Journal of Intelligent Systems and Internet of Things
Vol. 16, No. 02, PP. 13-25, 2025
DOI: https://doi.org/10.54216/JISIoT.160202

Greylag Goose Optimization for Diabetes Prediction: Feature Selection Meets Advanced Machine Learning

Gomaa Mohamed Ismail[1,*], **El-Sayed M. El-kenawy**[2,3,4], **Shady Y. El-Mashad**[1]

[1]Department of Computer, Systems Engineering, Faculty of Engineering at Shoubra, Benha University, Egypt
[2]Department of Communications and Electronics, Delta Higher Institute of Engineering and Technology, Mansoura, 35111, Egypt
[3]Applied Science Research Center. Applied Science Private University, Amman, Jordan
[4]Jadara University Research Center, Jadara University, Jordan

Emails: gomaa.esmail19@feng.bu.edu.eg; skenawy@ieee.org; Shady.elmashad@feng.bu.edu.eg

Abstract

Diabetes mellitus remains a global health concern, necessitating both accurate and effective diagnostic methodologies. This condition presents a significant challenge due to the high dimensionality of clinical datasets and the inherent complexity of diabetes classification. To address this problem, this study integrates feature selection and machine learning architectures to enhance diabetes prediction accuracy. A novel framework based on the Binary Greylag Goose Optimization (bGGO) algorithm is proposed to optimize feature selection, thereby improving classification performance. A comprehensive evaluation uses multiple classifiers, including Decision Trees, k-nearest Neighbors, Support Vector Machines, Random Forests, and Multilayer Perceptron (MLP). The experimental results demonstrate that bGGO significantly enhances feature selection quality, improving classification metrics, particularly for MLP, which achieves the highest classification accuracy of 95.98%. These findings underscore the efficacy of combining metaheuristic optimization with machine learning for diabetes diagnosis, offering a scalable and interpretable approach for real-world healthcare applications. The proposed methodology contributes to more precise risk estimation and the development of individualized intervention strategies, facilitating early diagnosis and effective disease management.

Received: January 01, 2025 Revised: February 05, 2025 Accepted: March 01, 2025

Keywords: Diabetes; Meta-heuristic Optimization; Feature Selection; Machine Learning Architectures; Greylag Goose Optimization; Multilayer Perceptron

1 Introduction

Diabetes mellitus is a chronic metabolic disorder resulting from elevated blood glucose levels caused by insufficient insulin production in the body or ineffective utilization of insulin. This is a significant global health problem affecting millions of patients worldwide, with devastating ramifications for their health, including cardiovascular diseases, neuropathy, nephropathy, and retinopathy [1, 2]. Diabetes has emerged as one of the most serious non-communicable diseases, exacerbated by modern lifestyles. The combination of sedentary habits, unhealthy eating patterns, genetic predispositions, and environmental conditions has contributed significantly to the prevalence of this disease. Given the alarming increase in diabetes cases, there is a pressing

need for improved diagnostic methods to enable timely detection and advanced computational techniques for early identification and classification.

Since diabetes is a heterogeneous disease with multiple subtypes, such as Type 1 diabetes (T1D), Type 2 diabetes (T2D), and gestational diabetes mellitus (GDM), its classification remains a complex challenge. Each subtype has unique pathophysiological characteristics, necessitating precise diagnostic approaches to facilitate effective disease management and intervention [3]. Accurate diabetes classification is crucial in tailoring treatment plans specific to each patient's needs and mitigating long-term complications. This study aims to extensively investigate diabetes classification techniques using computational intelligence and machine learning methods to improve prediction accuracy and enable automated diagnosis [4, 5].

Machine learning has emerged as a powerful tool in medical diagnostics, allowing the ability to analyze large biometric datasets and uncover hidden patterns contributing to disease progression. The incorporation of data-driven methodologies in diabetes classification enables the development of predictive models that assess an individual's likelihood of developing the disease based on clinical parameters, lifestyle attributes, and genetic factors [6]. These models can process multidimensional data, extract meaningful insights, and enhance diagnostic precision. The primary objective of this study is to design robust diagnostic algorithms that utilize machine learning techniques to predict diabetes susceptibility and optimize patient care. By implementing various classification techniques, we aim to develop a comprehensive framework that enhances clinical decision-making and assists healthcare professionals in identifying high-risk individuals [7].

