


Enhancing Early Diagnosis of Type II Diabetes through Feature Selection and Hybrid Metaheuristic Optimization Techniques



Sunil Upadhyay^{1,*}  and Yogesh Kumar Gupta¹ 

¹Department of Computer Science, Banasthali Vidyapith-304022, Banasthali, Rajasthan, India

Abstract:

Introduction: Type-II Diabetes Mellitus (T2DM) is a chronic metabolic disorder characterized by elevated blood glucose levels, posing a critical global health challenge. It is largely attributed to lifestyle changes, unhealthy dietary habits, and lack of awareness. If not diagnosed early, T2DM can lead to severe complications, including damage to vital organs such as the kidneys, heart, and nerves. While timely and accurate diagnosis is crucial, current diagnostic procedures are often costly and time-consuming, necessitating innovative approaches to improve early detection. This study aimed to enhance the early prediction of T2DM by leveraging advanced hybrid metaheuristic optimization algorithms to improve model efficiency, accuracy, and computational time. The objective of this study is to develop a robust and interpretable hybrid machine learning framework that combines feature selection and metaheuristic optimization techniques to enable early, accurate, and computationally efficient diagnosis of T2DM.

Method: The methodology employed in this study involved three key steps: feature selection and refinement, model optimization, and evaluation. For feature selection, SHAP (SHapley Additive exPlanations) was integrated with Support Vector Machines (SVMs) to identify the most significant predictive features. This was followed by Particle Swarm Optimization (PSO), which was utilized for feature refinement, ensuring a concise yet highly informative feature set. In the model optimization phase, Genetic Algorithms (GAs) were applied to optimize the hyperparameters of machine learning models, including Artificial Neural Networks (ANNs), Random Forest (RF), and SVM. Bayesian Optimization (BO) was then employed to further refine these hyperparameters, enhancing overall model performance. Finally, the models were evaluated using key classification metrics, such as accuracy, Receiver Operating Characteristic (ROC) curves, and F1 scores, to ensure the robustness and reliability of the proposed approach.

Result: Among all models, the hybrid Random Forest model incorporating SHAP, PSO, GA, and BO demonstrated superior performance with 99.0% accuracy, a 94.8% F1-score, and an AUC of 1.00. The model also maintained high performance on the PIDD dataset, confirming its robustness and generalizability.

Discussion: The hybrid metaheuristic framework significantly improved prediction accuracy and efficiency for early T2DM diagnosis compared to conventional models. These findings support the growing evidence for integrating feature selection and optimization in clinical prediction. However, the study is limited by the use of publicly available datasets and lacks clinical validation, which should be addressed in future work.

Conclusion: The proposed hybrid metaheuristic framework offers a reliable and scalable solution for early diabetes prediction. It advances the application of AI in healthcare by improving diagnostic accuracy and supporting timely medical interventions. Future work should include clinical deployment, real-time validation, and dataset expansion for greater generalizability.

Keywords: Hybrid model, Metaheuristic optimization, Machine learning, SHapley Additive exPlanations (SHAP), Particle Swarm Optimization (PSO), Genetic algorithm.

© 2025 The Author(s). Published by Bentham Open.

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Public License (CC-BY 4.0), a copy of which is available at: <https://creativecommons.org/licenses/by/4.0/legalcode>. This license permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

* Address correspondence to this author at the Department of Computer Science, Banasthali Vidyapith-304022, Banasthali, Rajasthan, India; E-mail: sunilhit120@yahoo.com

Cite as: Upadhyay S, Gupta Y. Enhancing Early Diagnosis of Type II Diabetes through Feature Selection and Hybrid Metaheuristic Optimization Techniques. Open Bioinform J, 2025; 18: e18750362382139. <http://dx.doi.org/10.2174/0118750362382139250502100340>



Received: January 13, 2025
Revised: March 17, 2025
Accepted: March 27, 2025
Published: May 09, 2025



Send Orders for Reprints to
reprints@benthamscience.net

1. INTRODUCTION

Diabetes Mellitus Type II is a long-lasting metabolic disease characterized by high glucose in the blood due to the body not working correctly with insulin or does not produce enough insulin [1]. Due to its rapid rise, which is mostly caused by dietary changes, physical inactivity, and rising obesity rates, this condition has become a major global health challenge. According to a World Health Organization report, millions of people worldwide suffer from this disease [2]. Many health-related issues, such as cardiovascular disease, neuropathy, and kidney disease, are significantly increased by these complications [3]. T2DM has been a global health concern over the past three decades. The diabetes population went up from 108 million in 1980 to 800 million in 2024, and was a cause of death for 1.5 million people in 2019 [4, 5]. Genetic predisposition, lifestyle decisions, levels of physical activity, and dietary practices all impact the condition. Early detection lowers the risk of complications like vision loss and heart or kidney disease by enabling prompt intervention. Early diagnosis is also more cost-effective and significantly improves the quality of life for affected individuals. Timely and precise assessment of this disease is critical for mitigating risks and enabling appropriate interventions. A delayed diabetes diagnosis may require additional tests, such as random plasma glucose, glycated hemoglobin (A1C), and fasting plasma glucose. However, many individuals postpone these tests until they develop symptoms like polyuria, polydipsia, and polyphagia [6]. Conventional diagnostic techniques, such as oral glucose tolerance testing and fasting glucose testing, are frequently expensive, slow, and prone to errors, especially in the earlier stages of the disease. Machine learning (ML) and advanced computational methods have appeared as promising solutions to improve the accuracy and efficiency of diabetes prediction models [7]. ML can process large datasets rapidly, enabling prior prognosis and management of diabetes [8]. Patients can now monitor their blood sugar levels in real-time using devices like continuous glucose monitoring devices, which improve diabetes care and quality of life. Researchers are leveraging machine learning models to analyze datasets and improve the accuracy of diabetes prognosis [9]. ML applications in healthcare range from robotic surgery to prescription drug recommendations. By using feature selection and metaheuristic optimization techniques, this research focuses on improving the prediction of type II diabetes. In ML, feature selection is one of the key steps, especially when working with high-dimensional medical datasets. It helps in data reduction while keeping the most pertinent data, which improves the interpretability and efficacy of models. In this research, SVC assists in the feature selection process, and SHAP values and PSO are utilized to evaluate feature importance. Metaheuristic techniques like GA are used to balance and optimize the parameters of models, such as SVM, RF, and ANN, to enhance model performance further. These methods make it possible to explore a larger search space for hyperparameter tuning, which produces more precise and

effective predictions. Furthermore, hyperparameters are refined further using Bayesian optimization to guarantee the best possible model performance. A new method for type II diabetes predictive modeling is introduced by combining feature selection and metaheuristic optimization. The hybrid metaheuristic optimization model, integrating random forest with SHAP, PSO, GA, and Bayesian optimization, demonstrated superior performance and achieved the highest accuracy. According to this study's comparative analysis using metrics, such as ROC and accuracy scores, the model appeared as the most effective strategy, offering unparalleled accuracy, robustness, and computational efficiency. The findings of this study help in creating a scalable and trustworthy medical support system for early T2DM diagnosis.

The structure is such that a literature review is provided in Section 2, dataset and pre-processing are covered in Section 3, feature selection and hybrid model construction are covered in Section 4, results and performance evaluation are provided in Section 5, and the findings are concluded with future research directions in Section 6.

2. LITERATURE REVIEW

Diabetes is a global health concern, and in many parts of the world, more than 70% of the population suffers from diabetes. To predict and manage diabetes symptoms, many researchers have made use of machine learning and data mining methods. Pima Indians Diabetes Dataset is a dataset widely used in diabetes prediction in scientific research. Researchers have explored many methods, such as machine learning, neural networks, hybrid methods, and data mining, to forecast diabetes better. ANN is among the methods widely used in diabetes prediction models [10]. For instance, Swapna *et al.* [11] made diabetes predictions based on electrocardiogram information with the assistance of deep learning. Feature extraction in their study was dependent on a convolutional neural network (CNN), and an SVM was utilized later to fine-tune features. Their system was extremely accurate, with a rate of 95.7%. In addition, fuzzy cognitive maps were applied in knowledge-based systems to model diabetes-related knowledge, consequently improving the prediction ability in these models. Rastogi *et al.* [12] explored diabetes prediction using data mining techniques by applying Random Forest, SVM, logistic regression, and Naive Bayes on a real dataset. It was found that logistic regression achieved the highest accuracy of 82.46% in comparison to other models. Sisodia *et al.* [13] made use of the Pima Indians Diabetes Dataset in designing three machine learning models, namely decision tree, support vector machine, and naive Bayes, to classify diabetes with naive Bayes having a classification rate of 76.3%. Data mining was used in a research paper presented by Wu *et al.* [14] to identify risk factors for developing type 2 diabetes with a classification rate of 95.42%. The results of their experiments were found to be sensitive to the initial seed point value, which had a direct effect on the

outcomes. Zeinalnezhad *et al.* integrated data mining and meta-heuristic techniques to predict early readmission of diabetic patients within 30 days of discharge using a dataset from the UC Irvine Machine Learning Repository [15]. They employed classification algorithms, including Random Forest, neural network, and support vector machine, with a Genetic Algorithm (GA) for hyperparameter tuning. Their results demonstrated that GA-SVM improved prediction accuracy by 1.12%, highlighting the potential of optimized models in managing diabetic patient readmissions. Dharmarathne *et al.* introduced a self-explanatory interface for diabetes diagnosis using machine learning, incorporating four classification models: Decision Tree, K-Nearest Neighbor, Support Vector Classification, and Extreme Gradient Boosting (XGB) [16]. SHAP was utilized to interpret model predictions, with XGB demonstrating the highest accuracy. The integrated interface not only predicted diabetes but also provided transparent explanations, enhancing user awareness and supporting medical professionals in decision-making. Mumjudar and Vaidehi were interested in the prediction of diabetes using machine learning classifiers. Logistic regression emerged as the best-performing model based on classification metrics without pipelining [17]. When a pipeline was utilized to control and automate workflow, the AdaBoost classifier outperformed other models in diabetes prediction, especially in the Pima Indian Diabetes Dataset. Nicolucci *et al.* constructed diabetes complications prediction models based on electronic medical record information [18]. Their supervised machine learning model trained on data on 148 patients over a 15-year observation horizon in 23 centers was able to identify high-risk diabetes complications successfully. Ganie and Malik explored the prediction of diabetes independent of insulin based on life and biological predictors [19]. They made a machine learning ensemble with synthetic minority oversampling (SMOTE) to address a dataset with a sample size of 1,939 and 11 life and biological predictors. Their study identified urination as a prominent feature in the prediction of diabetes independent of insulin, with a bagged decision tree classifier performing better than other models. Bhat *et al.* reported a diabetes prediction and risk analysis using the SMOTE technique on the PIMA Indian Diabetes Dataset [20]. They identified diabetes risk contributors like blood pressure, glucose level, and diabetes pedigree function, with weight being the least contributor. Decision tree was the best performing among the classifiers used, with precision (96%), accuracy (91%), recall (92%), and F1-score (94%). Since more than 60% of diabetic patients are unaware of their conditions, early diagnosis is important to reduce laboratory visits and in-hospital admissions. Singh *et al.* analyzed the impact of data preparation on machine learning algorithms for type 2 diabetes prediction using two datasets: LS (locally developed) and PIMA (from Kaggle) [21]. They evaluated five machine learning models with eight scaling strategies, observing that PIMA dataset accuracy improved from 46.99–69.88% (without preprocessing) to 77.92% (with scaling). Similarly, Arvind *et al.* [22] explored the integration of feature selection

