAP explainability, and domain-aware model validation, and was evaluated on the Pima Indians Diabetes, UCI Diabetes 130-US Hospitals, MIMIC-III, and a simulated longitudinal EHR dataset. The Random Forest + SHAP model achieved the best performance with an accuracy of 91% and an AUROC of 0.93, while identifying glucose, HbA1c, insulin, BMI, and blood pressure as the most influential clinical predictors. The proposed framework improved prediction accuracy, interpretability, and clinical applicability; however, its effectiveness requires further validation on large real-world multicenter EHR datasets to establish broader generalizability [46].

Presitha et al. proposed a deep learning framework for diabetic retinopathy (DR) detection and severity classification using retinal fundus images. The study evaluated Inception-ResNet-V2, EfficientNet-B4, and ConvNeXt-Small on the APTOS 2019 and DDR datasets, employing ordinal learning for severity classification, bootstrap confidence intervals, McNemar's significance test, and Grad-CAM for model explainability. Experimental results demonstrated that ConvNeXt-Small achieved superior cross-dataset generalization, while ordinal learning significantly improved severity classification accuracy and agreement. The proposed framework enhanced robustness, interpretability, and cross-dataset performance for DR screening; however, it requires validation on larger multicenter clinical datasets to establish its real-world applicability [47].

Soomro et al. proposed a leakage-safe machine learning framework for diabetes classification by incorporating fold-contained preprocessing and multi-seed stratified 10-fold cross-validation to ensure unbiased model evaluation. The framework was evaluated on the Pima Indians Diabetes Dataset and the Early-Stage Diabetes Risk Dataset using six machine learning models, including Logistic Regression, Random Forest, Feedforward Neural Network, Backpropagation Neural Network, General Regression Neural Network (GRNN), and Artificial Bee Colony Optimized Feedforward Neural Network (ABC-FFNN). Experimental results showed that GRNN achieved the best predictive performance with ROC–AUC values of 0.837 and 0.998 on the two datasets, respectively, while also demonstrating superior probability calibration. The proposed framework improved evaluation reliability and predictive performance by eliminating data leakage; however, it did not incorporate explainable artificial intelligence techniques to enhance model interpretability for clinical decision-making [48].

Seyedi et al. proposed an interpretable ensemble learning framework for diabetes prediction using clinical and demographic data collected from two regional datasets and their merged cohort. The framework incorporated class-aware imputation, data standardization, Synthetic Minority Oversampling Technique (SMOTE), and a soft-voting ensemble of Random Forest, Gradient Boosting, and Logistic Regression, while SHAP was employed to explain feature contributions. Experimental results achieved an average ROC–AUC above 0.95 with balanced precision and recall across all datasets, demonstrating stable cross-dataset performance and reliable risk estimation. The proposed framework enhanced prediction accuracy, interpretability, and reproducibility; however, its effectiveness was evaluated only on regional cohorts, and further validation on larger multicenter or federated datasets is required to establish broader clinical generalizability [49].

Recent studies have shown that machine learning and deep learning models combined with optimization techniques and explainable AI substantially improve diabetes prediction by enhancing feature selection, hyperparameter tuning, predictive accuracy, and interpretability. Although several studies employ ensemble learning, leakage-free validation, and multi-dataset evaluation, most primarily focus on predictive performance with limited emphasis on multi-objective optimization, explanation stability, comprehensive statistical validation, and clinically deployable decision support systems.

F. RESEARCH GAPS AND MOTIVATION

However, despite the great progress achieved in diabetes prediction via machine learning approaches, a few key problems are yet to be solved. Firstly, the majority of current models pay little attention to feature selection and computational efficiency while trying to achieve maximum performance. Secondly, the optimization-based feature selection approaches suffer from the lack of explainability mechanisms, making them unsuitable for use in a clinical setting. Thirdly, most of the studies used only one dataset and no cross-dataset testing was conducted. Finally, little attention has been paid to explainability, calibration, decision-curve analysis, and counterfactual reasoning.

To fill the research gap, the current paper suggests a Multi-Objective Adaptive Improved Whale Optimization Explainable Artificial Intelligence (MO-AIWOA-XAI) framework which will incorporate multi-objective feature selection, Optuna-optimized Extra Trees classification, SHAP, LIME, DiCE, calibration, decision curve analysis, and cross-dataset testing. It is expected that the proposed framework will allow to build an accurate, explainable, and clinically applicable tool for diabetes risk prediction with minimal number of features required.

III. PROPOSED MO-AIWOA-XAI FRAMEWORK AND RESULT ANALYSIS

A. DATASET DESCRIPTION

To assess the performance of the MO-AIWOA-XAI framework in terms of its efficiency, resilience, and generalization ability, experiments have been conducted using two different public databases for diabetes testing: the PIMA Indians Diabetes Database and the Early Stage Diabetes Risk Prediction Database. The above-mentioned databases are based on two diabetes test procedures – physiological parameters-based

diagnosis and symptoms-based early diabetes prediction. Using several different datasets makes it possible to get an unbiased evaluation of the algorithm's performance.

1) PIMA Indians Diabetes Dataset

PIMA Indians Diabetes Dataset is considered one of the most popular benchmarking data sets in diabetes predictions researches. This data set was first compiled by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and includes medical information about women of Pima Indian origin with age 21 and older.

In total there are 768 patients in the data set and eight clinical predictors along with one binary target attribute. The chosen clinical attributes are considered clinically important physiological measures concerning diabetes development and growth. These attributes include plasma glucose concentration, blood pressure, body mass index, insulin, skin fold thickness, age, diabetes pedigree function, and pregnancy status.

Several attributes have physiologically implausible zeros which are regarded as missing values. Thus, proper data preparation and imputation techniques were conducted before modeling. Due to clinical significance and popularity of the data set it can be considered a benchmark for diabetes predictions research and allows comparing results with already existing publications. The dataset is publicly available at:

https://archive.ics.uci.edu/dataset/34/diabetes.

2) Early Stage Diabetes Risk Prediction Dataset

The Early Stage Diabetes Risk Prediction Dataset has been created to assist in detecting diabetes at an early stage based on the assessment of symptoms. As opposed to the Pima dataset, which has been created with the emphasis on laboratory tests and physiological measurements, the current dataset highlights demographic information and the symptoms of diabetes.

The dataset includes 520 records of patients with 16 predictor variables and one target variable representing whether diabetes was detected as positive or negative. Age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity are some of the variables included in the current dataset.

As mentioned above, the current dataset can be used for assessing machine learning models in screening settings where laboratory tests cannot be performed right away. Thus, by including symptoms into consideration, the dataset offers additional information about patients compared to the physiological measurements provided by the Pima dataset. The dataset is publicly available at:

https://archive.ics.uci.edu/dataset/329/early+stage+diabetes+risk+prediction+dataset

Table 1 summarizes the key characteristics of both datasets. The PIMA dataset primarily captures physiological and biochemical measurements, whereas the Early Stage Diabetes dataset focuses on symptom-oriented clinical observations. The complementary nature of these datasets enables comprehensive evaluation of the proposed framework under different diagnostic conditions.

Figure 1 illustrates the class distributions of both datasets before preprocessing. The PIMA dataset exhibits a moderate class imbalance, while the Early Stage Diabetes dataset demonstrates a relatively higher proportion of positive cases. To mitigate the impact of class imbalance on predictive performance, appropriate resampling techniques were employed during preprocessing.

The use of two independent datasets enables rigorous evaluation of the proposed framework's predictive capability, interpretability, and generalization performance. Furthermore, cross-dataset validation experiments were conducted to investigate the transferability of the learned patterns across diverse patient populations and different diabetes prediction scenarios.

B. DATA PREPROCESSING

Data preprocessing plays a key role in building effective and resilient machine learning models, especially in health care scenarios where the quality of data impacts model prediction accuracy. In such a clinical dataset, some instances of missing data, measurement inconsistencies, class imbalance, and redundancy may arise, which can negatively impact model training and generalization processes. For this reason, a proper data preprocessing pipeline was employed for the diabetes datasets.

The overall preprocessing workflow adopted in the study is illustrated in Figure 2. The pipeline consists of missing value treatment, feature engineering, normalization, categorical encoding, class balancing, and correlation analysis.

1) Missing Value Handling

In general, missing data is quite frequent in the medical dataset and can either be a source of bias or lead to lower predictive performance. For example, there are several clinical measurements in the PIMA Indians Diabetes Dataset including glucose concentration, blood pressure, skin thickness, insulin level, and body mass index where zero values are physiologically impossible and therefore considered missing values.

K-Nearest Neighbor (KNN) imputation was chosen in order to maintain the data distribution and prevent any information loss from the dataset. The KNN imputation fills in the missing values according to the values of the samples that are close in terms of their features. As opposed to using median or mean imputation, KNN imputation helps to keep the data distribution intact.

For categorical data in the Early Stage Diabetes Dataset, missing values are replaced by the value with the highest frequency.

TABLE 1: Characteristics of the Diabetes Datasets Used in the Study

Characteristic	PIMA Indians Dataset	Early Stage Diabetes Dataset
Number of Instances	768	520
Number of Features	8	16
Target Classes	2	2
Data Type	Clinical Measurements	Symptoms and Demographics
Age Group	Adult Females	Mixed Population
Class Labels	Diabetes / Non-Diabetes	Positive / Negative
Primary Application	Clinical Diagnosis	Early Risk Screening
Feature Type	Numerical Clinical Variables	Categorical and Numerical Variables
Validation Purpose	Physiological Risk Assessment	Symptom-Based Risk Assessment

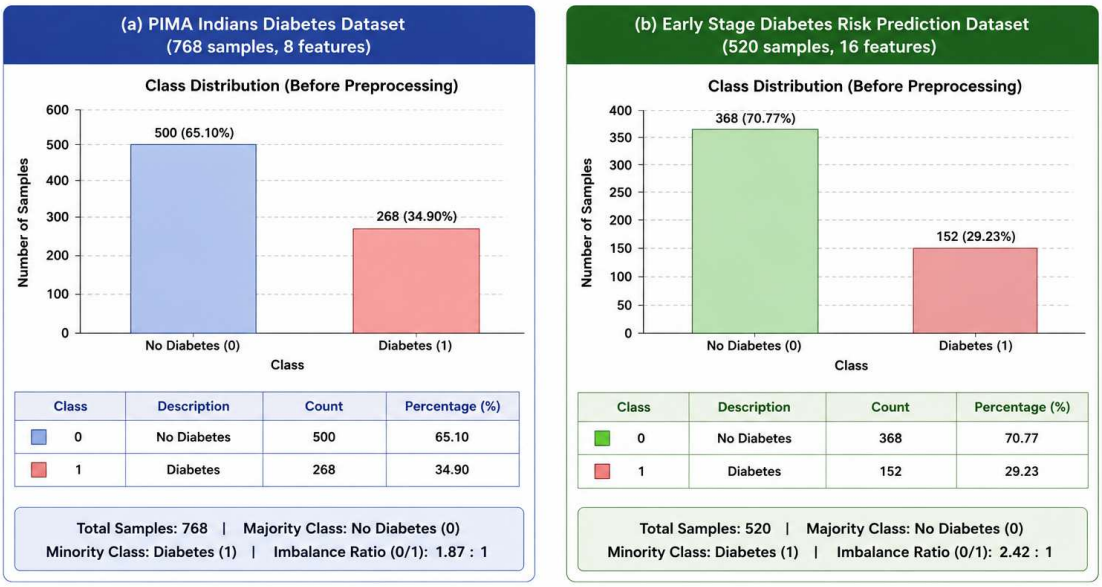


FIGURE 1: Class distribution of the PIMA Indians Diabetes Dataset and Early Stage Diabetes Risk Prediction Dataset before preprocessing.