One of the fundamental aspects of diabetes classification is the selection and application of machine learning algorithms, ranging from traditional statistical classifiers to contemporary deep learning models. Various conventional methods, such as logistic regression, decision trees, and support vector machines (SVM), have been widely used for diabetes prediction. However, recent advancements in deep learning architectures, including artificial neural networks (ANNs) and convolutional neural networks (CNNs), have demonstrated superior performance in capturing complex relationships within clinical data [8]. These advanced methodologies enable automated feature extraction, reducing the need for extensive manual preprocessing while improving classification accuracy. Furthermore, machine learning-driven classifiers facilitate the identification of intricate patterns and correlations within patient datasets, promoting early disease diagnosis and intervention [9].

Optimization techniques are crucial in refining classification models and enhancing their predictive performance. Integrating optimization methods within machine learning frameworks ensures fine-tuning of hyperparameters, selection of optimal feature subsets, and improved model generalizability. Hyperparameter tuning is one of the most widely utilized optimization strategies in machine learning, systematically adjusting learning rates, activation functions, and network architectures to achieve optimal performance [10]. Additionally, cross-validation techniques are employed to assess the reliability and stability of classification models, ensuring robust performance across diverse datasets. By implementing rigorous evaluation methods, we enhance the credibility and applicability of machine learning models in real-world healthcare scenarios [11, 12].

Feature selection is another pivotal component of diabetes classification, as it directly influences model efficiency, interpretability, and computational complexity. Feature selection techniques aim to identify the most relevant attributes contributing to accurate disease prediction while eliminating redundant and non-informative variables [13]. This problem is categorized as an NP-hard problem. It is typically solved using various approaches, such as filter-based techniques relying on statistical correlation, wrapper-based methods utilizing predictive modeling for feature ranking, and embedded approaches integrating feature selection within model training. By reducing the dimensionality of input datasets, these techniques enhance classification performance, minimize overfitting, and improve the generalization capabilities of predictive models [14].

Among the different machine learning architectures, the multilayer perceptron (MLP) stands out as a particularly effective model for diabetes classification. MLP is an artificial neural network comprising multiple layers of interconnected neurons, capable of capturing nonlinear relationships and intricate data patterns. Due to its ability to learn complex mappings between input features and target variables, MLP is well-suited for medical classification tasks [15]. Unlike traditional classifiers, MLP excels in handling high-dimensional data and extracting deep feature representations that contribute to improved diagnostic accuracy. The iterative training process of MLP, involving backpropagation and weight updates, enables continuous refinement of

model parameters, ultimately enhancing classification performance. By leveraging MLP's adaptive learning capabilities, we aim to develop a highly accurate and reliable diabetes prediction model [16–18].

This study seeks to integrate optimization techniques, feature selection methods, and MLP architectures to develop a robust diabetes classification framework. We combine these approaches to enhance diagnostic precision, improve model interpretability, and streamline computational efficiency. The proposed methodology aims to provide a scalable and adaptable solution for diabetes classification, ultimately contributing to better disease management and personalized treatment strategies. Through rigorous experimentation and evaluation, we endeavor to establish a state-of-the-art diagnostic system that equips healthcare professionals with reliable and actionable insights for diabetes diagnosis and intervention [19–21].

The key contributions of this study are as follows:

- Propose an optimized diabetes diagnosis framework integrating feature selection and advanced machine learning architectures.
- Utilize Greylag Goose Optimization (GGO) to refine feature selection and improve classification performance.
- Investigate the multilayer perceptron's (MLP) impact on effective diabetes classification by capturing complex feature interactions.
- Compare and evaluate multiple machine learning models, including Decision Trees, k-nearest Neighbors, Support Vector classifiers, Random Forests, and MLP, to determine the most effective classification approach.
- Employ rigorous evaluation metrics such as Positive Predictive Value (PPV), Negative Predictive Value (NPV), F1 Score, Accuracy, True Positive Rate (TPR), and True Negative Rate (TNR) to assess model performance.
- Propose a feature selection methodology combining the sigmoid function with binary Greylag Goose Optimization (bGGO) to enhance interpretability and model efficiency.
- Validate the proposed framework on real-world diabetes datasets, demonstrating significant improvements in classification accuracy and model robustness.

The remainder of this paper is structured as follows: Section 2 presents a comprehensive review of related works, highlighting previous approaches and methodologies in diabetes diagnosis and classification. Section 3 describes the dataset, pre-processing techniques, and machine learning methodologies, including feature selection and optimization techniques. Section 4 discusses the experimental results, comparing various classifiers and evaluating their performance based on multiple metrics. Finally, Section 5 concludes the study with a summary of findings, key contributions, and potential directions for future research.