techniques with machine learning algorithms for diabetes classification by applying SVM, Random Forest, KNN, and Naïve Bayes after genetic algorithm-based feature selection. Their proposed ensemble model achieved a classification accuracy of 93.82%, demonstrating its effectiveness in improving prediction accuracy. In another research, Reza *et al.* proposed two stacking-based models for diabetes classification using the PIMA Indian Diabetes dataset, simulated data, and locally collected healthcare data [23]. They combined classical and deep neural network stacking ensemble methods, achieving the highest accuracy of 95.50% with a 5-fold CV on the simulation study. Their findings highlight the effectiveness of stacking ensembles in enhancing diabetes prediction accuracy and robustness. Similarly, Upadhyay *et al.* developed a web-based hybrid machine-learning model for diabetes prediction using a dataset from a reputed Indian hospital [24]. They proposed two hybrid models: one combining Support Vector Machine with bootstrap bagging and Reduced Error Pruning (SVMBBREP) and another integrating SVM with a genetic algorithm, with feature selection performed using the MRMR method. The SVMBBREP model achieved the highest accuracy of 99.67% and was incorporated into a web-based system, enabling real-time diabetes risk assessment and aiding early detection and management. Patile *et al.* proposed an ML-based framework, Improved Ensemble Learning with Dimensionality Reduction Model (IELDR), for early type 2 diabetes prediction [25]. The IELDR model, integrating an autoencoder-based feature extraction method with ensemble learning, was evaluated using the LS_diabetes dataset and validated on the Diabetes_2019 and PIMA diabetes datasets. The model achieved a high accuracy of 98.67%, outperforming other datasets, and demonstrated its effectiveness in predicting diabetes risk based on lifestyle patterns, aiding in early diagnosis and prevention. Nadesh and Arivuselvan developed a deep neural network and reported a performance level of 94.16% by employing feature reduction and feature selection to remove features with a potential negative contribution to the execution time. Feature importance was determined with decision trees and random forest models [26]. Spearman's correlation was adopted by Olisah *et al.* to select the most significant features in the PIMA dataset in a quest to improve the performance level of their diabetes prediction system [27]. However, it was found that Spearman's correlation was not efficient in treating nonlinear relations and multiple feature combinations. Ejiyi *et al.* addressed this limitation by using the SHAP algorithm to assess feature importance for diabetes prediction, achieving an accuracy of 94% on the limited dataset with eight attributes [28]. This highlights the critical role of feature selection in improving prediction models. Singh *et al.* developed a novel predictive model for diabetes, integrating advanced techniques to improve accuracy and robustness [29]. Their approach utilized IDBMI for missing value imputation, MFLOF for outlier detection, ASENN for class balancing, and a Multi-Model FusionNet Classifier for enhanced prediction. Validated on NHANES and PIMA Indian Diabetes datasets, the model achieved high accu-

racy (97.88% and 97.95%, respectively), demonstrating its effectiveness in addressing key challenges in diabetes detection (Table 1).

Due to its increasing prevalence, significant economic impact, and the enormous amount of data generated by its various forms and related complications, diabetes is predicted to become a major focus of future global health research. This emphasizes the necessity of effective techniques for handling and examining sizable datasets with lots of features. By using feature selection techniques and employing advanced metaheuristic optimization techniques, our study aimed to increase prediction accuracy. A thorough method was designed for big datasets to overcome the drawbacks mentioned in earlier

studies and allow for the early diagnosis of T2D. Particle Swarm Optimization and SHAP were utilized for feature selection after pre-processing the data. By keeping only the most pertinent features, computational complexity was decreased, and productivity was increased. Afterward, hybrid models were created by combining machine learning methods with genetic algorithms and Bayesian optimization to optimize hyperparameters before prediction. After a thorough evaluation of these models, the algorithm with the best performance was chosen based on important performance indicators. The objective of this approach was to enhance the accuracy, reliability, and computational efficiency of the type II diabetes prognosis systems (Table 2).

Table 1. Comparison of accuracy (%) of different classification algorithms using different feature selection techniques.

Authors/Refs.	Classifications Model	Feature Selection Methods	Accuracy (%)
Ananya <i>et al.</i> [30]	SVM, RF	Step forward and backward	81.4
Astuti <i>et al.</i> [31]	ANN, NB	BWOA	70
Amit <i>et al.</i> [32]	LR, KNN, NB, RF, SVM	Fast correlation-based filter feature selection	97.81
Saxena <i>et al.</i> [33]	DT, KNN, RF	RF	79
Rubaiat <i>et al.</i> [34]	ANN	RF	77
Tuan <i>et al.</i> [35]	SVM, DT, KNN, NBC RFC, LR	Wrapper-based feature selection utilizing Grey Wolf Optimization and an Adaptive Particle Swam Optimization	96

Note: FCBF: Fast Correlation-Based filter, BWOA: Binary Whale Optimization Algorithm, ANN: Artificial Neural Network, GA: Genetic Algorithm, RF: Random Forest, DT: Decision Tree, KNN: K Nearest Neighbour, SVM: Support Vector Machine.

Table 2. Comparison of previous works with findings of this study.

Authors/Refs.	Data-processing	Feature selection methods	Classification models	Optimization method to find the best parameter	Hyperparameter optimization	Hybrid model
Alam <i>et al.</i> [36]	Median, binning	PCA	RF, clustering, ANN, and association rule	NA	NA	NA
Kaur <i>et al.</i> [37]	KNN imputation, outlier removal	Boruta method	SVM (Linear), SVM-RBF, KNN, and ANN	NA	NA	NA
Zou <i>et al.</i> [38]	NA	PCA, mRMR	DT, RF, ANN	NA	NA	NA
Ahmed <i>et al.</i> [39]	Inter quartile range (IQR) and label encoding	Correlation and chi-square	DT, NB, KNN, RF, GB, LR, and SVM	NA	NA	NA
Selim <i>et al.</i> [40]	Min-max normalization	Data reduction unit	Gradient boosting machine	GBM-DRU	NA	Ensemble
Zhang <i>et al.</i> [41]	Mean, variance, median, quartile	Pearson correlation coefficient	RF, bagging, and boosting	Harmony search algorithm	NA	Hybrid
Ahmed <i>et al.</i> [42]	Pearson correlation, mutual information, SMOTE	Standardization, min max scaler, robust scaler	KNN, DT, RF, NB, LR, SVM, Ada boost, Extra Tree, Gradient boosting, LDA, ANN	NA	NA	Ensemble learning
Tanti <i>et al.</i> [43]	Mean, IQR	PCA	DT, GB, SVM, KNN, NB, RF	Cross-validation	NA	NA
Hossain <i>et al.</i> [44]	Median, outlier	mRMR, RFE	LR, RF, SVM, KNN, XGB	NA	NA	Hybrid
Gollapalli <i>et al.</i> [45]	Min max scaler, KNN imputer, SMOT	Correlation	SVM, RF, DT, KNN	NA	NA	Ensemble
Ganie <i>et al.</i> [46]	Outlier, SMOT	Correlation	Bagging, boosting, and voting	NA	NA	Bagging
Jamal <i>et al.</i> [47]	Mean, median	PCA	DT, RF, stochastic gradient boosting	Cross-validation	NA	NA
Our Research	Label encoding, correlation	SHAP, PSO	SVM, RF, ANN	GA	BO	Hybrid

Note: RF: Random Forest, ANN: Artificial Neural Network, SVM (Linear)- Support Vector Machine Linear, SVM-RBF: Support Vector Machine- Radial-Based Function, KNN: K Nearest Neighbour, ANN: Artificial Neural Network, DT: Decision Tree, mRMR: Minimum Redundancy Maximum Relevance, NB: Naive Bayes, GB: Gradient Boosting, LR: Logistic Regression, LDA: Linear Discriminant Analysis, RFE: Recursive Feature Elimination, XGB: Extreme Gradient Boosting, SMOT: Synthetic Minority Oversampling Techniques, BO: Bayesian Optimization.

3. MATERIALS AND METHODOLOGY

3.1. Data Collection

The dataset (Tables 3-6) used in this study, sourced from a Kaggle repository, was comprised of 520 records with 17 attributes, where 16 features were categorical, and “age” was the only numeric variable. The target variable, “class,” indicated whether an individual has diabetes. The dataset included key attributes, such as age, sex, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle incoordination, alopecia, obesity, and class, providing a comprehensive representation of diabetes-related symptoms. While the dataset offers valuable insights, its relatively small size may impact its generalizability. To enhance robustness, the Pima Indians Diabetes dataset was incorporated for external validation, allowing for a cross-dataset performance evaluation across diverse population groups. Stratified k-fold cross-validation (k=5) was also applied to improve model reliability and minimize dataset-dependent biases. To further strengthen the analysis, summary statistics and feature correlation analysis were carried out. They provided deeper insights into the dataset’s structure and relationships among features, ensuring a more rigorous assessment of the proposed hybrid metaheuristic optimization model.

3.2. Sample Size Determination

The adequacy of the sample size in this study is justified using stratified k-fold cross-validation (k=5), effect size considerations, and performance consistency analysis. Since the study relies on publicly available benchmark datasets, the Kaggle Diabetes Dataset (520 records, 17 features) and the Pima Indians Diabetes Dataset (768 records, 8 features), a formal power analysis was not conducted. However, the use of stratified k-fold cross-validation ensures that the dataset is effectively

utilized, reducing variability in performance metrics and enhancing reliability. Additionally, the dataset size aligns with prior studies in machine learning-based diabetes prediction, where a moderate to high effect size (Cohen’s d) is expected due to the strong relationship between diabetes risk factors and classification outcomes. Furthermore, the performance consistency across validation splits was analyzed to confirm that the results were not overly dependent on specific dataset partitions. The minimal variance in accuracy, precision, recall, and AUC across different folds indicated that the dataset size was sufficient for statistical reliability. Future studies can further strengthen statistical justification by conducting a formal power analysis by applying bootstrapping techniques and validating results on larger, real-world clinical datasets to improve generalizability.

3.3. Clinical Relevance of AI-Based Diabetes Prediction

In clinical practice, a critical question is whether AI-based diagnostic methods provide a significant advantage over traditional tests like fasting blood sugar and HbA1c, which are widely used for diabetes detection. While these conventional tests are effective, they primarily diagnose diabetes at later stages, when symptoms have already developed. In contrast, AI-driven models, such as the proposed hybrid metaheuristic approach, enable the early identification of high-risk individuals before symptoms appear. This proactive approach supports preventive healthcare strategies, allowing for early intervention and potentially reducing long-term diabetes-related complications. Additionally, AI models have the capability to analyze large-scale patient data and detect hidden patterns that might not be evident through conventional testing. By integrating AI with clinical diagnostics, healthcare professionals can enhance risk stratification and ensure that individuals at higher risk receive timely and targeted interventions. Rather than replacing traditional

Table 3. Dataset description used for research.