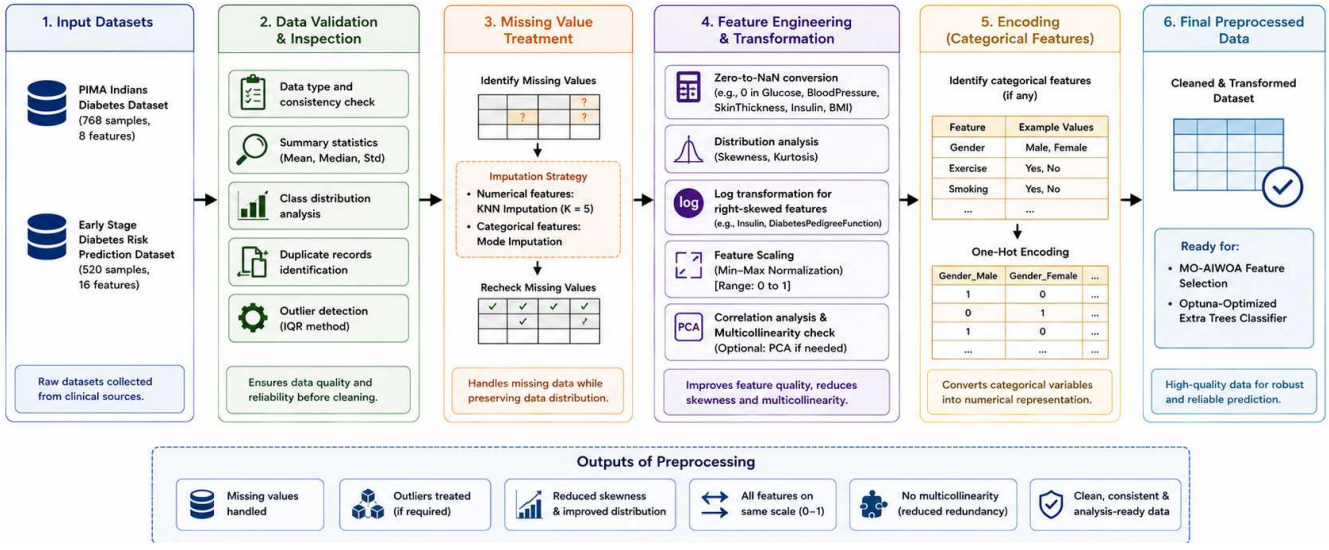


FIGURE 2: Data preprocessing workflow adopted in the proposed diabetes prediction framework.

2) Feature Engineering

Feature engineering was performed to capture nonlinear interactions among clinical variables and enhance the predictive capability of the classification model. Several interaction-based attributes were generated from clinically relevant features.

Examples include:

- Age × BMI
- Glucose × BMI
- Age × Glucose

The engineering of these traits provides further information concerning the joint impact of physiological and demographic characteristics in terms of the risk of diabetes. The use of the interaction variables enables the model to incorporate interactions which would otherwise remain hidden if considered separately.

3) Data Normalization

The variables in the datasets vary based on their measurement scale and unit. For instance, the variables of glucose concentration, insulin, age, and body mass index have varying values. The performance of optimization and learning algorithms may get affected because of this.

In order to normalize these variables, Min-Max Normalization was performed using:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where X denotes the original feature value, while X_{min} and X_{max} represent the minimum and maximum values of the corresponding feature.

This transformation scales all numerical variables to the interval [0,1], improving optimization efficiency and enhancing classifier convergence.

4) Categorical Encoding

Several categorical features have been included in the Early Stage Diabetes Dataset, which include symptoms and demographics. Categorical data cannot be used by machine learning algorithms unless they are converted to numbers. In this dataset, binary features like polyuria, polydipsia, obesity, and weakness have been encoded to numerical format.

This ensures that the data remains meaningful and does not have any artificial ordering between the different classes.

5) Class Imbalance Handling

Class imbalance can be a problem in healthcare data, and this might lead to a bias towards the majority class in the predictive model. In order to resolve this problem, the Synthetic Minority Over-sampling Technique along with Tomek Links (SMOTE-Tomek) technique was used.

The SMOTE technique produces artificial examples of the minority class by generating examples between neighboring examples. Tomek Links then eliminates the overlapping and borderline examples in the vicinity of the class borders.

Following resampling, the balanced dataset contained:

- Diabetes cases (Positive): 491
- Non-diabetes cases (Negative): 491

The balanced class distribution reduces classification bias and improves the model's ability to identify both diabetic and non-diabetic individuals.

6) Correlation Analysis

Prior to feature optimization, correlation analysis was performed to investigate relationships among clinical variables and identify potential multicollinearity issues [50].

Pearson correlation coefficients were computed for all numerical attributes, and the resulting relationships were visualized using a correlation heatmap, as shown in Figure 3.

Correlation analysis gives us important clues regarding dependency and redundancy among features. High correlation between attributes can add to unnecessary complexity and may harm model interpretability. Correlation analysis outcomes act as the initial guidance prior to the feature selection phase based on optimization techniques such as MO-AIWOA algorithm.

7) Dataset Splitting

After pre-processing, the data sets were split into training and testing data sets in an 80:20 split. Training data sets were used for feature selection, parameter tuning, and the development of models whereas testing data sets were only used to evaluate model performance.

To further measure the robustness of the model and eliminate any bias in the evaluation process, we used stratified 10-fold cross-validation while training our model.

8) Final Dataset Characteristics

Following the preprocessing stage, the datasets were converted to be machine learning ready by filling in any missing values, encoding categorical variables, normalizing numerical attributes, generating interaction features, and handling class imbalance using SMOTE-Tomek resampling. These datasets formed a strong base for carrying out the feature selection, classification, and explainability analysis experiments in the research.

C. PROPOSED MULTI-OBJECTIVE ADAPTIVE IMPROVED WHALE OPTIMIZATION ALGORITHM (MO-AIWOA)

The process of feature selection is an essential part of any medical data mining analysis since the presence of redundant and irrelevant features may impair the quality of classification and increase computation costs. Classical methods of feature selection do not pay attention to feature reduction, concentrating only on improving the prediction performance.

In order to improve the existing problems, we offer a method called MO-AIWOA which is used to extract the most significant subset of features from the set of diabetes features. Our approach is a modification of Whale Optimization Algorithm, which includes an adaptive mechanism and multi-objective fitness function. The overall architecture of the proposed MO-AIWOA framework is illustrated in Figure 4.

1) Standard Whale Optimization Algorithm

The Whale Optimization Algorithm (WOA) is a population-based metaheuristic optimization technique inspired by the

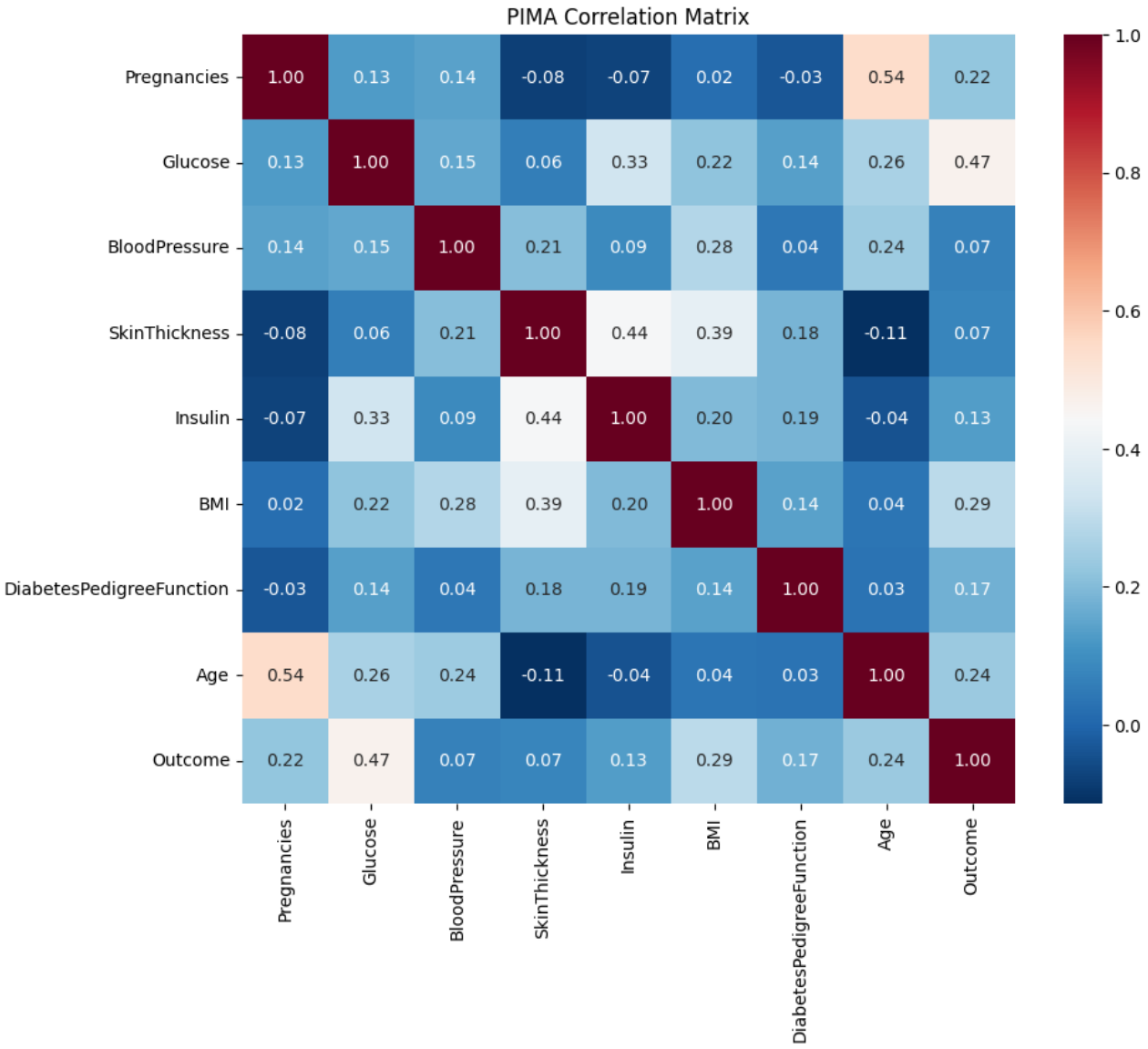


FIGURE 3: Correlation heatmap illustrating relationships among diabetes-related clinical features.