2 Related Works

The landscape of diabetes diagnostic research is enriched with various methodologies aimed at improving predictive interpretation through refined techniques and diverse optimization strategies. Researchers have recently explored multiple computational and machine learning techniques to enhance classification accuracy by balancing key performance metrics such as precision, recall, and overall predictive reliability [22]. Beyond these efforts, optimization techniques have significantly improved model performance, enhanced generalization capabilities, and ensured robustness when dealing with real-world medical datasets. Consequently, these approaches have guided research toward exploring novel algorithmic strategies to maintain efficiency and interpretability in the diagnostic process.

Optimization processes encompass various techniques, from traditional brute-force search methods to sophisticated nature-inspired algorithms. These methodologies include parametric and non-parametric procedures

for model selection and hyperparameter determination, often leading to optimal solutions for one performance objective but suboptimal for another. One of the earliest and most straightforward techniques is grid search, which systematically evaluates predefined sets of hyperparameter combinations to identify the optimal configuration for a given model [23]. Despite its simplicity, this approach suffers from computational inefficiency, particularly in high-dimensional hyperparameter spaces. Researchers have increasingly relied on metaheuristic optimization techniques to address these limitations, such as Genetic Algorithms (GA) and Particle Swarm Optimization (PSO). These methods provide more efficient search strategies by simulating biological evolution and swarm intelligence principles, respectively [24]. These algorithms converge towards optimal solutions by leveraging adaptive exploration and exploitation mechanisms, making them highly effective for fine-tuning machine learning models in complex classification tasks, including diabetes diagnosis.

The selection of an appropriate algorithm and the careful configuration of its hyperparameters are crucial steps in optimizing machine learning models for medical diagnostics [25]. Inadequate parameter selection can lead to overfitting or underfitting, reducing the model's diagnostic efficacy. Unlike grid search, more advanced methods such as Bayesian Optimization and Evolutionary Computation have been explored to enhance the efficiency of hyperparameter tuning. These techniques prioritize informative sampling strategies, thereby reducing computational overhead while maintaining high accuracy [26]. Such optimization frameworks have been widely applied in diabetes classification tasks, where precise and reliable predictions are critical for medical decision-making.

Feature selection methods also play a pivotal role in enhancing the predictive performance of machine learning models, particularly in medical applications where datasets often contain high-dimensional and redundant variables [27]. Feature selection techniques are broadly categorized into three groups: filter methods, wrapper methods, and embedded methods. Filter methods, such as correlation analysis and mutual information, evaluate the importance of individual features based on statistical criteria, enabling computationally efficient feature ranking. On the other hand, Wrapper methods iteratively assess different feature subsets by training and validating the model on each combination, thereby identifying the most informative feature set. While wrapper methods generally achieve superior predictive performance, they are computationally intensive. Embedded methods, such as LASSO (Least Absolute Shrinkage and Selection Operator) and tree-based feature importance rankings, integrate feature selection directly into the model training process, optimizing feature selection and classification simultaneously [28]. These methods have been widely adopted in diabetes prediction research, where reducing dimensionality enhances model interpretability and mitigates overfitting risks.

With the rise of deep learning techniques, neural network architectures have gained significant traction in diabetes classification studies. Among these, multilayer perceptrons (MLPs) have demonstrated remarkable effectiveness in capturing complex patterns within medical data [29]. MLPs consist of multiple hidden layers that facilitate the extraction of intricate relationships between input features, making them particularly suited for heterogeneous and multidimensional datasets, such as those encountered in medical diagnostics. Additionally, convolutional neural networks (CNNs) have been increasingly utilized in diabetes classification tasks, particularly for analyzing medical imaging data, including retinal scans and other diagnostic images. The ability of CNNs to automatically learn hierarchical feature representations has made them indispensable in medical image-based disease detection, providing a powerful tool for early diabetes diagnosis.

Moreover, hybrid optimization approaches have recently gained attention in the literature. In these techniques, multip combines multiple optimization techniques and their complementary strengths. For instance, hybrid models incorporating swarm intelligence algorithms and gradient-based optimization have been explored to improve classification accuracy and efficiency. These hybrid techniques enable dynamic adjustments to hyperparameter settings while maintaining robust convergence properties, ensuring high diagnostic reliability.