Feature Name	Description	Data Type
Age	Age of the individual	Integer
Sex	Gender of the individual	Categorical
Polyuria	Frequent urination	Categorical
Polydipsia	Excessive thirst	Categorical
Sudden Weight Loss	Unexpected weight loss	Categorical
Weakness	Persistent weakness or fatigue	Categorical
Polyphagia	Excessive hunger	Categorical
Genital Thrush	Fungal infection in the genital area	Categorical
Visual Blurring	Blurry vision	Categorical
Itching	Persistent skin itching	Categorical
Irritability	Increased irritability	Categorical
Delayed Healing	Slow healing of wounds	Categorical
Partial Paresis	Muscle weakness or partial paralysis	Categorical
Muscle Steadiness	Lack of muscle control	Categorical
Alopecia	Hair loss	Categorical
Obesity	BMI indicates obesity	Categorical
Class (Target)	Diabetes diagnosis (Positive/Negative)	Categorical

Table 4. Statistical summary of features.

-	Age	Gender	Polyuria	Polydipsia	Sudden Weight Loss	Weakness	Polyphagia	Genital Thrush	Visual Blurring	Itching	Irritability	Delayed Healing	Partial Paresis	Muscle Stiffness	Alopecia	Obesity	Class
Count	520	520	520	520	520	520	520	520	520	520	520	520	520	520	520	520	520
Mean	48.03	0.63	0.50	0.45	0.42	0.59	0.46	0.22	0.45	0.49	0.24	0.46	0.43	0.38	0.34	0.17	0.62
Std	12.15	0.48	0.50	0.50	0.49	0.49	0.50	0.42	0.50	0.50	0.43	0.50	0.50	0.48	0.48	0.38	0.49
Min	16.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.25	39.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.50	47.50	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
0.75	57.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00	1.00
Max	90.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 5. Pima indians diabetes dataset description.

Feature Name	Description	Data Type
Pregnancies	Number of times pregnant	Integer
Glucose	Plasma glucose concentration (mg/dL)	Numeric
Blood Pressure	Diastolic blood pressure (mm Hg)	Numeric
Skin Thickness	Triceps skinfold thickness (mm)	Numeric
Insulin	2-Hour serum insulin (mu U/ml)	Numeric
BMI	Body Mass Index (kg/m ²)	Numeric
Diabetes Pedigree	Diabetes pedigree function (Genetic Risk Factor)	Numeric
Age	Age of the individual (years)	Integer
Outcome (Target)	Diabetes diagnosis (0 = No, 1 = Yes)	Categorical

Table 6. PIMA dataset statistical summary of features.

Feature	Count	Mean	Std Dev	Min	0.25	0.50	0.75	Max
Pregnancies	768	3.85	3.37	0	1.00	3.00	6.00	17.00
Glucose	768	120.89	31.97	0	99.00	117.00	140.25	199.00
Blood Pressure	768	69.11	19.36	0	62.00	72.00	80.00	122.00
Skin Thickness	768	20.54	15.95	0	0.00	23.00	32.00	99.00
Insulin	768	79.80	115.24	0	0.00	30.50	127.25	846.00
BMI	768	31.99	7.88	0	27.30	32.00	36.60	67.10
Diabetes Pedigree Function	768	0.47	0.33	0.078	0.24	0.37	0.63	2.42
Age	768	33.24	11.76	21	24.00	29.00	41.00	81.00
Outcome (Target)	768	0.35	0.48	0	0.00	0.00	1.00	1.00

diagnostic tests, AI serves as a complementary tool that improves decision-making in diabetes screening, particularly in large-scale population studies and personalized medicine applications.

3.4. Data Pre-Processing

The data pre-processing phase ensures that the dataset is clean, balanced, and optimized for machine learning models. Since no missing values were found, imputation was not required. Categorical variables were converted into numerical representations using label encoding to facilitate model training. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, generating synthetic samples for the minority class to enhance model performance and prevent biased predictions. Feature scaling was performed using min-max scaling to normalize the data within the range of [0,1], ensuring consistency across features. Additionally, correlation analysis using the Pearson correlation coefficient was conducted to identify and remove highly correlated features, thereby reducing redundancy and improving model efficiency. The

correlation with diabetes (Fig. 1) indicates that polyuria and polydipsia are the most significant markers of diabetes. Conversely, features, such as gender and alopecia, appear to be weaker indicators and may negatively impact classification accuracy. For the PIMA Indian Diabetes dataset, similar pre-processing steps were applied to maintain consistency. Since the dataset primarily consisted of numerical values, min-max scaling was utilized to standardize feature distributions. Class imbalance was addressed using SMOTE to ensure an equal representation of positive and negative diabetes cases. Correlation analysis was also performed to eliminate redundant features, ensuring that only the most relevant attributes were retained for classification. These pre-processing steps enhanced data quality, leading to better generalization and improved performance of the machine learning models.

3.4.1. Synthetic Minority Over-sampling Technique (SMOTE)

Class imbalance is a common issue in medical datasets, where the number of positive (diabetic) and negative

(non-diabetic) cases is significantly different. This imbalance can lead to biased model predictions, where the classifier favors the majority class. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was implemented in this study. Instead of merely duplicating instances of the minority class, SMOTE generates synthetic samples by interpolating feature values between

existing minority class instances. This approach helps the model learn meaningful patterns from both classes, improving classification performance while preventing overfitting. By integrating SMOTE, the model can better distinguish between diabetic and non-diabetic cases, reducing bias and enhancing predictive performance (Fig. 2).

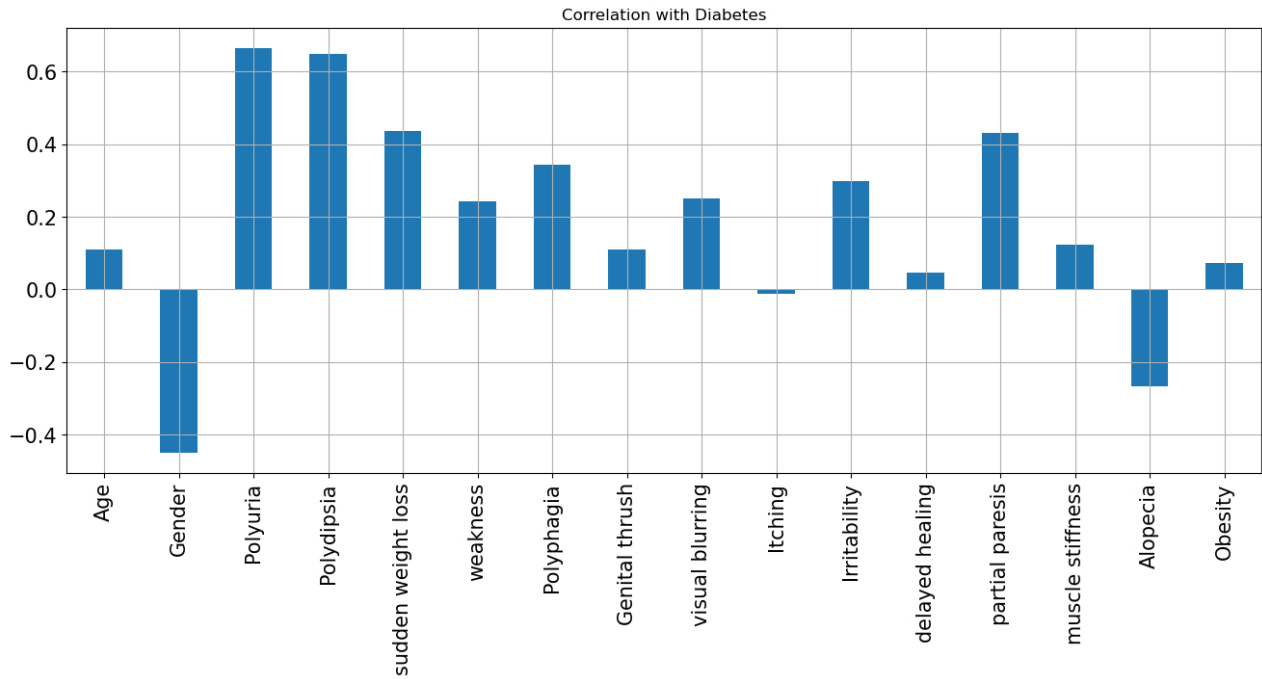


Fig. (1). Correlation with diabetes.

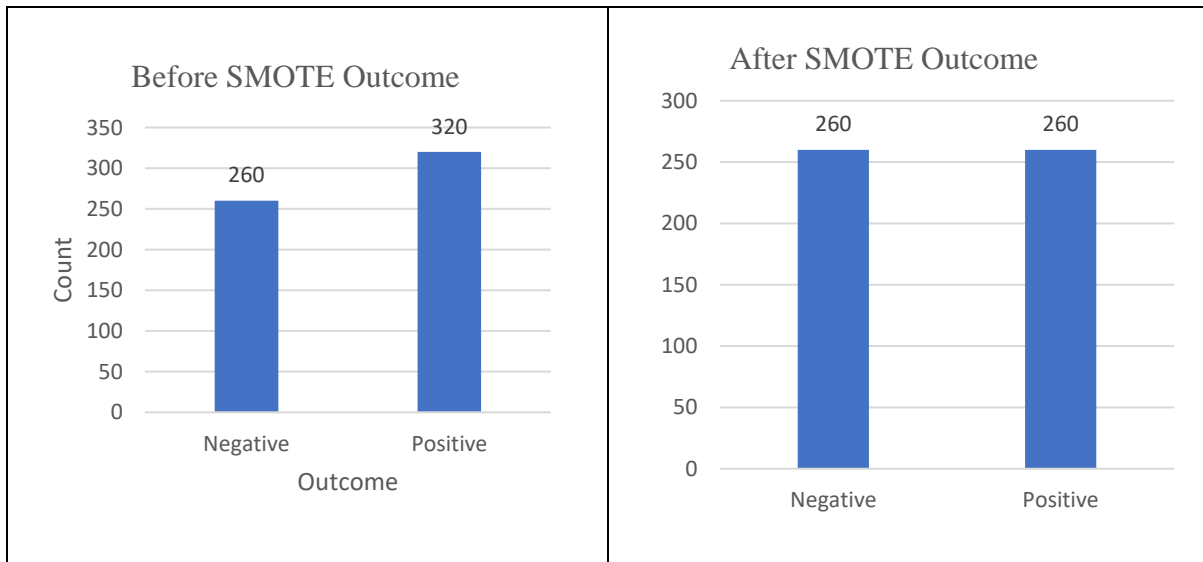


Fig. (2). SMOTE Outcome: Before and after.