bubble-net hunting strategy of humpback whales. The algorithm mimics the social hunting behavior of whales through three primary operations:

1) Encircling prey
2) Bubble-net attacking mechanism
3) Random search for prey

The position update mechanism can be expressed as:

$$\mathbf{D} = |\mathbf{C} \cdot \mathbf{X}^*(t) - \mathbf{X}(t)| \tag{2}$$

$$\mathbf{X}(t+1) = \mathbf{X}^*(t) - \mathbf{A} \cdot \mathbf{D} \tag{3}$$

where:

- $\mathbf{X}^*(t)$ represents the current best solution.
- $\mathbf{X}(t)$ denotes the position of a whale.
- $\mathbf{A}$ and $\mathbf{C}$ are coefficient vectors controlling exploration and exploitation.

Although WOA demonstrates effective global search capability, it may suffer from premature convergence and reduced population diversity when dealing with complex medical datasets.

2) Adaptive Improvements in MO-AIWOA

To enhance optimization performance, several modifications were incorporated into the conventional WOA framework.

a: Adaptive Exploration-Exploitation Control

The two phases of exploration and exploitation can be dynamically tuned during the entire process of optimization. In initial iterations, the algorithm puts more emphasis on global exploration to explore various parts of the search space.

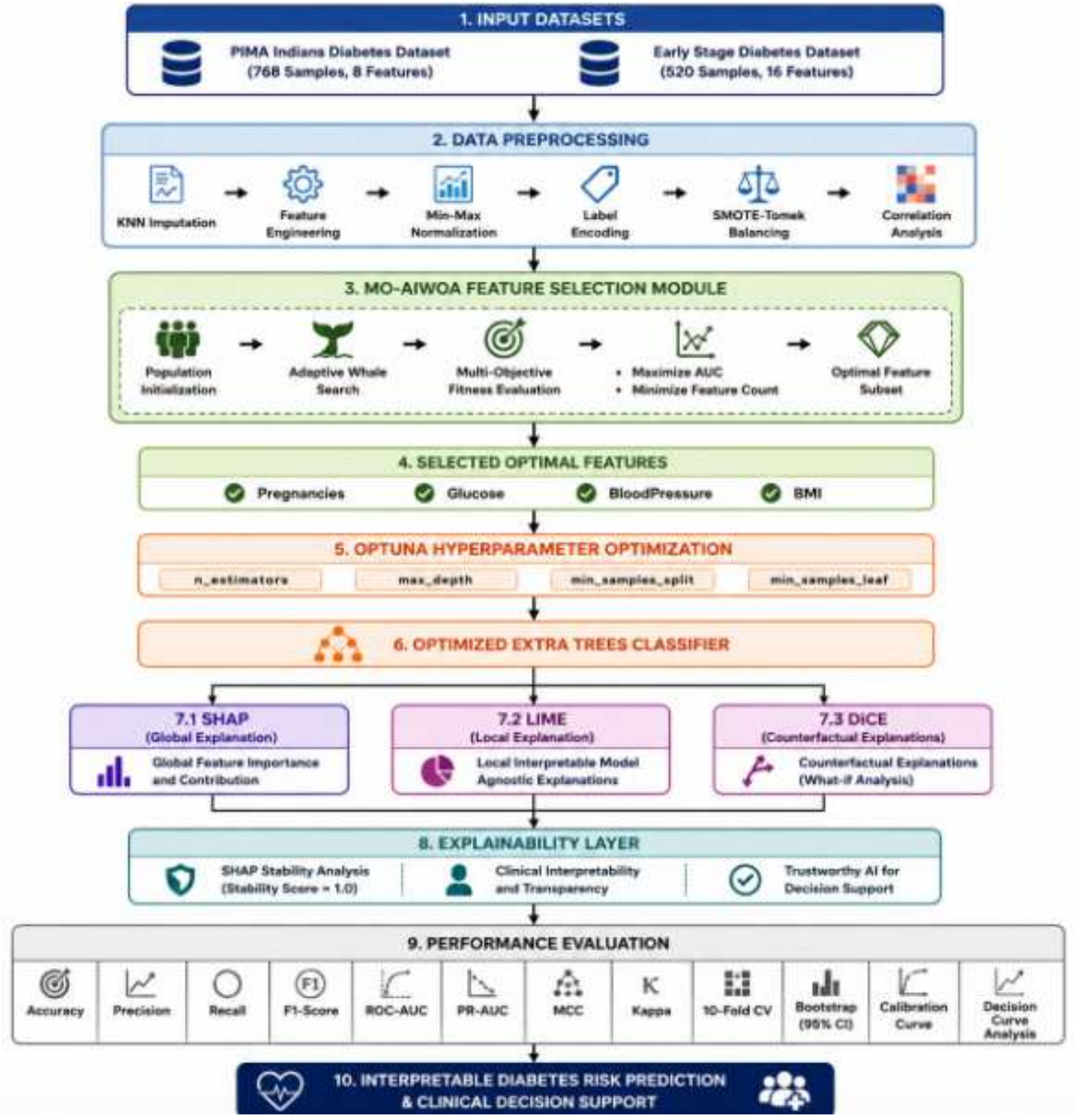


FIGURE 4: The overall architecture of the MO-AIWOA-XAI framework that is proposed. The framework encompasses the following components: data preprocessing, MO adaptive whale optimization-based improved feature selection, Optuna-based Extra Trees classifier, SHAP, LIME, and DiCE explainability components, and validation methods in order to predict the diabetes risk.

However, exploitation starts dominating in further stages of optimization.

b: Dynamic Position Updating

In order to avoid getting stuck and improving the convergence properties, adaptive position update strategies have been developed. The step size is updated based on the current optimization stage to navigate the feature space.

c: Random Perturbation Strategy

A random perturbation scheme is used to improve the population's diversity and prevent the possibility of being stuck

at the local optimum points. Random perturbations are introduced when there is little improvement in the population for successive iterations.

d: Binary Feature Representation

Each whale represents a candidate feature subset using a binary vector:

$$X_i = [x_1, x_2, x_3, \ldots, x_m] \tag{4}$$

where:

$$x_j = \begin{cases} 1, & \text{if feature } j \text{ is selected} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

and m denotes the total number of available features.

3) Multi-Objective Fitness Function

Unlike traditional feature selection methods that optimize a single objective, the proposed MO-AIWOA simultaneously considers predictive performance and feature reduction.

The fitness function is defined as:

$$Fitness = 0.7 \times AUC + 0.3 \times FeatureReduction \tag{6}$$

where:

$$FeatureReduction = 1 - \frac{N_{selected}}{N_{total}} \tag{7}$$

and:

- AUC represents the Area Under the Receiver Operating Characteristic Curve.
- $N_{selected}$ denotes the number of selected features.
- N_{total} denotes the total number of available features.

The formulation encourages the selection of compact feature subsets while maintaining high predictive performance.

4) Optimization Procedure

The proposed MO-AIWOA operates according to the following steps:

1) Initialize whale population randomly.
2) Generate binary feature subsets.
3) Train the classifier using selected features.
4) Compute AUC score.
5) Evaluate fitness using Equation (6).
6) Update whale positions using adaptive search mechanisms.
7) Apply random perturbation when necessary.
8) Update the global best solution.
9) Repeat until maximum iterations are reached.
10) Return the optimal feature subset.

The workflow of the proposed algorithm is presented in Figure 5.

5) Selected Feature Subset

The optimization process identified the most informative diabetes-related features from the original feature space. The selected subset consisted of:

- Pregnancies
- Glucose
- BloodPressure
- BMI

The chosen features performed dimensionality reduction by about 60%, while still performing well as predictors. Reduced set of features enhances model explainability, reduces costs of computations, and helps clinicians comprehend the most significant predictors of diabetes.

6) Experimental Settings

The MO-AIWOA algorithm was configured using the parameter settings summarized in Table 2.

TABLE 2: Experimental Settings for MO-AIWOA

Parameter	Value
Population Size	30
Maximum Iterations	50
Optimization Objective	Multi-Objective
Fitness Function	Eq. (6)
Classifier for Evaluation	Extra Trees
Cross-Validation	10-Fold
Random State	42

The resulting feature subsets were subsequently utilized by the Optuna-optimized Extra Trees classifier for diabetes risk prediction and explainability analysis.

D. OPTUNA-OPTIMIZED EXTRA TREES CLASSIFIER

After performing feature selection via the use of the new MO-AIWOA approach, an Extra Trees (ET) classifier was used to predict diabetes risks. The ET technique is a type of ensemble machine learning algorithm which involves training multiple random trees and making their collective predictions to form the final prediction. Because of its ability to deal with nonlinear problems, robustness to noise, and low variance, the ET algorithm has proven to perform very well on structured health care data sets.