The related work underscores that diabetes classification remains a challenging problem, necessitating an interdisciplinary approach that integrates optimization techniques, feature selection strategies, and advanced machine learning architectures. The primary objective is to develop models that are not only highly accurate but also interpretable and computationally efficient, ensuring their practical applicability in real-world healthcare settings. Building on insights from previous studies, future research will focus on refining model architectures and optimization strategies to further enhance the predictive capabilities of diabetes diagnosis models, ultimately contributing to improved patient care and disease management.

3 Material and Methods

3.1 data collection

The dataset [30] is a collection of medical data primarily aimed at predicting whether a patient has diabetes based on diagnostic measurements. It originates from the National Institute of Diabetes and Digestive and Kidney Diseases. It's essential to note that the dataset focuses on female patients at least 21 years old and of Pima Indian heritage. Each instance in the dataset comprises several features that are potential indicators of diabetes. These features include:

- Pregnancies: The number of times a patient has been pregnant. This can be a significant factor in assessing diabetes risk, as pregnancy can affect insulin sensitivity and glucose metabolism.
- Glucose: The plasma glucose concentration measured two hours after an oral glucose tolerance test. Elevated glucose levels are a hallmark of diabetes and are crucial in diagnosing the condition.
- Blood Pressure: The diastolic blood pressure, measured in millimeters of mercury (mm Hg). High blood pressure is often associated with diabetes and can be both a cause and a consequence of the disease.
- Skin Thickness: The thickness of the triceps skin fold, measured in millimeters. While not a direct measure of diabetes, abnormal skin thickness can indicate underlying metabolic issues.
- Insulin: The serum insulin level measured two hours after glucose intake, expressed in micro International Units per milliliter (mu U/ml). Insulin resistance, characterized by elevated insulin levels, is a hallmark of type 2 diabetes.
- BMI: The body mass index, calculated as weight in kilograms divided by the square of height in meters. Obesity and higher BMI are significant risk factors for diabetes due to their association with insulin resistance.
- Diabetes Pedigree Function: A function that scores the likelihood of diabetes based on family history. Genetic predisposition plays a crucial role in the development of diabetes, and this feature captures that aspect.
- Age: The age of the patient in years. Age is a critical factor in diabetes risk assessment, as the prevalence of diabetes increases with age.
- Outcome: This is the target variable, indicating whether the patient has diabetes (1) or not (0).

These features can provide insights into the underlying patterns and risk factors associated with diabetes among the Pima Indian female population aged 21 and above. However, it's important to interpret the findings considering the dataset's limitations, such as its specific demographic focus and potential biases in the data collection process.

3.2 Data Pre-Processing

This is an essential step that must be completed before the whole analysis. Its main goal is to verify whether the dataset is good quality and reliable. At this point, the process includes cleaning and upgrading the nascent data to meet the analytical requirement. For instance, the diabetes classification data from the National Institute of Diabetes, Digestive and Kidney Diseases should undergo a series of preprocessing before being optimized for specific tasks.

1. Handling Missing Values: An effective way is to consider the problem of non-combative integrity of the real-world data, which is a source of deterioration in algorithms' effectiveness. In this case, the dataset may have a missing value in the feature set, which can be Glucose, Pressure, Skin thickness, Insulin, or BMI.

2. Handling Outliers: Outliers are soapy values that could increasingly deform data distribution and distort the model performance. To analyze datasets with disparities in data point values, techniques like z-score normalization and the interquartile range (IQR) method can be used to detect and overcome outliers.

3. Feature Scaling: Since the models are affected by dissimilar dimensions and units in the dataset, it is necessary to carry out feature scaling to ensure that all features contributing to the model have the same importance. Several scaling techniques, such as standardization (z-score normalization) and min-max scaling, are usually used, which change the features to a common scale while maintaining their relative differences.

4. Splitting the Dataset: In most cases, the initial stage before using machine learning algorithms appears to consist of dividing the dataset into parts called training and testing sets to see the model performance. The training set is learning for the model, whereas the test set is used to test its generalization ability.

The initial raw dataset is optimized for analysis so that the various basic data-cleaning techniques are implemented to extract the primary data efficiently.

3.3 Descriptive Analysis of the Dataset

The diabetes dataset is analyzed in depth using descriptive analysis techniques to decipher the variables' features and distribution. This analysis will be the starting point for the in-depth investigation of the data and inform our modeling choices. The inset of demographic and clinical data, consisting of several variables, helps identify different risk factors for the development of diabetes.