3.5. Feature Selection

A crucial phase in data analysis is feature selection. Choosing the most relevant features from the initial set by predetermined evaluation criteria entails lowering the dimensionality of a dataset. By removing unnecessary features, this procedure streamlines the dataset while preserving crucial information. The set of attributes N includes n attributes, $\{n_1, n_2, n_3, \dots, n_k\}$ [48, 49]. It is the goal of feature selection to select k -relevant attributes from the set. Subset generation is the first of the three required procedures in the feature selection process. It is the process of constructing unique subsets of attributes from the given set. The evaluation phase follows, during which each subset's quality is evaluated based on predetermined evaluation criteria to ascertain its significance and relevance. When an ideal subset is found, the search is stopped using stopping rules, and the chosen features are then validated to make sure they are appropriate for the predictive model. This ends the process. By ensuring that only the most relevant features are used, this hierarchical method enhances the accuracy and productivity of the ensuing predictive model.

3.5.1. Features Selection by SHAP with SVM

Support Vector Machines, in conjunction with SHAP, allow for efficient feature selection by locating the most important dataset features. To maximize class separation, a hyperplane is optimized before an SVM model is trained on all features. Kernel SHAP, which measures each feature's contribution to predictions, is then used to calculate SHAP values [50]. A global ranking of feature importance is obtained by aggregating these values. The SVM model is retrained using the top k features, which lowers dimensionality and improves computational efficiency without compromising accuracy [51]. This integration highlights important aspects of decision-making and produces an accurate, interpretable model.

Mathematical notation is presented in Eq. (1), which is as follows:

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (1)$$

Where: $\phi_i(f, x)$ is the SHAP value of the feature i for the prediction $f(x)$, N is the set of all features, S is a subset of the features that exclude the feature i , $|S|!$ is the factorial of the size of the subset S , $|N|!$ is the factorial of the total number of features, $f(S \cup \{i\})$ is the model's prediction using the features in S plus feature i , and $f(S)$ is the model's prediction using only the features in S .

Feature selection with SHAP involves calculating the mean absolute SHAP values across all training examples, providing a global ranking of feature importance influencing the model's predictions (Eq. 2).

$$\text{Mean SHAP value for feature } i = \frac{1}{n} \sum_{j=1}^n |\phi_i(f, x_j)| \quad (2)$$

Where n is the total number of training instances, $\phi_i(f,$

$x_j)$ is the SHAP value of feature i , for instance x_j , and $f(x_j)$ is the model's prediction, for instance x_j .

According to the SHAP value displayed in Fig. (3), the top 10 features selected include polyuria, polydipsia, gender, partial paresis, itching, age, genital thrush, irritability, sudden weight loss, visual blurring, and weakness.

3.5.2. Particle Swarm Optimization Feature Selection Method

Particle Swarm Optimization, which finds the most relevant features in a dataset, was inspired by the collective behaviour of flocking birds. PSO effectively finds the best feature subsets to enhance the productivity of models, and it works especially well in high-dimensional spaces [52]. Each particle in the process represents a possible feature subset encoded as a binary vector (1 for selected features, 0 otherwise), and the population is initialized at random at the start of the process. Fitness scores obtained from model metrics, such as AUC, are used to assess particles. Particles explore the feature space and migrate toward the optimal subset for the optimal model's efficiency by iteratively updating their position as well as their velocity using their optimal position individually as well as the optimal position of the swarm (Eq. 3) [53].

$$v_i(t+1) = w \cdot v_i(t) + c_1 \cdot r_1 \cdot (p_{\text{best},i} - x_i(t)) + c_2 \cdot r_2 \cdot (g_{\text{best}} - x_i(t)) \quad (3)$$

Where, $v_i(t+1)$ is the velocity of particle i at time $t+1$, w is the inertia weight to control the exploration and exploitation balance, c_1 , c_2 is the cognitive and social learning coefficients, controlling the influence of personal and global best positions, r_1 , r_2 is the random value between 0 and 1, $p_{\text{best},i}$ is the personal best position of particle i , g_{best} is the best global position among all particles, and $x_i(t)$ is the current position of particle i .

The position (selected features) of each particle is updated based on the updated velocity (Eq. 4).

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (4)$$

A sigmoid function is typically applied to convert continuous velocities to binary decisions for feature selection (Eq. 5).

$$S(v_i(t+1)) = \frac{1}{1 + e^{-v_i(t+1)}} \quad (5)$$

The new position is determined by comparing $S(v_i(t+1))$ to a random number, with values closer to 1 leading to the feature being selected.

The process repeats for a fixed number of iterations or until convergence is achieved (*i.e.* when the particles no longer improve their positions significantly).

The top 10 features selected include gender, polyuria, polydipsia, weakness, polyphagia, itching, irritability, delayed healing, partial paresis, and muscle stiffness.

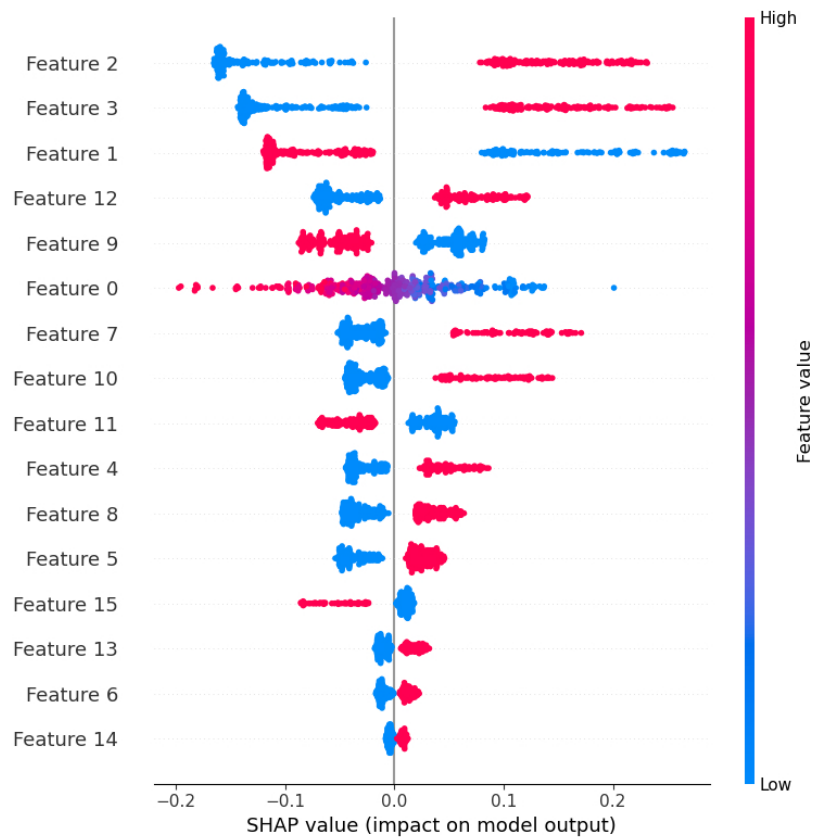


Fig. (3). SHAP value for feature selection.

3.6. Framework of the Proposed Methodology

The proposed framework consists of six key stages: data collection, data pre-processing, feature selection, model optimization, classification modeling, and model evaluation, as illustrated in Fig. (4).

3.6.1. Step 1

Data Collection: Publicly available datasets, including the Kaggle Diabetes dataset and Pima Indians Diabetes Dataset, are used.

3.6.2. Step 2

Data Pre-processing: This includes handling missing values, encoding categorical variables, feature scaling, and class balancing using SMOTE. Correlation analysis is performed to remove redundant features.

3.6.3. Step 3

Step 3- Feature Selection: SHAP (SHapley Additive exPlanations) ranks feature importance, while Particle Swarm Optimization (PSO) selects the most relevant subset for model training.

3.6.4. Step 4

Step 4- Model Optimization: Genetic Algorithm (GA) and Bayesian Optimization refine model hyperparameters, while stratified k-fold cross-validation (k=5) ensures model reliability and reduces overfitting.

3.6.5. Step 5

Step 5- Classification Models: The study implements Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest (RF) models, along with their optimized hybrid versions.

3.6.6. Step 6

Step 6- Model Evaluation: Performance is assessed using accuracy, precision, recall, F1-score, and AUC-ROC. Additionally, cross-validation and external validation with PIDD confirm the generalizability of the hybrid metaheuristic approach.

4. MACHINE LEARNING MODELS

4.1. Artificial Neural Network (ANN)

ANN is a type of model that imitates brain activity to process data and produce predictions. It is modeled after biological neural networks. The three fundamental elements include an output layer making the ultimate classifications or predictions, an input layer taking in the input data, and the hidden layers carrying out the intermediate transformations [54]. The layers are all connected with weights that adjust with the training process to maximize accuracy. During the learning of intricate patterns, every neuron calculates a weighted sum of the input along with utilizing an activation function. This creates non-linearity. ANNs reduce prediction errors through iterative training and optimization methods, gradually improving performance.

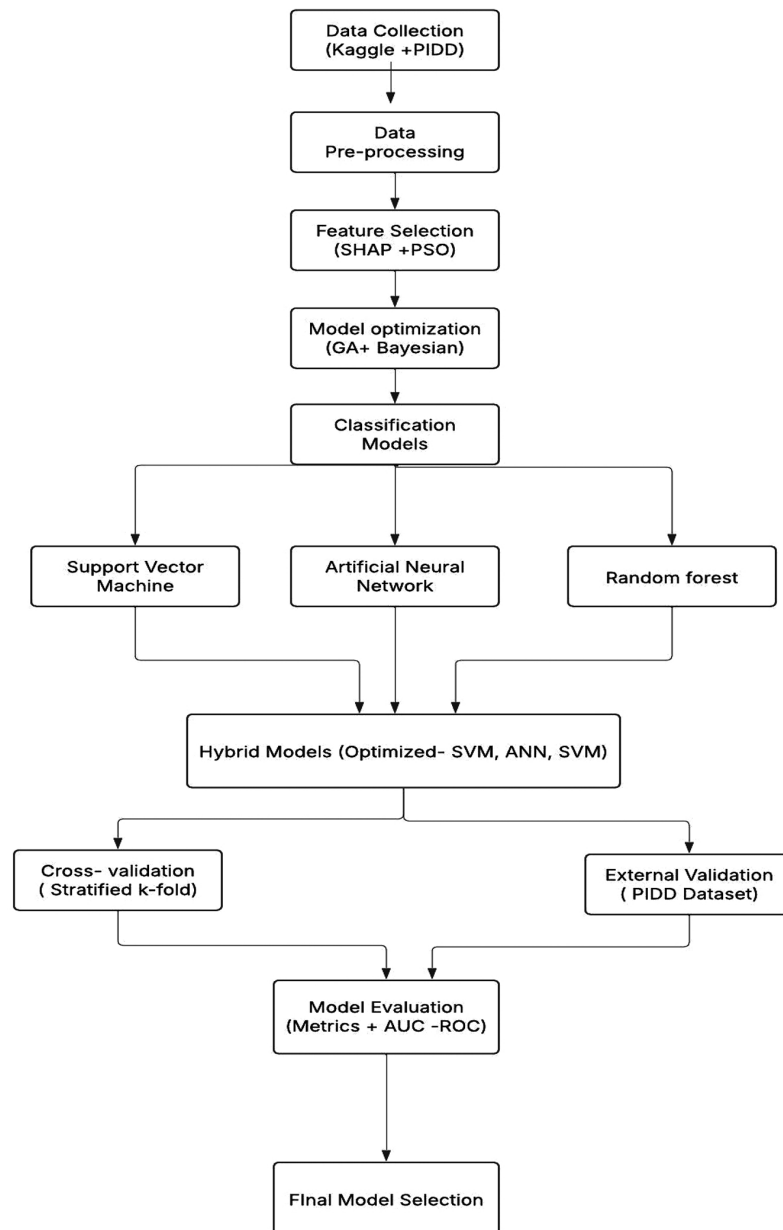


Fig. (4). Proposed framework for model.