Unlike the conventional machine learning algorithms like the Logistic Regression and Decision Trees, ET adds more randomness to the process of tree construction through random splitting of nodes. This makes the model more generalized and helps to prevent overfitting when working with heterogeneous data. Also, the algorithm can efficiently handle numerical and categorical data without any data assumptions.

1) Hyperparameter Optimization Using Optuna

Predictive accuracy of machine learning algorithms largely depends on proper setting of hyperparameters. Classical methods of hyperparameter tuning, including Grid Search and Random Search, frequently consume considerable computational power and cannot efficiently cover the large search space. Hence, Optuna, one of the most advanced frameworks

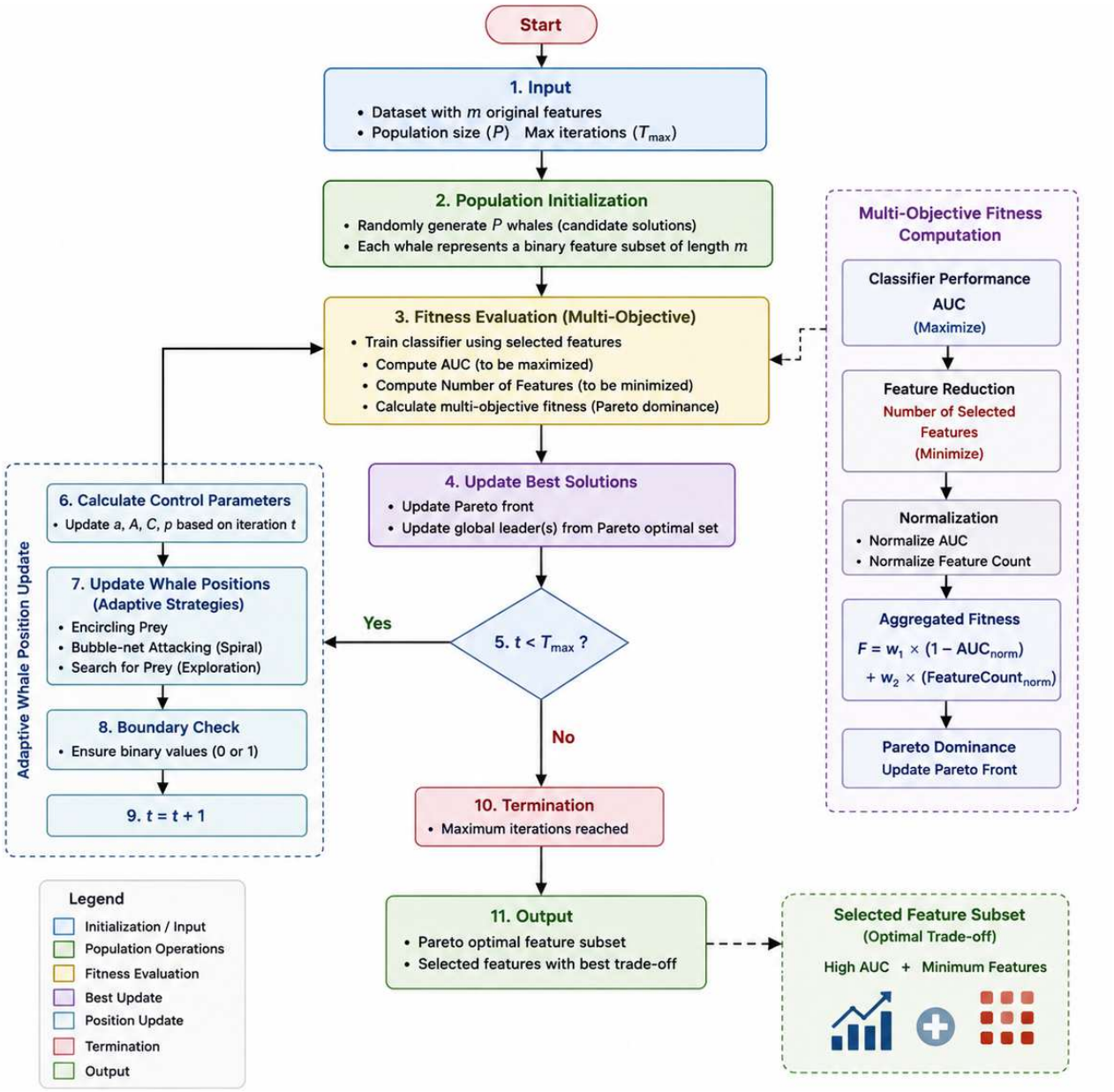


FIGURE 5: Workflow of the proposed MO-AIWOA feature selection algorithm.

for automatic hyperparameter tuning, was used to tune the parameters of the Extra Trees classifier.

Optuna makes use of a Bayesian optimization approach, which relies on the Tree-structured Parzen Estimator (TPE). While performing the optimization, the framework was testing multiple configurations in terms of their cross-validation results and then selected the best one.

The following hyperparameters were tuned:

- Number of Trees (`n_estimators`)
- Maximum Tree Depth (`max_depth`)
- Minimum Samples Required for Splitting (`min_samples_split`)
- Minimum Samples Required at Leaf Nodes (`min_samples_leaf`)

The optimal hyperparameter configuration obtained through Optuna is presented in Table 3.

TABLE 3: Optimal Hyperparameters Obtained Using Optuna

Parameter	Optimal Value
Number of Trees ($n_estimators$)	729
Maximum Depth (max_depth)	23
Minimum Samples Split ($min_samples_split$)	4
Minimum Samples Leaf ($min_samples_leaf$)	1

The optimized configuration was subsequently employed for all experiments and validation procedures reported in the study.

2) Model Training Procedure

The enhanced Extra Trees classifier was developed based on the features selected by the MO-AIWOA method. Prior to developing the model, the data were split into training and testing subsets in the ratio of 80:20. The training subset was used to develop the model while the testing subset was only used for evaluation.

In order to reduce bias in sampling and thus obtain reliable results, stratified sampling was used in the process of data splitting.

The complete classification workflow is illustrated in Figure 6.

3) Cross-Validation Strategy

To assess the ability of the proposed model to generalize, stratified 10-fold cross validation was carried out. The data set was divided into ten equal parts. In each iteration, nine parts were used to train the model, and one part was kept aside for validation. This process was repeated till all parts were used once for validation.

Mean Cross Validation Accuracy of Proposed Framework:

$$Accuracy_{CV} = 80.12\% \pm 3.78\% \tag{8}$$

The relatively small standard deviation indicates stable performance across different data partitions and demonstrates the robustness of the proposed framework.

4) Model Evaluation Metrics

Performance of the optimized classifier was tested using a number of statistical and clinical performance measures. Accuracy alone might not be sufficient to describe the classifier performance in medical applications; thus, other performance measures were utilized.

a: Accuracy

Accuracy measures the proportion of correctly classified instances:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

b: Precision

Precision quantifies the proportion of positive predictions that are actually correct:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

c: Recall

Recall evaluates the model's ability to correctly identify positive cases:

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

d: F1-Score

The F1-score represents the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

e: Area Under the ROC Curve (AUC)

The Area Under the Receiver Operating Characteristic Curve (AUC) measures the classifier's ability to distinguish between diabetic and non-diabetic patients across varying classification thresholds.

f: Matthews Correlation Coefficient (MCC)

MCC provides a balanced evaluation metric that considers all elements of the confusion matrix and is particularly suitable for healthcare applications:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{13}$$

g: Cohen's Kappa

Cohen's Kappa assesses agreement between predicted and actual class labels while accounting for agreement occurring by chance.

h: Precision-Recall Area Under Curve (PR-AUC)

PR-AUC provides additional insight into classifier performance, particularly when evaluating minority-class detection capability.

5) Statistical Validation

To further assess the reliability of the proposed framework, bootstrap resampling was employed to estimate confidence intervals for predictive performance. A 95% confidence interval was computed using 1000 bootstrap iterations.

The resulting confidence interval for classification accuracy was:

$$95\%\ CI = [77.16\%, 87.31\%] \tag{14}$$

The tight confidence interval implies that there is stability in the predictability behavior, and therefore the reliability of the developed framework is achieved.

Finally, the implementation of the MO-AIWOA algorithm for feature selection with Optuna tuning of the Extra Trees classifier achieves a good trade-off among accuracy, efficiency, and robustness.

E. EXPLAINABLE ARTIFICIAL INTELLIGENCE FRAMEWORK

While state-of-the-art models in machine learning demonstrate high predictive capability, they are still rarely used in the field of medicine since there is no sufficient transparency and interpretability of such models. The process of clinical

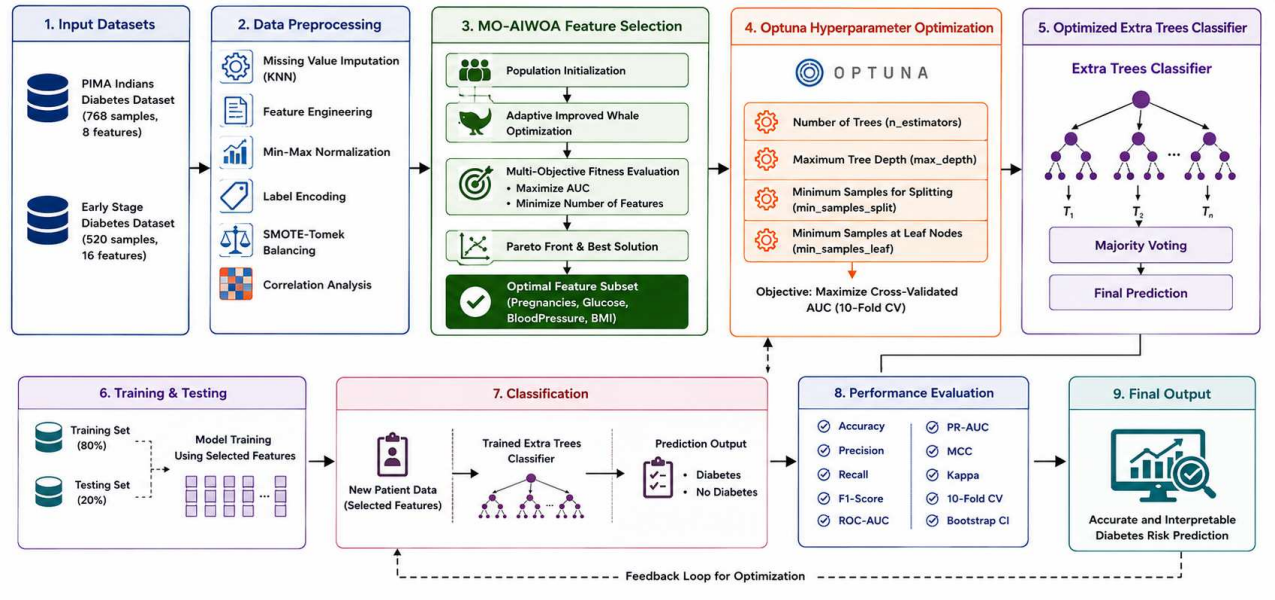


FIGURE 6: Classification workflow using MO-AIWOA feature selection and Optuna-optimized Extra Trees classifier.

decision making implies not only correct prediction but also explanations of results to help health workers trust the decisions made by such models. To solve this problem, the Explainable Artificial Intelligence framework was implemented in the MO-AIWOA-XAI architecture.