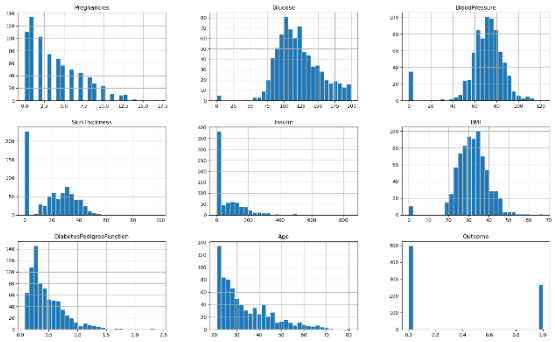

Figure 1: Histogram of the dataset

The accountancy of descriptive statistics includes an elaborate computation of summary statistics for each variable, which gives us the mean, median, standard deviation, minimum, and maximum values. As a result, researchers can identify any trends and pick outliers. Figure 1 displays all the variables in a histogram, a graphical presentation of the frequency distribution. In addition, correlation analysis is implemented to assess the level and nature of relations between pairs of variables. These analytical techniques provide a means to detect multicollinearity problems in which the relationship between variables is very tight, which will probably affect the stability and interpretability of modeling methods. The heatmap of the dataset is shown in Figure 2.

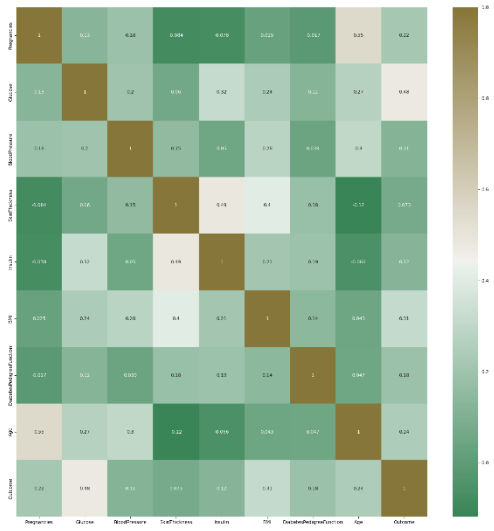

Figure 2: Heatmap of the dataset

Moreover, the analyses included univariate and bivariate studies and showed stratification by the outcome variable to see if the variable distribution differs between diabetic and non-diabetic individuals. Through this comparative analysis, the researchers can point out aspects of informative value in the feature selection stage and might take part in developing appropriate classification algorithms. Figure 3 explores the percentage of people who have diabetes and those who do not use the data to visualize the distribution of outcome variables over the dataset.

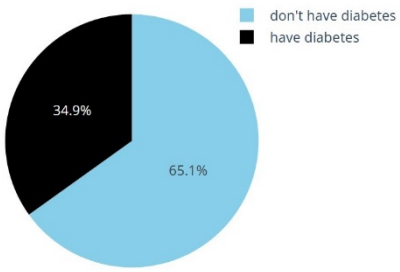

Figure 3: Percentage of people wPercentageabetes vs. those who don't have diabetes

Essentially, this descriptive analysis provides the foundation for examining the diversity and constructs in the diabetes dataset. By examining the relationships between different variables, researchers can obtain vital information regarding the structure of diabetes diagnostics and treatments, which will subsequently be used for inferential and provocative analyses of the whole system.

3.4 Multilayer Perceptron

The multilayered perceptron (MLP) is an efficient and flexible neural network used for multiple tasks in diagnosing diabetes. Since a multilayered network of neurons constitutes the layers of MLP, it quickly learns intricate data patterns. This iterative process mainly revolves around refining the weight and bias feature through training, which leads to the accurate prediction of the presence or absence of diabetes among the clinical variables. Advances in weight regularization methods and activation functions have made MLP a ready solution for diabetes diagnosis as early as possible and encouraged intervention and personalized treatment methods. Figure 4 shows the different layers in a multilayer perceptron: the input, the hidden and the output layers, and the neuron connections.