For calculating a neuron i in the hidden layer, Eq. (6) is used, which is as follows:

$$z_i = \sum_{j=1}^n w_{ij} x_j + b_i \quad (6)$$

Where, x_j is the input to the neuron, w_{ij} is the weight associated with the inputs, b_i is the biased term, and z_i is the weighted sum.

After computing the weighted sum, an activation function is applied to introduce non-linearity to the model. Common activation functions include Sigmoid, which is shown in Eq. (7).

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (7)$$

Output: The final output is computed by applying the same process to the output layer. Training an ANN involves adjusting the weights and biases using backpropagation and an optimization method like gradient descent. The goal is to minimize the loss function. The weights are updated using the following Eq. (8):

$$w_{ij} = w_{ij} - \eta \cdot \frac{\partial L}{\partial w_{ij}} \quad (8)$$

4.2. Bayesian Optimization

This method is used for optimizing problems where the objective function is difficult to evaluate and has an unknown form. It is especially helpful for adjusting hyperparameters in machine learning. The surrogate model, acquisition function, and objective function are the

three main elements of the process. A performance metric (such as accuracy or loss) that depends on hyperparameters x is represented by the objective function, $f(x)$. This model uses a surrogate model to approximate the function based on previous evaluations because evaluating $f(x)$ is computationally costly. To choose the next point, the acquisition function, such as the upper confidence bound or expected improvement, balances exploration and exploitation. The algorithm iteratively assesses $f(x)$, updates the surrogate model, and refines the solution with each iteration [55].

Let $f(x)$ be the objective function where x is the set of hyperparameters. Bayesian optimization aims to find the value of x that minimizes $f(x)$ Eq. (9):

$$x^* = \operatorname{argmin} f(x) \quad (9)$$

Since $f(x)$ is expensive to compute, we model $f(x)$ using a Gaussian Process $g(x)$, and then choose the next point to evaluate by maximizing the acquisition function $a(x)$ Eq. (10):

$$x_{\text{next}} = \operatorname{argmax} a(x) \quad (10)$$

The acquisition function guides the search process, deciding where to explore next based on the current model's uncertainty and predictions.

4.3. Random Forest

A widely used ensemble learning algorithm for classification and regression tasks, Random Forest is widely used for its robustness and ability to handle high-dimensional data. Multiple decision trees are constructed during training, and the ultimate prediction is made using majority voting (for classification) or averaging (for regression) [56]. This ensemble approach helps reduce overfitting and improves accuracy. Two fundamental types of randomness are used to build each Random Forest decision tree: Random feature selection, which considers only an arbitrary portion of features at each split, and Bagging, which trains each tree on an arbitrary portion of the data. This reduces variance and strengthens the model by decorating the trees. Random feature selection further increases tree diversity, while bootstrapping guarantees that every tree is exposed to various subsets of data. With the majority voting for classification or averaging for regression, Random Forest merges the predictions of all trees in the final step, leading to a more robust and accurate model that performs very well on complex datasets [57].

Let $h_i(x)$ be the prediction of the i -th decision tree for input x , and let N be the number of trees. The final prediction for Random Forest is the majority vote, as shown in Eq. (11):

$$H(x) = \operatorname{mode}(h_1(x), h_2(x), \dots, h_N(x)) \quad (11)$$

4.4. Genetic Algorithm

This is a natural selection-inspired enhancing method used to identify the best answers in challenging search spaces. It begins with a population of chromosomes, each

representing a potential solution. Genes, which are decision variables and are frequently encoded as binary, integer, or real values, make up chromosomes [58]. Each person's performance is assessed by a fitness function, which favours those with higher fitness scores to guide selection. To produce children, a selected group of people (parents) go through crossover, exchanging chromosome segments. Small, random changes are introduced by mutation to preserve diversity. To effectively identify the best answers, this process iterates over generations, striking a balance between exploration and exploitation [59].

The mathematical representation of this process is as follows:

Chromosome: $x = (x_1, x_2, \dots, x_n)$, where each x_i represents a gene (decision variable), Fitness Function: $f(x)$, where f evaluates the quality of the solution x , Crossover: Two parent chromosomes x^1 and x^2 are combined to form offspring x^{new} , often using a point or uniform crossover, and Mutation: With a small probability, some genes in x^{new} are altered randomly.

4.5. Hybrid Metaheuristic Optimization Methods and Experimentation Process

This study explores three hybrid metaheuristic optimization techniques that enhance classification performance for diabetes prediction by integrating SHAP, Particle Swarm Optimization (PSO), Genetic Algorithm (GA), and Bayesian Optimization with Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest (RF). Each method follows a structured experimental process involving feature selection, feature refinement, and hyperparameter tuning.

4.5.1. Hybrid Metaheuristic Optimization Integrating SVM with SHAP, PSO, GA, and Bayesian Optimization

In the first approach, the hybrid metaheuristic optimization integrates SVM with SHAP, PSO, GA, and Bayesian Optimization to optimize diabetes classification. The process begins with SHAP, which analyzes feature importance using a linear SVM model, allowing the selection of the most relevant features. These selected features are then refined using PSO, where each particle represents a subset of features and is evaluated based on the AUC score of an SVM model trained on that subset. PSO iteratively converges to the best feature subset, improving classification performance. Next, GA is used to optimize hyperparameters, such as C and γ , by evolving parameter sets across generations. Bayesian Optimization is then applied to fine-tune these hyperparameters, balancing exploration and exploitation to maximize the AUC score. The final optimized SVM model, trained with an RBF kernel, is evaluated using a Receiver Operating Characteristic (ROC) curve and classification metrics, including accuracy, precision, recall, and F1-score.

4.5.1.1. Experimental Process**4.5.1.1.1. Feature Selection Using SHAP**

A linear SVM model is trained, and SHAP (SHapley Additive exPlanations) values are computed to determine feature importance.

Only the most significant features are selected to reduce dimensionality and improve efficiency.

4.5.1.1.2. Feature Refinement Using PSO

Each particle represents a subset of features, and the fitness of each subset is evaluated using the AUC score of an SVM trained on it.

PSO iteratively converges toward the optimal feature subset for better classification.

4.5.1.1.3. Hyperparameter Optimization Using GA

The genetic algorithm is applied to optimize SVM hyperparameters (C and gamma).

The best hyperparameter sets are selected through iterative evolution, enhancing classification accuracy.

4.5.1.1.4. Fine-tuning with Bayesian Optimization

Bayesian Optimization further refines the hyperparameters, balancing exploration and exploitation to maximize AUC.

4.5.1.1.5. Model Evaluation

The final optimized SVM model (with an RBF kernel) is trained on the selected features and fine-tuned hyperparameters.

Performance is assessed using a Receiver Operating Characteristic (ROC) curve and classification metrics, including precision, recall, F1-score, and accuracy.

4.5.2. Hybrid Metaheuristic Optimization Integrating ANN with SHAP, PSO, GA, and Bayesian Optimization

The second approach enhances ANN performance using a similar hybrid metaheuristic optimization framework. Initially, SHAP is employed to rank feature importance in a linear SVM model, selecting the top 10 most relevant features. These features are further refined using PSO, where particles represent feature subsets, and their fitness is determined based on the AUC score of an ANN trained on them. Through iterative optimization, PSO identifies the best-performing feature subset. GA is then utilized to optimize key ANN hyperparameters such as the number of neurons, batch size, and learning rate by evolving parameter combinations over multiple generations. Bayesian Optimization further fine-tunes these hyperparameters to ensure optimal performance. The ANN model, trained with the optimized features and hyperparameters, is assessed using an AUC score, ROC curve, and classification report summarizing accuracy, precision, recall, and F1-score.

4.5.2.1. Experimental Process**4.5.2.1.1. Feature Selection using SHAP**

A linear SVM model is trained, and SHAP values are computed to rank feature importance.

The top 10 most significant features are selected for better model efficiency.

4.5.2.1.2. Feature Refinement using PSO

Each particle represents a subset of features, and its fitness is determined by the AUC score of an ANN trained on that subset.

PSO iteratively searches for the optimal feature subset to enhance predictive performance.

4.5.2.1.3. Hyperparameter Optimization using GA

The Genetic Algorithm is employed to optimize key ANN parameters such as the number of neurons, batch size, and learning rate.

The most effective hyperparameter combinations are selected through evolutionary processes.

4.5.2.1.4. Fine-tuning with Bayesian Optimization

Bayesian Optimization further fine-tunes ANN hyperparameters by balancing exploration and exploitation.

4.5.2.1.5. Model Evaluation

The ANN model is trained using the optimal hyperparameters and PSO-optimized features.

Performance is measured using the ROC curve, AUC score, and a classification report summarizing accuracy, precision, recall, and F1-score.

4.5.3. Hybrid Metaheuristic Optimization Integrating Random Forest with SHAP, PSO, GA, and Bayesian Optimization

The third approach integrates Random Forest with SHAP, PSO, GA, and Bayesian Optimization to enhance diabetes prediction. SHAP is first applied to a linear SVM model to determine feature importance, selecting the top 10 influential features. These features undergo further refinement through PSO, where each particle represents a subset, and its fitness is evaluated using the AUC score of a Random Forest model trained on that subset. PSO iteratively optimizes the feature selection process, leading to improved classification accuracy. GA is then applied to fine-tune Random Forest hyperparameters, such as the number of trees and maximum depth, by evolving optimal parameter sets. Finally, Bayesian Optimization further adjusts these hyperparameters, balancing exploration and exploitation for improved AUC scores. The final optimized Random Forest model is evaluated using the ROC curve, AUC score, and a classification report highlighting accuracy, precision, recall, and F1-score.

4.5.3.1. Experimental Process

4.5.3.1.1. Feature Selection Using SHAP

A linear SVM model is trained, and SHAP values are calculated to determine feature importance.

The top 10 most influential features are selected to reduce dimensionality while preserving model accuracy.

4.5.3.1.2. Feature Refinement Using PSO

Each particle represents a subset of features, and its fitness is evaluated based on the AUC score of a Random Forest model trained on that subset.

PSO iteratively optimizes the feature subset to improve classification performance.

4.5.3.1.3. Hyperparameter Optimization Using GA

Genetic Algorithm optimizes Random Forest hyperparameters, such as the number of trees and maximum depth.