The XAI approach involves three approaches – SHAP, LIME, and DiCE.

The explainability workflow adopted in the study is illustrated in Figure 7.

1) SHAP-Based Global and Local Explanations

SHAP is a model-agnostic interpretability technique derived from cooperative game theory. SHAP quantifies the contribution of each feature toward the final prediction by computing Shapley values. These values represent the average marginal contribution of a feature across all possible feature combinations.

For a prediction model $f(x)$, the SHAP explanation can be represented as:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{15}$$

where:

- ϕ_0 denotes the baseline prediction,
- ϕ_i represents the contribution of feature i,
- M denotes the total number of features.

SHAP provides two levels of interpretability:

1) Global Interpretability
2) Local Interpretability

Global explanations identify the most influential diabetes risk factors across the entire dataset, while local explanations reveal how specific features contribute to individual patient predictions.

The SHAP summary plot shown in Figure 8 presents the overall importance of selected features. The analysis revealed that Glucose, BMI, BloodPressure, and Pregnancies were the most influential attributes contributing to diabetes prediction.

2) SHAP Stability Analysis

Explanation approaches must provide consistent interpretations for different data samples. For assessing robustness in the interpretation, SHAP stability test was conducted through resampling procedures.

In each experiment, SHAP scores were calculated for random sub-samples from the testing dataset, and then the feature rankings were evaluated based on the Spearman rank correlation coefficient. The average SHAP stability score is:

$$SHAP_{stability} = 1.0 \tag{16}$$

This result indicates perfect consistency of feature rankings across repeated experiments, demonstrating the reliability of the generated explanations.

3) LIME-Based Local Explanations

Local Interpretable Model-Agnostic Explanations (LIME) offer patient-level explanations through approximating the machine learning model behavior through a locally interpretable proxy model.

In LIME, for a particular prediction, multiple perturbed samples are generated and a linear model is learned in the vicinity of that sample. The importance of each feature can then be calculated in terms of its impact on the local prediction.

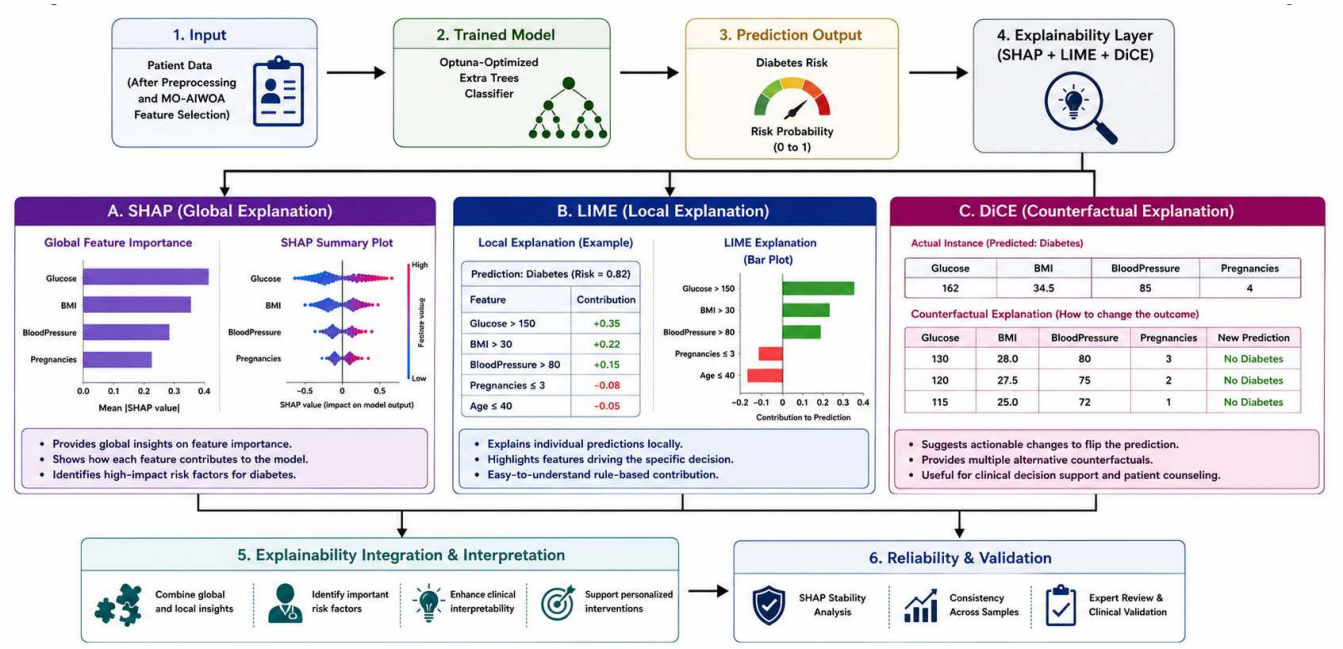


FIGURE 7: Explainability framework integrating SHAP, LIME, and DiCE for diabetes risk prediction.

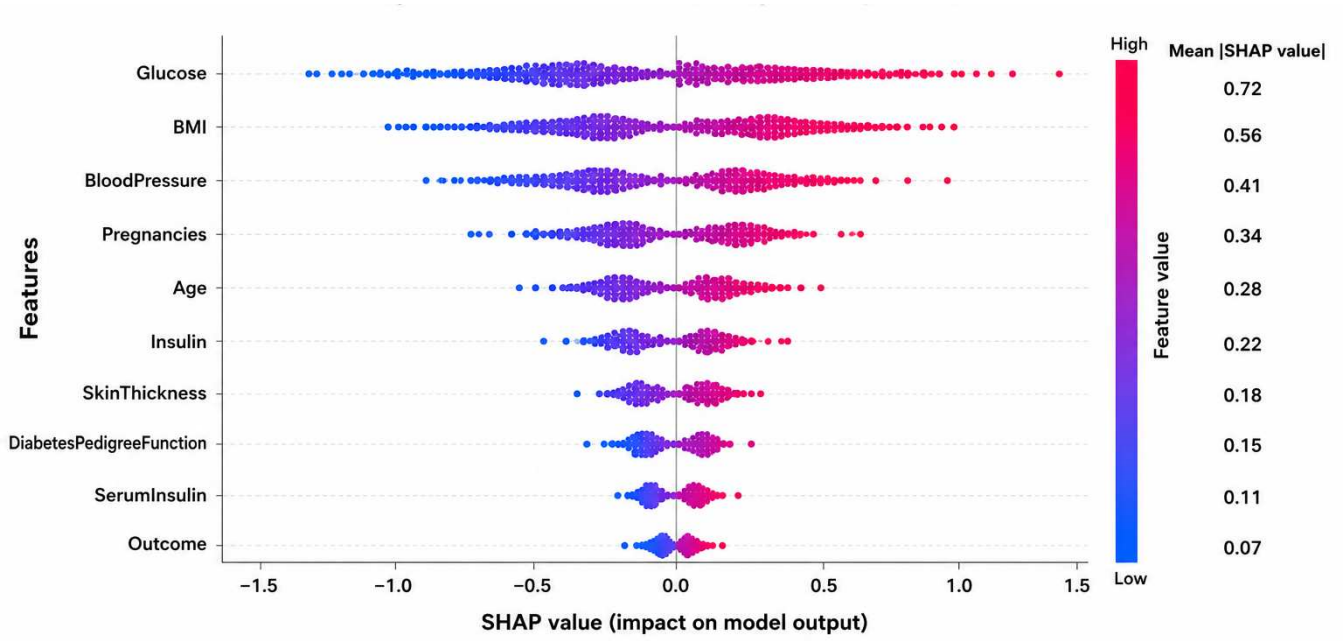


FIGURE 8: SHAP summary plot illustrating global feature importance.

The optimization problem of LIME can be formulated as:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{17}$$

where:

- f is the original prediction model,
- g is the interpretable surrogate model,
- L measures local fidelity,
- Ω controls model complexity.

Figure 9 presents an example LIME explanation for an individual patient prediction. Positive feature contributions increase diabetes risk, whereas negative contributions reduce the likelihood of diabetes.

The local explanations generated by LIME assist clinicians in understanding patient-specific risk factors and provide transparency for individual predictions.