Through multiple generations, GA selects the best-performing hyperparameter combinations.

4.5.3.1.4. Fine-tuning with Bayesian Optimization

Bayesian Optimization further refines hyperparameters by striking a balance between exploration and exploitation to maximize AUC scores.

4.5.3.1.5. Model Evaluation

The final Random Forest model is trained using the optimized features and fine-tuned hyperparameters.

Performance is assessed using the ROC curve, AUC score, and a classification report detailing precision, recall, F1 score, and accuracy.

5. RESULTS AND DISCUSSION

This section presents and discusses the outcomes obtained from the experimental setup using various models for diabetes prediction.

5.1. Correlation Coefficient Analysis

The correlation matrix provides valuable insights into the relationships between features in the diabetes dataset using Pearson's correlation coefficient (Fig. 5). A value close to +1 signifies a strong positive correlation, -1 represents a strong negative correlation, and 0 indicates no correlation. The heatmap, which illustrates feature correlations within the dataset after pre-processing, visually represents these relationships. Darker shades in the heatmap indicate stronger correlations between features. Key observations reveal strong positive correlations between features, such as polyuria and polydipsia, as well as sudden weight loss and weakness, suggesting their tendency to occur together. Conversely, features like age and alopecia or obesity and partial

paresis exhibit weak or negative correlations, while others, such as visual blurring and genital thrush, show little to no association. The rightmost column of the correlation matrix highlights the relationship of each feature with the diabetes outcome, where polyuria, polydipsia, and partial paresis demonstrate strong positive correlations, making them crucial predictors. The correlation matrix is instrumental in feature selection, as it helps identify redundant variables, minimize multicollinearity, and enhance model efficiency. By eliminating highly correlated features, the risk of overfitting is reduced, ultimately leading to more accurate diabetes predictions.

5.2. Performance Evaluation of Machine Learning Models

A comprehensive comparison between standard machine learning models (SVM, ANN, and RF) and their hybrid optimized counterparts (Hybrid-SVM, Hybrid-ANN, and Hybrid-RF) was carried out to assess improvements in predictive performance. Key evaluation metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, were analyzed both before and after optimization to highlight the impact of the applied hybrid techniques. To ensure the reliability and generalizability of the models, stratified k-fold cross-validation was implemented, allowing performance assessment across multiple subsets of the dataset and reducing bias. Additionally, external validation using the Pima Indians Diabetes dataset was performed to verify model robustness beyond the Kaggle dataset, ensuring that the optimized models maintain high performance across diverse data sources. The results demonstrated that hybrid-optimized models consistently outperformed their standard counterparts, highlighting the effectiveness of optimization techniques in improving diabetes prediction accuracy (Table 7).

The optimized models exhibited a notable performance boost compared to their non-optimized versions. Hybrid-SVM and Hybrid-ANN demonstrated significant improvements, while hybrid-RF was the most effective model, achieving 99.0% accuracy and an AUC-ROC of 1.00 (Figs. 6-7), indicating exceptional predictive capability for diabetes diagnosis. Furthermore, the application of cross-validation ensured that the models maintained generalizability and robustness, effectively minimizing the risk of overfitting across different dataset partitions.

External Validation on PIDD Dataset

To evaluate the robustness and generalizability of the models, they were tested on the Pima Indians Diabetes dataset. The results confirmed that the hybrid models maintained high predictive performance even on an independent dataset, further validating the effectiveness of the hybrid optimization approach (Table 8).

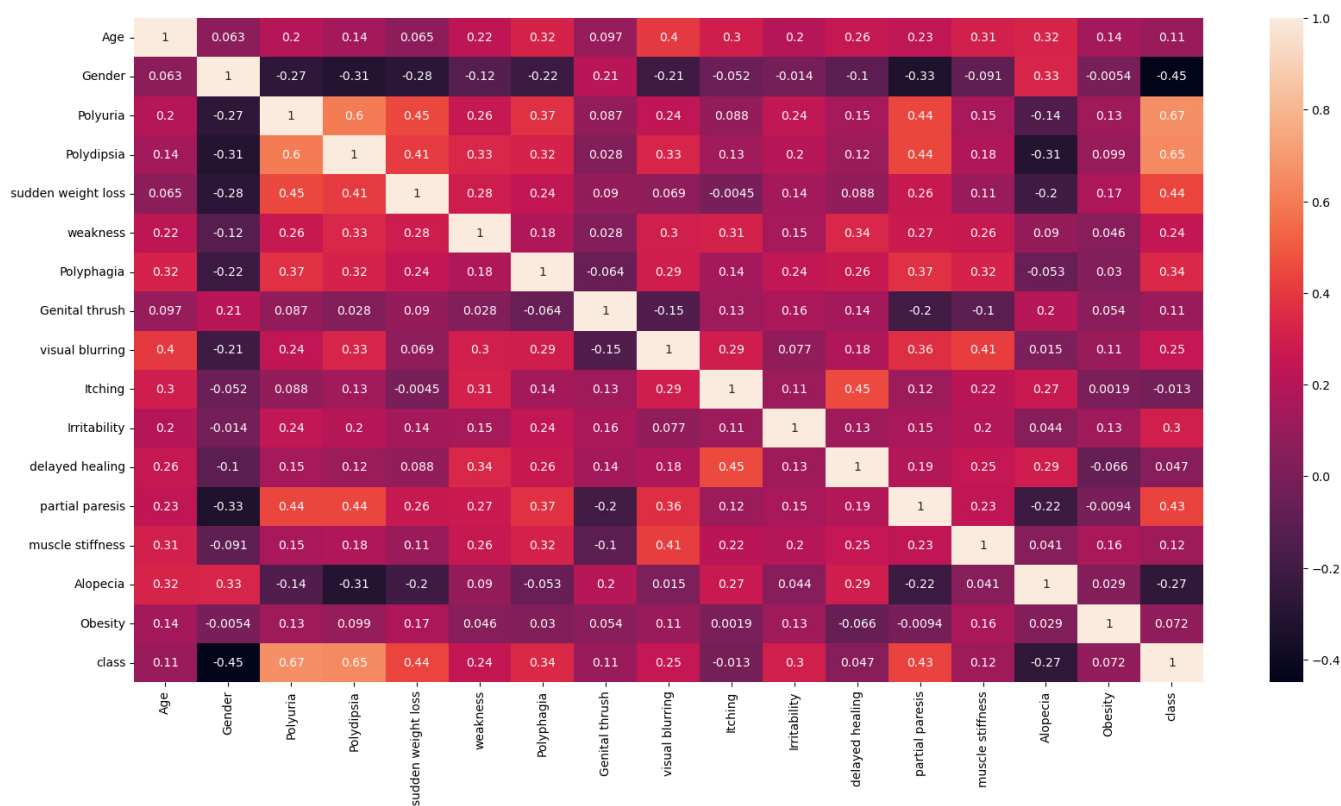


Fig. (5). Correlation matrix of diabetes.

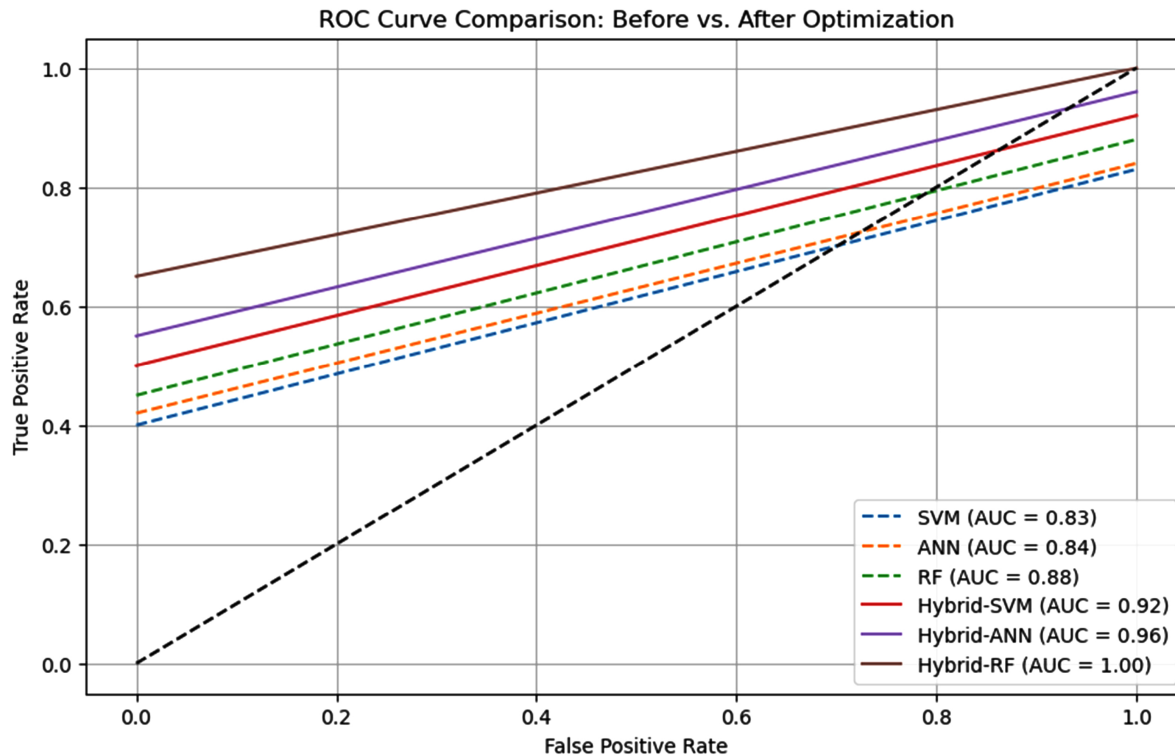


Fig. (6). ROC curve before and after optimization on the diabetes dataset.

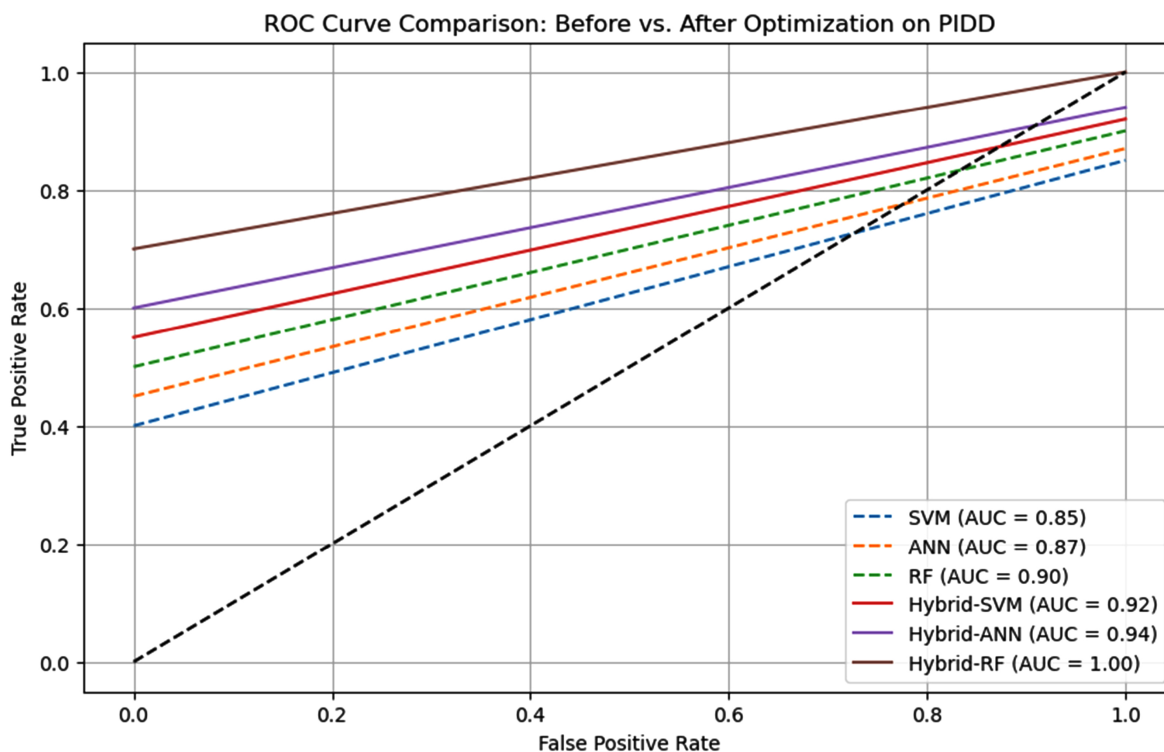


Fig. (7). ROC curve before and after optimization on the PIMA dataset.

Table 7. A comparative analysis of the models before and after optimization on the dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
SVM (without optimization)	80.0	79.0	80.5	82.7	0.83
ANN (without optimization)	82.2	80.8	81.0	84.4	0.84
RF (without optimization)	85.0	82.2	85.8	86.2	0.88
Hybrid-SVM (optimized after cross-validation)	94.0	95.2	94.4	94.8	0.92
Hybrid-ANN (optimized after cross-validation)	93.0	92.8	93.6	93.2	0.96
Hybrid-RF (optimized after cross-validation)	99.0	93.9	92.5	93.2	1.00

Table 8. Performance metrics comparison on PIDD dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
SVM (without optimization)	82.5	81.2	80.9	81.0	0.85
ANN (without optimization)	84.3	83.0	83.2	83.1	0.87
RF (without optimization)	87.1	85.8	86.2	86.0	0.90
Hybrid-SVM (optimized after cross-validation)	89.0	90.0	89.3	89.6	0.92
Hybrid-ANN (optimized after cross-validation)	89.8	89.5	90.1	89.8	0.94
Hybrid-RF (optimized after cross-validation)	99.0	98.7	99.5	99.1	1.00

The results of standard machine learning models (SVM, ANN, and RF) and their hybrid optimized counterparts (Hybrid-SVM, Hybrid-ANN, and Hybrid-RF) were evaluated to assess improvements in predictive performance by considering the following aspects [60].

- Accuracy: Measures the proportion of correct predictions.

$$\text{Accuracy} = (\text{True Positive} + \text{True Negative}) / (\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative})$$

- Precision: Evaluates the ability to avoid labeling negative cases as positive.

$$\text{Precision} = \text{True Positive} / (\text{True Positive} + \text{False Positive})$$

- Recall: Measures the ability to correctly identify positive cases.

$$\text{Recall} = \text{True Positive} / (\text{True Positive} + \text{False Negative})$$

- F1 Score: Provides a harmonic mean of precision and recall.

$$\text{F1-Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

CONCLUSION

Diabetes is a chronic disease affecting millions worldwide. This study evaluated the effectiveness of hybrid metaheuristic optimization techniques in enhancing diabetes prediction using machine learning models. A comprehensive comparison between conventional models (SVM, ANN, and RF) and their optimized counterparts (Hybrid-SVM, Hybrid-ANN, and Hybrid-RF) demonstrated notable improvements in classification performance. The integration of feature selection using SHAP and PSO, class balancing through SMOTE, and performance validation *via* stratified k-fold cross-validation contributed to enhanced accuracy, precision, and generalizability. Among the optimized models, Hybrid-RF achieved the highest accuracy of 99.0% with an AUC-ROC score of 1.00, underscoring the effectiveness of the proposed approach. Despite these advancements, several limitations must be considered.

STUDY LIMITATIONS

The study relies on publicly available datasets (Kaggle Diabetes dataset and Pima Indians Diabetes dataset), which may not fully capture diverse populations due to the absence of detailed patient demographics and clinical variables. This limitation may affect the generalizability of the model. Additionally, feature selection using SHAP and PSO, while improving model efficiency, may inadvertently exclude subtle yet clinically relevant features. Another challenge is the potential dataset bias, which has not been explicitly assessed and could influence the model's predictive reliability. Moreover, the proposed model has not undergone real-world clinical validation, raising concerns about its practical applicability, physician interpretability, and seamless integration into healthcare systems. The computational complexity associated with

hybrid metaheuristic optimization also increases processing requirements, making real-time deployment in resource-constrained environments challenging. Although 5-fold cross-validation enhances performance estimation, further external validation using larger, real-world hospital datasets is crucial to confirm the robustness and clinical utility of the model.

Future research should address these limitations by incorporating more diverse and extensive datasets, conducting real-world testing, and optimizing computational efficiency. Additionally, exploring federated learning can improve model generalizability while maintaining data privacy. The proposed framework can also be extended to predict the likelihood of diseases at early stages. In the future, mobile and web applications based on this model could assist healthcare providers in early diabetes detection and prediction, ultimately improving patient outcomes and supporting timely medical interventions.

AUTHORS' CONTRIBUTIONS

The authors confirm contribution to the paper as follows: S.U: Writing Paper; Y.K.G: Conceptualization;. All authors reviewed the results and approved the final version of the manuscript.

LIST OF ABBREVIATIONS

T2DM	= Type-II Diabetes Mellitus
SVMs	= Support Vector Machines
PSO	= Particle Swarm Optimization
GAs	= Genetic Algorithms
RF	= Random Forest
ANNs	= Artificial Neural Networks
BO	= Bayesian Optimization
ROC	= Receiver Operating Characteristic
SMOTE	= Synthetic Minority Over-sampling Technique

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The data supporting the findings of the article is available in the Kaggle Repository at <https://www.kaggle.com/datasets/andrewmvd/early-diabetes-prediction-dataset>, reference number: KAG-EDP-520. Additionally, the Pima Indians Diabetes dataset used for validation is available at <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>, reference number: KAG-PIMA-768

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

- [1] Daliya VK, Ramesh TK. A cloud based optimized Ensemble model for risk prediction of diabetic progression-An Azure Machine Learning perspective. *IEEE Access* 2025; 13: 11560-75. <http://dx.doi.org/10.1109/ACCESS.2025.3528033>
- [2] Shojaee-Mend H, Velayati F, Tayefi B, Babaee E. Prediction of diabetes using data mining and machine learning algorithms: A cross-sectional study. *Healthc Inform Res* 2024; 30(1): 73-82. <http://dx.doi.org/10.4258/hir.2024.30.1.73> PMID: 38359851
- [3] Dagliati A, Marini S, Sacchi L, *et al.* Machine learning methods to predict diabetes complications. *J Diabetes Sci Technol* 2018; 12(2): 295-302. <http://dx.doi.org/10.1177/1932296817706375> PMID: 28494618
- [4] Diabetes. 2025. Available from: <https://www.who.int/news-room/fact-sheets/detail/diabetes>
- [5] King J, Aubert RE, Herman WH. Global Burden of Diabetes, 1995-2025: Prevalence, numerical estimates, and projections. *Diabetes Care* S 2025; 21(9): 1414.S <http://dx.doi.org/10.2337/diacare.21.9.1414>
- [6] Zhang J, Fan X, Xu Y, *et al.* Association between inflammatory biomarkers and mortality in individuals with type 2 diabetes: NHANES 2005-2018. *Diabetes Res Clin Pract* 2024; 209: 111575. <http://dx.doi.org/10.1016/j.diabres.2024.111575> PMID: 38346591
- [7] Sharma T, Shah M. A comprehensive review of machine learning techniques on diabetes detection. *Visual Comput Indust Biomedicine Art* 2021; 4(1): 30. <http://dx.doi.org/10.1186/s42492-021-00097-7> PMID: 34862560
- [8] Wee BF, Sivakumar S, Lim KH, Wong WK, Juwono FH. Diabetes detection based on machine learning and deep learning approaches. *Multimedia Tools Appl* 2023; 83(8): 24153-85. <http://dx.doi.org/10.1007/s11042-023-16407-5>
- [9] Abdulhadi N, Al-Mousa A. Diabetes detection using machine learning classification methods. 2021 International Conference on Information Technology (ICIT). Amman, Jordan, 14-15 July 2021, pp. 350-354 <http://dx.doi.org/10.1109/ICIT52682.2021.9491788>
- [10] Hasan MK, Alam MA, Das D, Hossain E, Hasan M. Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access* 2020; 8: 76516-31. <http://dx.doi.org/10.1109/ACCESS.2020.2989857>
- [11] Swapna G, Vinayakumar R, Soman K P. Diabetes detection using deep learning algorithms. *ICT Express* 2018; 4(4): 243-6. <http://dx.doi.org/10.1016/j.icte.2018.10.005>
- [12] Rastogi R, Bansal M. Diabetes prediction model using data mining techniques. *Measurement: Sensors* 2023; 25: 100605. <http://dx.doi.org/10.1016/j.measen.2022.100605>
- [13] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. *Procedia Comput Sci* 2018; 132: 1578-85. <http://dx.doi.org/10.1016/j.procs.2018.05.122>
- [14] Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. *Inform Med Unlock* 2018; 10: 100-7. <http://dx.doi.org/10.1016/j.imu.2017.12.006>
- [15] Zeinalnezhad M, Shishehchi S. An integrated data mining algorithms and meta-heuristic technique to predict the readmission risk of diabetic patients. *Healthcare Analyt* 2024; 5: 100292. <http://dx.doi.org/10.1016/j.health.2023.100292>
- [16] Dharmarathne G, Jayasinghe TN, Bogahawaththa M, Meddage DPP, Rathnayake U. A novel machine learning approach for diagnosing diabetes with a self-explainable interface. *Healthcare Analyt* 2024; 5: 100301. <http://dx.doi.org/10.1016/j.health.2024.100301>
- [17] Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. *Procedia Comput Sci* 2019; 165: 292-9. <http://dx.doi.org/10.1016/j.procs.2020.01.047>
- [18] Nicolucci A, Romeo L, Bernardini M, *et al.* Prediction of complications of type 2 diabetes: A machine learning approach. *Diabetes Res Clin Pract* 2022; 190: 110013. <http://dx.doi.org/10.1016/j.diabres.2022.110013> PMID: 35870573
- [19] Ganie SM, Malik MB. An ensemble machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators. *Healthcare Analyt* 2022; 2: 100092. <http://dx.doi.org/10.1016/j.health.2022.100092>
- [20] Bhat SS, Banu M, Ansari GA, Selvam V. A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms. *Healthcare Analyt* 2023; 4: 100273. <http://dx.doi.org/10.1016/j.health.2023.100273>
- [21] Singh K, Barak D. Healthcare performance in predicting type 2 diabetes using machine learning algorithms. *Driving Smart Medical Diagnosis Through AI-Powered Technologies and Applications*. IGI global 2024; pp. 130-41. <http://dx.doi.org/10.4018/979-8-3693-3679-3.ch008>
- [22] Lohani BP, Dagur A, Shukla D. Feature selection based hybrid machine learning classification model for diabetes mellitus type-II. *Artificial Intelligence, Blockchain, Computing and Security*. CRC Press 2023; pp. 96-101. <http://dx.doi.org/10.1201/9781032684994-16>
- [23] Reza MS, Amin R, Yasmin R, Kulsum W, Ruhi S. Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data. *Heliyon* 2024; 10(2): e24536. <http://dx.doi.org/10.1016/j.heliyon.2024.e24536> PMID: 38312584
- [24] Upadhyay S, Gupta YK. Development of web-based novel machine learning model using boosting techniques for early prediction of diabetes in Indian adults. 2023 12th International Conference on System Modeling and Advancement in Research Trends (SMART). Moradabad, India, 22-23 December 2023, pp. 592-602 <http://dx.doi.org/10.1109/SMART59791.2023.10428549>
- [25] Patil R, Anant Patil , Surekha Janrao , Sandip Bankar , Kamal Shah . A framework for prediction of type II diabetes through ensemble stacking model. *J Electron Electromed Eng Med Inform* 2024; 6(4): 459-66. <http://dx.doi.org/10.35882/jeeemi.v6i4.497>
- [26] P BMK, R SP, R K N, K A. Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier. *Int J Cogn Comput Eng* 2020; 1: 55-61. <http://dx.doi.org/10.1016/j.ijcce.2020.10.002>
- [27] Olisah CC, Smith L, Smith M. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Comput Methods Programs Biomed* 2022; 220: 106773. <http://dx.doi.org/10.1016/j.cmpb.2022.106773> PMID: 35429810
- [28] Ejiyi CJ, Qin Z, Amos J, *et al.* A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms. *Healthcare Analytics* 2023; 3: 100166. <http://dx.doi.org/10.1016/j.health.2023.100166>
- [29] Singh Y, Tiwari M. A comprehensive machine learning approach for early detection of diabetes on imbalanced data with missing and outlier values. *SN Computer Science* 2025; 6(3): 213. <http://dx.doi.org/10.1007/s42979-025-03751-6>
- [30] Sivaranjani S, Ananya S, Aravinth J, Karthika R. Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. *J Artificial Intell Capsule Network*. 5(2): 190. <http://dx.doi.org/10.1109/ICACCS51430.2021.9441935>
- [31] Astuti LW, Saluza I, Yulianti E, Dhamayanti D. Feature selection