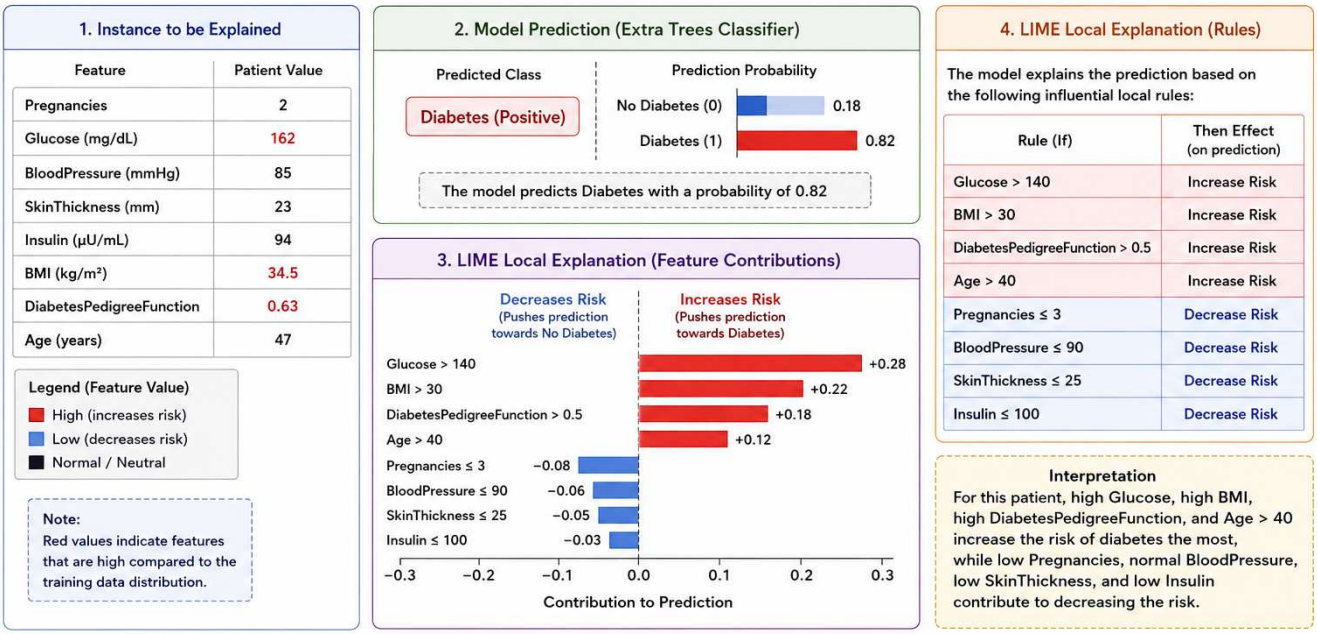


FIGURE 9: LIME explanation for an individual diabetes prediction.

4) DiCE Counterfactual Explanations

Even though SHAP and LIME provide information about the reason for the prediction being made, neither of them suggests how the prediction can be changed. Hence, DiCE were introduced to offer practical insights.

DiCE produces counterfactual instances that require minimal changes to be made to the input features to change the result of the prediction. This type of explanation is useful especially in the health-care sector because it suggests interventions.

In the case of a diabetic prediction for an individual, DiCE will suggest the minimum number of changes to make him/her non-diabetic.

An example counterfactual explanation generated by the proposed framework is presented in Table 4.

TABLE 4: Example Counterfactual Explanations Generated Using DiCE

Glucose	Age	Age_BMI	Outcome
0.698	2.788	1.157	0
0.698	2.377	0.795	0
0.698	2.055	-0.718	0

These counterfactual instances provide clinicians with interpretable recommendations regarding potential changes associated with lower diabetes risk.

5) Benefits of the Proposed XAI Framework

The integration of SHAP, LIME, and DiCE offers several advantages:

- Global understanding of feature importance.
- Patient-specific explanation of predictions.
- Actionable counterfactual recommendations.
- Improved transparency and trustworthiness.
- Enhanced clinical decision support.
- Improved acceptance of AI-assisted diagnosis.

As such, by utilizing these complementary explanation methods together, the framework offers a complete analysis of the model while maintaining its predictive performance. This makes the framework more applicable to healthcare in the real world since both interpretability and predictive performance are important.

F. EVALUATION METRICS AND VALIDATION STRATEGY

The proposed framework of MO-AIWOA-XAI was evaluated using a comprehensive evaluation approach for testing its predictive performance, robustness, reliability, and applicability in the clinical domain. As prediction systems in healthcare need to be accurate as well as reliable, a variety of metrics were used for the evaluation.

1) Performance Metrics

The classification performance of the proposed framework was evaluated using several widely accepted metrics derived from the confusion matrix.

a: Accuracy

Accuracy measures the proportion of correctly classified instances among all observations and is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

b: Precision

Precision quantifies the proportion of positive predictions that are correct:

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

c: Recall

Recall measures the ability of the classifier to correctly identify positive cases:

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

d: F1-Score

The F1-score combines precision and recall into a single metric:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{21}$$

e: Area Under the ROC Curve (AUC)

The Area Under the Receiver Operating Characteristic Curve evaluates the discriminative ability of the classifier across different decision thresholds.

f: Precision-Recall Area Under Curve (PR-AUC)

PR-AUC is particularly useful for healthcare applications because it emphasizes the model's ability to correctly identify positive cases.

g: Matthews Correlation Coefficient (MCC)

MCC provides a balanced assessment even when class distributions are unequal:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{22}$$

h: Cohen's Kappa

Cohen's Kappa evaluates the agreement between actual and predicted class labels while accounting for chance agreement.

i: Balanced Accuracy

Balanced Accuracy computes the average of sensitivity and specificity and provides a robust measure when evaluating medical datasets.

2) Cross-Validation Analysis

To evaluate model stability and generalization capability, stratified 10-fold cross-validation was performed. The dataset was divided into ten approximately equal partitions while preserving class distribution.

During each iteration:

1) Nine folds were used for training.
2) One fold was used for validation.
3) The process was repeated ten times.
4) Results were averaged across all folds.

The proposed framework achieved:

$$\text{Accuracy}_{\text{CV}} = 80.12\% \pm 3.78\% \tag{23}$$

The low standard deviation indicates stable predictive performance across different subsets of the data.

3) Bootstrap Confidence Interval Analysis

The assessment of statistical validity of the gained results was carried out through the procedure of bootstrap resampling, which was done for 1000 replications.

The bootstrap procedure allows sampling observations with repetitions and constructing an approximation to the distribution of the classifier accuracy.

The 95% confidence interval obtained for the proposed framework was:

$$95\%\ CI = [77.16\%, 87.31\%] \tag{24}$$

The relatively narrow interval demonstrates the consistency and robustness of the proposed model.

4) Calibration Analysis

Well-calibrated prediction probabilities in healthcare applications are just as important as classification accuracy itself. Calibration testing determines if the predicted probabilities match the frequency of the outcomes.

The calibration plot was obtained by analyzing the predicted probabilities vs. observed frequencies for various probability intervals.

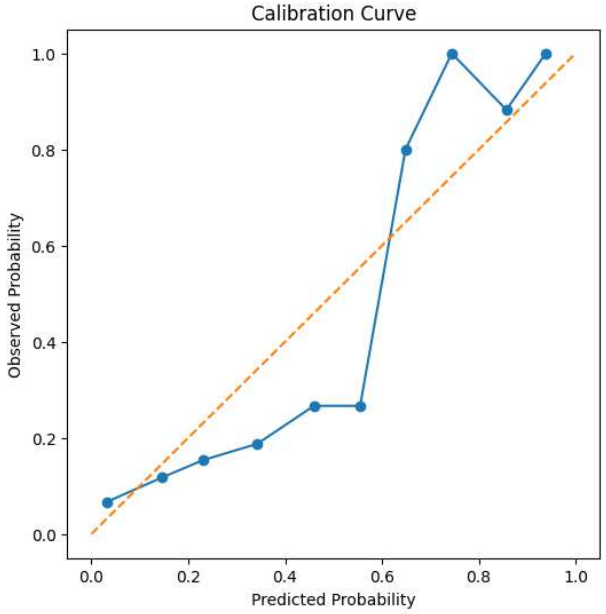


FIGURE 10: Calibration curve of the proposed MO-AIWOA-XAI framework.

The calibration results indicate that the predicted probabilities closely align with observed outcomes, demonstrating reliable probability estimation and supporting the use of the proposed framework in risk assessment applications.

5) Decision Curve Analysis

The Decision Curve Analysis (DCA) method was used to test the clinical utility of the proposed framework at various decision thresholds.

This method differs from conventional methods because it calculates the net clinical benefit by taking into account the balance between true positives and false positives.

The net benefit is calculated as:

$$\text{Net Benefit} = \frac{TP}{N} - \left(\frac{FP}{N} \times \frac{p_t}{1 - p_t} \right) \tag{25}$$

where:

- N denotes the total number of observations.
- p_t represents the threshold probability.

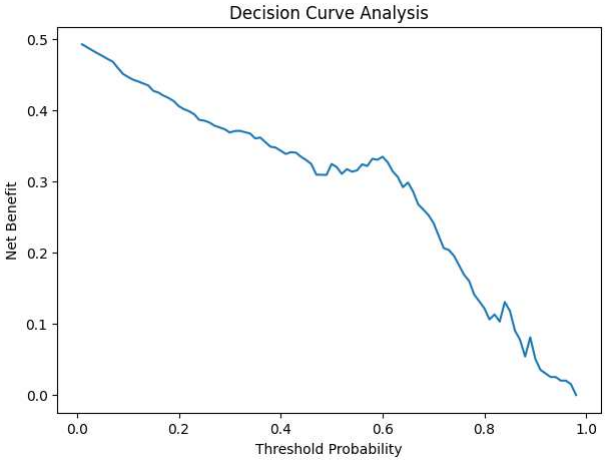


FIGURE 11: Decision Curve Analysis of the proposed diabetes prediction framework.

The decision curve demonstrates positive net benefit across a wide range of threshold probabilities, suggesting that the proposed model can provide clinically meaningful support for diabetes risk assessment.

6) Cross-Dataset Validation

In order to validate the effectiveness of the model in heterogeneous groups of patients, cross-validation procedures were performed based on two datasets, namely the PIMA Indians Diabetes Dataset and the Early Stage Diabetes Risk Prediction Dataset.

Both frameworks were tested separately on each dataset to verify the level of generalization for various diagnostics settings. It should be noted that the PIMA dataset is mostly concerned with physiological parameters, while the Early Stage dataset considers symptoms.

The results achieved show high performance of prediction in both cases.

7) Comparative Feature Selection Evaluation

To validate the effectiveness of the proposed MO-AIWOA algorithm, its performance was compared against several established feature selection approaches, including:

- Particle Swarm Optimization (PSO)
- Genetic Algorithm (GA)
- Whale Optimization Algorithm (WOA)
- No Feature Selection

The comparison considered:

- Classification Accuracy
- Area Under Curve (AUC)
- Number of Selected Features
- Feature Reduction Percentage

The analysis provides insight into the capability of MO-AIWOA to identify compact yet highly informative feature subsets.

8) Experimental Environment

All experiments were done via Python programming language and were run inside the Google Colab environment. Machine learning, optimization, and explainability experiments were realized with Scikit-learn, Optuna, SHAP, LIME, DiCE, XGBoost, LightGBM, CatBoost, and associated scientific packages.

The whole experimental process has been organized in order to guarantee reproducibility, transparency, and consistency of all experiments.