- menggunakan binary wheel optimization algorithm (BWOA) pada klasifikasi penyakit diabetes. *Global Inform Sci J* 2022; 13(1) <http://dx.doi.org/10.36982/jiig.v13i1.2057>
- [32] Kishor A, Chakraborty C. Early and accurate prediction of diabetics based on FCBF feature selection and SMOTE. *Int J Syst Assur Eng Manag* 2024; 15(10): 4649-57. <http://dx.doi.org/10.1007/s13198-021-01174-z>
- [33] Saxena R, Sharma SK, Gupta M, Sampada GC. A novel approach for feature selection and classification of diabetes mellitus: Machine learning methods. *Comput Intell Neurosci* 2022; 2022(1): 1-11. <http://dx.doi.org/10.1155/2022/3820360> PMID: 35463255
- [34] Rubaiat SY, Rahman MM, Hasan MK. Important feature selection & accuracy comparisons of different machine learning models for early diabetes detection. 2018 International Conference on Innovation in Engineering and Technology (ICIET). Dhaka, Bangladesh, 27-28 December 2018, pp. 1-6, <http://dx.doi.org/10.1109/ICIET.2018.8660831>
- [35] Le TM, Vo TM, Pham TN, Dao SVT. A novel wrapper-based feature selection for early diabetes prediction enhanced with a metaheuristic. *IEEE Access* 2021; 9: 7869-84. <http://dx.doi.org/10.1109/ACCESS.2020.3047942>
- [36] Mahboob Alam T, Iqbal MA, Ali Y, *et al.* A model for early prediction of diabetes. *Inform Medicine Unlocked* 2019; 16: 100204. <http://dx.doi.org/10.1016/j.imu.2019.100204>
- [37] Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Comput Inform* 2020; 8: 90-100. <http://dx.doi.org/10.1016/j.aci.2018.12.004>
- [38] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. *Front Genet* 2018; 9: 515. <http://dx.doi.org/10.3389/fgene.2018.00515> PMID: 30459809
- [39] Ahmed N, Ahammed R, Islam MM, *et al.* Machine learning based diabetes prediction and development of smart web application. *Int J Cogn Comput Eng* 2021; 2: 229-41. <http://dx.doi.org/10.1016/j.jicce.2021.12.001>
- [40] Althobaiti T, Althobaiti S, Selim MM. An optimized diabetes mellitus detection model for improved prediction of accuracy and clinical decision-making. *Alex Eng J* 2024; 94: 311-24. <http://dx.doi.org/10.1016/j.aej.2024.03.044>
- [41] Zhang Z, Lu Y, Ye M, *et al.* A novel evolutionary ensemble prediction model using harmony search and stacking for diabetes diagnosis. *J King Saud Univ, Comp Info Sci* 2024; 36(1): 101873. <http://dx.doi.org/10.1016/j.jksuci.2023.101873>
- [42] Linkon AA, Noman IR, Islam MR, *et al.* Evaluation of feature transformation and machine learning models on early detection of diabetes mellitus. *IEEE Access* 2024; 12: 165425-40. <http://dx.doi.org/10.1109/ACCESS.2024.3488743>
- [43] Triandi B, Tanti L, Puspasari R, Elhijas MA. Optimizing diabetes prediction using machine learning with data deviation. 2024 6th International Conference on Cybernetics and Intelligent System (ICORIS). Surakarta, Indonesia, 29-30 November 2024, pp. 1-6 <http://dx.doi.org/10.1109/ICORIS63540.2024.10903969>
- [44] Hossain MM, Swarna RA, Mostafiz R, *et al.* Analysis of the performance of feature optimization techniques for the diagnosis of machine learning-based chronic kidney disease. *Machin Learning Appl* 2022; 9: 100330. <http://dx.doi.org/10.1016/j.mlwa.2022.100330>
- [45] Gollapalli M, Alansari A, Alkhorasani H, *et al.* A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM. *Comput Biol Med* 2022; 147: 105757. <http://dx.doi.org/10.1016/j.combiomed.2022.105757> PMID: 35777087
- [46] Ganie SM, Malik MB. An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators. *Healthcare Analytics* 2022; 2: 100092. <http://dx.doi.org/10.1016/j.health.2022.100092>
- [47] Jamal H, Azzimani K, Bihri H, Salma A, Charaf MEH. A comparative analysis of random forest and decision tree classifiers for predicting type 2 Diabetes using K-fold cross-validation. 6th International Symposium on Advanced Electrical and Communication Technologies (ISAECT). Alkhobar, Saudi Arabia, Alkhobar, Saudi Arabia, 2024, pp. 1-4,, pp. 1-4 <http://dx.doi.org/10.1109/ISAECT64333.2024.10799515>
- [48] Prendin F, Pavan J, Cappon G, Del Favero S, Sparacino G, Facchinetti A. The importance of interpreting machine learning models for blood glucose prediction in diabetes: An analysis using SHAP. *Sci Rep* 2023; 13(1): 16865. <http://dx.doi.org/10.1038/s41598-023-44155-x> PMID: 37803177
- [49] Kutlu M, Donmez TB, Freeman C. Machine learning interpretability in diabetes risk assessment: a SHAP analysis. *Comput Electron Medicine* 2024; 1(1): 1. <http://dx.doi.org/10.69882/adba.cem.2024075>
- [50] Revathy J, Jayanthi SK. A diabetes diagnosis model using optimized long short-term memory based on improved particle swarm optimization. *Int Res J Multidiscip Tech* 2025; 7(1): 47-70. <http://dx.doi.org/10.54392/irjmt2514>
- [51] Pang K. A comparative study of explainable machine learning models with Shapley values for diabetes prediction. *Healthcare Analyt* 2025; 7: 100390. <http://dx.doi.org/10.1016/j.health.2025.100390>
- [52] Choubey D K, Tripathi S, Kumar P, Shukla V, Dhandhanian V K. Classification of diabetes by kernel based SVM with PSO. *Recent Adv Comput Sci Commun* 2021; 7: 1242-55. <http://dx.doi.org/10.2174/2213275912666190716094836>
- [53] Ulutas H, Günay RB, Sahin ME. Detecting diabetes in an ensemble model using a unique PSO-GWO hybrid approach to hyperparameter optimization. *Neural Comput Appl* 2024; 36(29): 18313-41. <http://dx.doi.org/10.1007/s00521-024-10160-y>
- [54] Bukhari MM, Alkhamees BF, Hussain S, Gumaei A, Assiri A, Ullah SS. An improved artificial neural network model for effective diabetes prediction. *Complexity* 2021; 2021(1): 5525271. <http://dx.doi.org/10.1155/2021/5525271>
- [55] Kurt B, Gürlek B, Keskin S, *et al.* Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques. *Med Biol Eng Comput* 2023; 61(7): 1649-60. <http://dx.doi.org/10.1007/s11517-023-02800-7> PMID: 36848010
- [56] Gündoğdu S. Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique. *Multimedia Tools Appl* 2023; 82(22): 34163-81. <http://dx.doi.org/10.1007/s11042-023-15165-8> PMID: 37362660
- [57] Palimkar P, Shaw RN, Ghosh A. Machine learning technique to prognosis diabetes disease: Random forest classifier approach. *Advanced computing and intelligent technologies: Proceedings of ICACIT 2021*. Singapore: Springer Singapore 2021; pp. 219-44. http://dx.doi.org/10.1007/978-981-16-2164-2_19
- [58] Abdollahi J, Nouri-Moghaddam B. Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction. *Iran J Computer Sci* 2022; 5(3): 205-20. <http://dx.doi.org/10.1007/s42044-022-00100-1>
- [59] Al-Tawil M, Mahafzah BA, Al Tawil A, Aljarah I. Bio-inspired machine learning approach to Type 2 diabetes detection. *Symmetry* 2023; 15(3): 764. <http://dx.doi.org/10.3390/sym15030764>
- [60] Saxena S, Mohapatra D, Padhee S, Sahoo GK. Machine learning algorithms for diabetes detection: A comparative evaluation of performance of algorithms. *Evol Intell* 2023; 16(2): 587-603. <http://dx.doi.org/10.1007/s12065-021-00685-9>