G. DISCUSSION

Proposed MO-AIWOA-XAI model is able to efficiently overcome a number of serious problems related to diabetes risk prediction, such as feature redundancy, predictive performance, interpretability, and clinical feasibility. In contrast to traditional machine learning models which concentrate mostly on achieving high classification accuracy, the proposed framework is able to optimize predictive performance and feature reduction at the same time using the multi-objective optimization approach. This allows for the discovery of a small set of clinical features along with good diagnostic ability.

The feature selection procedure showed the efficiency of the proposed MO-AIWOA model, reducing the feature space from the original one by about 60%. Pregnancies, Glucose, Blood Pressure, and BMI are recognized as clinical markers of diabetes and are supported by current medical literature.

The classification model optimized using Optuna showed high prediction ability for both datasets. The accuracy of 83.76%, AUC of 0.9004, and F1-score of 0.84 were obtained for the PIMA Indians Diabetes Dataset. Also, in cross-validation with another dataset, the Early Stage Diabetes Risk Prediction Dataset, the accuracy of 95.19% and AUC of 0.992 were obtained.

One of the main advantages of the proposed framework is its ability to provide explanations. Unlike other high-performing models, which work as black boxes, SHAP,

LIME, and DiCE algorithms help to understand predictions in detail. SHAP gave both local and global explanations, showing the importance of features for predictions, while LIME created personalized explanations for each patient, giving the reasons behind the model's decision. DiCE, in turn, created counterfactual explanations to understand what is necessary to change in order to change the outcome of prediction. The SHAP stability score was equal to 1.0, indicating high stability of explanations.

The designed framework was tested for its validity through several methods of statistical evaluations. The result of 10-fold cross-validation accuracy being $80.12\% \pm 3.78\%$ suggests that the framework performs reliably on any kind of data partitioning. The 95% confidence interval of bootstrap analysis was found to be in the range [77.16%, 87.31%]. It can be noted that calibration analysis suggested that the predicted probabilities were well aligned with the observations, and positive net benefit was obtained from decision curve analysis at many probability thresholds.

1) Computational Complexity Analysis

The computational complexity of the proposed framework is primarily influenced by the MO-AIWOA feature selection stage and the Extra Trees classification process. Let P denote the population size, T the number of optimization iterations, and m the number of features. The approximate complexity of the feature selection process can be expressed as:

$$O(P \times T \times m) \tag{26}$$

since each candidate solution represents a feature subset that must be evaluated using a classification model.

The complexity of the Extra Trees classifier depends on the number of trees and training samples and can be approximated as:

$$O(n_{trees} \times n \log n) \tag{27}$$

where n denotes the number of training instances.

While there is an extra amount of computational effort involved in feature selection through optimization-based approaches in comparison to the filter-based approach, the lower feature dimensionality helps to reduce classification costs. Additionally, the SHAP, LIME, and DiCE explanations will be produced in the interpretability phase only, not impacting the model training process.

2) Limitations and Future Work

In spite of its good results, the developed framework has several weaknesses. Firstly, the experiments were performed on public benchmark data, which do not fully cover the heterogeneity of clinical populations. Secondly, the optimization process used for the selection of the features requires more computational resources than plain feature ranking methods. Thirdly, the techniques of explainability considered in the work are post-hoc approaches and may not reflect the logic of the inner classifier fully.

Further research should be devoted to the validation of the framework based on larger multi-center clinical datasets of patients with heterogeneous background. Further research may also be directed to the application of federated learning algorithms that would allow for developing the models in collaboration without compromising patient privacy. Another direction may be related to the use of temporal medical records, data from wearable devices, and deep learning representations for improving the predictive power of the model and performing personalized risk assessment of diabetes.

IV. CONCLUSION

In the research, a new framework for Explainable Diabetes Risk Prediction (MO-AIWOA-XAI) based on Multi-Objective Adaptive Improved Whale Optimization Algorithm (MO-AIWOA), Extra Trees classification optimization using Optuna library, and explainable Artificial Intelligence was developed. The introduced framework aimed at solving several problems related to healthcare-oriented Machine Learning applications, such as improving the prediction capabilities, minimizing the number of input features, and increasing the model interpretability.

The introduced MO-AIWOA algorithm managed to select a set of relevant features that were used to train the classifier, and reached around 60% feature selection while keeping high prediction performance. The optimization process was successful in finding a robust classification model that showed an accuracy of 83.76%, AUC of 0.9004, and F1-Score of 0.84 on the PIMA Indians Diabetes Dataset. In addition, testing on the Early Stage Diabetes Risk Prediction Dataset achieved an accuracy of 95.19% and AUC of 0.992.

The inclusion of SHAP, LIME, and DiCE contributed to improving model transparency through global, local, and counterfactual explanations. The SHAP stability score of 1.0 ensured the consistency and validity of the generated explanations. Validation using cross-validation, bootstrap confidence intervals, calibration, and decision curve analysis ensured that the proposed framework was robust, reliable, and potentially clinically useful.

The MO-AIWOA-XAI framework has proven to be capable of closing the gap between prediction performance and explainability in diabetes risk prediction. The use of optimization-based feature selection, ensemble learning, explainable artificial intelligence, and validation make the proposed framework a good candidate for future clinical decision support systems and intelligent healthcare applications.

For future research, the focus will be on the clinical validation of the proposed system with large-scale datasets, integrating longitudinal patient records and data from wearable sensors, developing federated learning frameworks that guarantee patient data privacy, and deploying the framework within clinical decision support environments.

V. AVAILABILITY AVAILABILITY

The datasets analyzed during the current study are publicly available. The PIMA Indians Diabetes Database is

available from the UCI Machine Learning Repository: https://archive.ics.uci.edu/dataset/34/diabetes

The Early Stage Diabetes Risk Prediction Dataset is available from the UCI Machine Learning Repository: https://archive.ics.uci.edu/dataset/329/early+stage+diabetes+risk+prediction+dataset

All data used in this study were obtained from these publicly accessible repositories, and no new datasets were generated during the current study.

VI. AUTHOR CONTRIBUTIONS

Sumalatha conceived the research idea, designed the methodology, implemented the proposed framework, performed the experiments, analyzed the results, and prepared the original manuscript draft. Dr. Krishna Prakasha K supervised the research, provided technical guidance, reviewed the methodology and experimental results, and critically revised and edited the manuscript. Both authors read and approved the final version of the manuscript.

VII. FUNDING

The authors declare that no funding was received for this research, authorship, or publication of this article.

REFERENCES

[1] P. Chandrasekaran and R. Weiskirchen, "The role of obesity in type 2 diabetes mellitus—an overview," *International Journal of Molecular Sciences*, vol. 25, 2024.

[2] X. Lu, Q. Xie, X. Pan, R. Zhang, X. Zhang, G. Peng, Y. Zhang, S. Shen, and N. Tong, "Type 2 diabetes mellitus in adults: pathogenesis, prevention and therapy," *Signal Transduction and Targeted Therapy*, vol. 9, 2024.

[3] Y. Zhang, M. Liu, C. Zhou, Z. Ye, Y. Zhang, S. Yang, P. He, X. Gan, and X. Qin, "Social isolation, loneliness, and the risk of incident type 2 diabetes mellitus by glycemic status." *Diabetes & metabolism*, p. 101517, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267074311

[4] K. A. Lee, J. S. Kim, Y. J. Kim, I. Goak, H. Y. Jin, S. Park, H. Kang, and T. S. Park, "A machine learning-based prediction model for diabetic kidney disease in korean patients with type 2 diabetes mellitus," *Journal of Clinical Medicine*, vol. 14, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:277345713

[5] T. Li, J. Chen, X. Zhang, K. Wang, X. Zhao, Y. Cao, Z. Xu, S. Wang, P.-Y. Su, X. He, Y. Yang, X. Cao, X. Liang, and D. Ma, "A machine learning model for predicting the risk of diabetic nephropathy in individuals with type 2 diabetes mellitus," *Frontiers in Endocrinology*, vol. 16, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:282193064

[6] S. Raina and A. Kaul, "A metaheuristic-driven deep feature selection approach for hyperspectral image classification," *Remote Sensing Applications: Society and Environment*, p. 101916, 2026.

[7] A. Ziane, H. E. Bouhissi, and T. Hanne, "Intelligent diabetes prediction system based on hybrid pso-gwo feature selection and optimized machine learning," *Information*, 2026. [Online]. Available: https://api.semanticscholar.org/CorpusID:288915224

[8] A. K. Yadav, M. Khanna, and D. Anand, "A novel and efficient statistical and soft-computing intelligence integrated feature selection technique for human chronic diseases prediction," *Multimedia Tools and Applications*, vol. 84, pp. 41 853 – 41 896, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:277720587

[9] D. Bhatotia, A. Sharma, A. G. Dutt, S. Sharma, and S. Ghosh, "Xai machine learning framework for noninvasive cardiac output estimation using impedance cardiography," *2025 International Conference on Artificial Intelligence for Computing, Astronomy and Renewable Energy (AICARE)*, pp. 1–6, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:286013833

[10] Q. Liu, T. Hang, and Y. Wu, "Unveiling differential impacts of multidimensional urban morphology on heat island effect across local climate zones: Interpretable catboost-shap machine learning model," *Building and Environment*, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:275552244

[11] M. Azad, M. F. K. Khan, and S. A. El-Ghany, "Xai-enhanced machine learning for obesity risk classification: A stacking approach with lime explanations," *IEEE Access*, vol. 13, pp. 13 847–13 865, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:275745064

[12] V. Bakir, P. Goktas, and S. Özögür-Akyüz, "Dice-extended: A robust approach to counterfactual explanations in machine learning," *ArXiv*, vol. abs/2504.19027, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:278164725

[13] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 423–443, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:10137425

[14] Q. Xu, R. Yu, H. Qiu, Y. Jiang, J. Ball, C. Xu, and J. Sun, "Machine learning–based prediction model construction for type 2 diabetes mellitus: A comparison of algorithms and multilevel risk factor analysis," *Journal of Diabetes Research*, vol. 2026, no. 1, p. 4525736, 2026.

[15] A. Aich, M. M. Murshed, S. Hewage, and A. Mayeaux, "A copula based supervised filter for feature selection in machine learning driven diabetes risk prediction," *Scientific Reports*, vol. 16, no. 1, p. 12132, 2026.

[16] S. Yang, R. Mo, W. Wang, P. Zhen, and W. Han, "Development of a machine learning-based model for prediction of diabetes risk in patients with metabolic dysfunction-associated steatotic liver disease (masld)," *BMJ open*, vol. 16, no. 1, p. e107239, 2026.

[17] R. Srinithi, R. Yazhini, A. Sharmila, R. Sushmitha *et al.*, "Machine learning-based diabetes risk detection and diet suggestion system," in *2026 4th International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)*. IEEE, 2026, pp. 1–8.

[18] N. Mohankumar, P. Vimala, R. J. Kannan, S. Sujatha *et al.*, "Hybrid deep learning approach for accurate early-stage diabetes risk prediction," in *2026 International Conference on Visual Analytics and Data Visualization (ICVADV)*. IEEE, 2026, pp. 491–497.

[19] Y. Zhao, S. Lu, J. Lu, L. Yang, C. W. Lo, M. K. Wong, T. Li, H. Ren, X. Li, L. Xu *et al.*, "Risk prediction of chronic kidney disease progression in type 2 diabetes mellitus across diverse populations," *npj Digital Medicine*, vol. 9, no. 1, p. 250, 2026.

[20] M. Bernardini, M. di Cosmo, G. Barone, L. Romeo, and E. Frontoni, "Machine learning-based clinical decision support system for hepatic fibrosis risk prediction in general practice," *ACM Transactions on Computing for Healthcare*, 2026.

[21] Y. Özkan, Y. B. Özçelik, and A. Altan, "Optimization-driven hybrid machine learning framework for brain tumor classification in mri with metaheuristic feature selection," *Diagnostics*, vol. 16, no. 5, p. 819, 2026.

[22] G. T. Francis, A. Souri, and N. Inanç, "A systematic review of metaheuristic based feature selection strategies for cyber-attack detection in the iiot," *Artificial Intelligence Review*, 2026.

[23] S. Mudassar, A. Zameer, and M. A. Z. Raja, "Gwo-dagru: A hybrid deep learning framework with metaheuristic feature selection and self-weighted context gru for short-term wind power forecast," *Expert Systems with Applications*, p. 131279, 2026.

[24] A. Ziane, H. E. Bouhissi, and T. Hanne, "Intelligent diabetes prediction system based on hybrid pso-gwo feature selection and optimized machine learning," *Information*, vol. 17, no. 6, p. 533, 2026.

[25] P. Sarker, A.-A. Nahid, K. Choi, and M. A. Samad, "Feature reduction using swarm optimization and random forest classifiers for early diabetes risk prediction," *Scientific Reports*, 2026.

[26] S. Adamu, H. Alhussian, S. J. Abdulkadir, M. M. E. Eltahir, S. O. Khairy, G. Hyder, F. Ali, U. A. Ismaíla, S. Abdullahi, H. Mamman *et al.*, "Interpretable hybrid metaheuristic optimization with iteration-level behavior analysis for clinical feature selection in heart disease prediction," *Journal of King Saud University Computer and Information Sciences*, 2026.

[27] S. Sokoot, F. Zamani-Boroujeni, F. Davami, and P. Derakhshan-Barjoei, "Improving diabetes diagnosis using metaheuristic-based ensemble classification method," *Discover Computing*, vol. 29, no. 1, p. 261, 2026.

[28] K. Middha and A. Mittal, "Optimization enabled feature selection and hybrid deep learning approach for detection of type 2 diabetes mellitus," *Biomedical Materials & Devices*, pp. 1–26, 2026.

[29] S. Anam, D. E. Ratnawati, S. H. Wijoyo, and N. J. Paat, "Random forest classifier with binary grey wolf optimization feature selection for predicting diabetes mellitus." *International Journal of Mathematical, Engineering & Management Sciences*, vol. 11, no. 2, p. 804, 2026.

[30] A. Y. Ahmed, "Personalized glycemic control modeling using explainable artificial intelligence for optimizing diabetes management," *Scientia. Technology, Science and Society*, vol. 3, no. 1, pp. 85–96, 2026.
[31] S. Mukherjee and D. De, "Xai-iot based real-time diabetes prediction using tabtransformer," *Medicine in Novel Technology and Devices*, p. 100435, 2026.
[32] D. Rossi, A. Auriemma Citarella, F. De Marco, L. Di Biasi, H. Zheng, and G. Tortora, "Dream: diabetes risk via explainable ai modeling," *Multimedia Tools and Applications*, vol. 85, no. 2, p. 145, 2026.
[33] Y.-C. Wang, M. M. Islam, A. N. Okere, T. N. Poly, and M.-C. Lin, "Social determinants of health to predict health-related quality of life in diabetes patients: explainable machine learning approaches," *Endocrine*, vol. 91, no. 1, p. 57, 2026.
[34] S. L. Cichosz, S. S. Olesen, and M. H. Jensen, "Explainable machine-learning models to predict weekly risk of hyperglycemia, hypoglycemia, and glycemic variability in patients with type 1 diabetes based on continuous glucose monitoring," *Journal of Diabetes Science and Technology*, vol. 20, no. 3, pp. 836–847, 2026.
[35] M. Ghaderzadeh, Z. Rafie, and C. Salehnasab, "Explainable extratreeclassifier model for early detection of type 2 diabetes: evidence from the persian dena cohort," *BMC Medical Informatics and Decision Making*, vol. 26, no. 1, p. 36, 2026.
[36] M. G. Kim, K. C. Lee, K. Lee, H. U. Kim, Y. W. Seo, and S. W. Chae, "Explainable artificial intelligence approaches for predicting depression by combining feature selection methods and machine learning classifiers," *Digital Health*, vol. 12, p. 20552076251411968, 2026.
[37] L. R. Kundu and A. K. Majumder, "Hyperparameter-tuned machine learning techniques for the prediction of diabetes status among bangladeshi adults," *Advances in Public Health*, vol. 2026, no. 1, p. 8525279, 2026.
[38] M. N. I. Ulum and J. Unjung, "Enhancing diabetes classification performance using xgboost integrated with smote and bayesian hyperparameter optimization," *Journal of Soft Computing Exploration*, vol. 7, no. 1, pp. 9–18, 2026.
[39] S. Dhanka, A. Sharma, A. Kumar, S. Maini, and H. Vundavilli, "Advancements in hybrid machine learning models for biomedical disease classification using integration of hyperparameter-tuning and feature selection methodologies: A comprehensive review," *Archives of Computational Methods in Engineering*, vol. 33, no. 1, pp. 289–324, 2026.
[40] M. N. J. Resma, M. A. Kayum, P. Bhowmik, M. E. A. Khondaker, and M. K. Kibria, "Explainable ensemble machine learning for predicting diabetes mellitus and identifying key risk factors: a population-based study in northern bangladesh," *Scientific Reports*, 2026.
[41] M. A. Alam, J. Chisty, M. R. Nayeem, S. K. Ghosh, A. Biswas, and M. F. Alam, "A healthcare-oriented machine learning framework for early detection of type 2 diabetes," in *2026 7th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*. IEEE, 2026, pp. 60–66.
[42] M. Valilou, S. Valilou, and F. S. Gharehchopogh, "An enhanced diabetes prediction using an improved hybrid deep learning algorithm with mountain gazelle optimizer," *Journal of Diabetes & Metabolic Disorders*, vol. 25, no. 1, p. 46, 2026.
[43] A. Muntaha and U. Fatima, "A comprehensive machine learning framework for predicting diabetes, heart and kidney diseases with cross dataset validation and risk assessment," in *2026 Global Conference on Wireless and Optical Technologies (GCWOT)*. IEEE, 2026, pp. 1–6.
[44] Y. Hudumula, "Multi-dataset diabetes prediction using stacking ensemble learning and explainable artificial intelligence," 2026, manuscript.
[45] Z. A. Ansari, K. K. Kumar, S. Fatima, S. Siddiqui, and S. W. Mohsin, "Explainable artificial intelligence for cross domain evaluation of predictive models in multi-disease diagnosis," *Discover Computing*, vol. 29, no. 1, p. 163, 2026.
[46] M. Khan, P. Chakrabarty, B. Unhelkar, A. K. Lodhi, and A. Hussain, "An interpretable hybrid machine learning framework for robust type ii diabetes prediction using electronic health records," *International Journal of Artificial Intelligence and Computer Electronics*, vol. 2, no. 1, pp. 1–9, 2026.
[47] M. Presitha and K. Somasundaram, "Cross-dataset binary and ordinal diabetic retinopathy classification with deep learning and statistical evaluation," in *2026 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*. IEEE, 2026, pp. 1–6.
[48] A. Soomro and O. Ata, "A leakage-free and statistically validated machine learning evaluation framework for diabetes risk prediction across multiple datasets," *Research Square*, Jun. 2026, preprint, Version 1. [Online]. Available: https://doi.org/10.21203/rs.3.rs-9678084/v1
[49] A. Seyedi, M. Pournameti, and M. Mosavi, "Ensembleshield-dm: Interpretable, shift-robust ensemble learning for type 2 diabetes prediction across multi-regional cohorts," in *2026 Fourth International Conference on Distributed Computing and High Performance Computing (DCHPC)*, 2026, pp. 1–17.
[50] S. Seth, K. Chaudhary, and S. Ramachandran, "Type 2 diabetes mellitus associated pancreatic cancer prediction using combinations of machine learning models," *Biomed. Signal Process. Control.*, vol. 111, p. 108240, 2026.



SUMALATHA U received her B.E. and M.Tech. degrees from Viswesvaraya Technological University, Belagavi, and Ph.D. degree in biometric security from Manipal Academy of Higher Education (MAHE) University, Manipal, India. She is currently working as an Assistant Professor at the Department of Computer Science and Engineering, Mangalore Institute of Technology and Engineering, Moodabidri. Her current research interests are deep learning, biometrics, multimodal biometric systems, biometric template security, privacy enhancing technologies, cryptographic protocols, and their applications.



KRISHNA PRAKASH K (Senior Member, IEEE) received his B.E. and M.Tech. degrees from Viswesvaraya Technological University, Belagavi, and Ph.D. degree in network security from Manipal Academy of Higher Education (MAHE) University, Manipal, India. He is currently working as an Additional Professor at the Department of Information and Communication Technology, Manipal Institute of Technology, MAHE. He has more than 30 publications in national and international conferences and journals. His current research interests include information security, network security, algorithms, real-time systems, and wireless sensor networks.