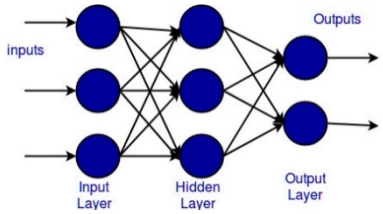

Figure 4: Multilayer Neural Networks Structure

3.5 Feature Selection

Feature selection using the sigmoid function and bGGO (binary Greylag Goose Optimization) is a new method that enables the most relevant features for diabetes diagnosis to be found. The sigmoid function, a nonlinear activation function typically used in neural networks, is deliberately utilized to compute the relevance of the features to the target variable, which is, in this case, diabetes outcome. Large weights are attributed to those features that are believed to be more effective in predicting diabetes risk, whereas small weights correspond with those features that are less likely to be helpful. Together with the sigmoid function, bGGO offers a hybrid optimization technique for solving complex problems that are based on the collective behavior of Greylag geese. bGGO is a version of a genetic algorithm that does not mimic the natural flocking behavior of geese. Instead, it iteratively explores different feature spaces to identify the best subset of features that results in the highest classification performance of the model. It is due to the use of an evolutionary approach that the features that contribute the most to the model's accuracy are not only chosen, but also the model is stable and robust across multiple datasets and various experiments. The sigmoid function and the bGGO, taken together, form a solid platform for feature selection in a diabetes classifier. This approach reduces the number of features while minimizing the problem of overfitting. It is very promising because it improves the interpretability and generalization ability of diabetes classification models, thus providing more precise risk assessment and individual intervention strategies for diabetes patients.

$$x_d^{(t+1)} = \begin{cases} 1 & \text{if Sigmoid}(m) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Sigmoid}(m) = \frac{1}{1+e^{-10(m-0.5)}}$$

3.6 Binary Greylag Goose Optimization Algorithm

Our Binary Greylag Goose Optimization Algorithm (bGGO) is a special metaheuristic optimization algorithm dedicated to resolving complicated optimization problems, such as feature selection in diabetes classification. Algorithm 1, derived from population-based algorithms inspired by the flocking behavior of Greylag geese, explores solution space while identifying the optimal feature subset for classification.

Algorithm 1 bGGO Algorithm

```
Initialize GGO population, objective function, and GGO parameters
Convert solution to binary [0 or 1]
Calculate the objective function for each agent and get the best agent position
Update Solutions in exploration group and exploitation group
while t ≤ t_max do
    for i = 1 to n_1 do
        if t%2 == 0 then
            if r_3 < 0.5 then
                if |A| < 1 then
                    Update position of current search agent in exploration group
                else
                    Update position of current search agent based on three random search agents
                end if
            else
                Update position of current search agent
            end if
        else
            Update individual positions
        end if
    end for
    for i = 1 to n_2 do
        if t%2 == 0 then
            Update position of current search agent in exploitation group
        else
            Update position of current search agent
        end if
    end for
    Convert updated solution to binary
    Calculate objective function
    Update parameters
    Adjust beyond the search space solutions
    Update Solutions in exploration group and exploitation group
end while
return best agent
```

Then, the system is defined by initializing the population of GGO agents, objective function, and parameters corresponding to the bGGO algorithm. Then, each solution in the population is converted to binary format depicting the presence of traits or lack thereof in classification. The next step is to calculate an objective function for all agents, after which the best agent position is found according to its performance. Exploring and exploiting the algorithm are iteratively performed until a termination condition (represented by the maximum number of iterations (tmax)) is reached. In each iteration, the algorithm determines the solutions of the varying exploration and exploitation groups with the help of condition statements set according to the iteration index (t) and random variables (r3). As these updates are designed to balance exploration and exploitation, the algorithm can utilize its feature space efficiently and eventually converge to an optimum feature subset.

4 Experimental Results

The accuracy of feature selection for diabetes classification was evaluated through various metrics that looked into its effectiveness, as shown in Table. 1 illustrates that the criteria incorporate Best Fitness, Worst Fitness, Mean Error, Mean Fitness, Fitness Size (range), and Standard deviation.

Table 1: Criteria for evaluating feature selection results.

Metric	Formula
Best Fitness	$\min_{i=1}^{M} S_i^*$
Worst Fitness	$\max_{i=1}^{M} S_i^*$
Average Error	$\frac{1}{M}\sum_{j=1}^{M}\frac{1}{N}\sum_{i=1}^{N}\text{mse}(\hat{V}_i - V_i)$
Average Fitness	$\frac{1}{M}\sum_{i=1}^{M} S_i^*$
Average Fitness Size	$\frac{1}{M}\sum_{i=1}^{M}\text{size}(S_i^*)$
Standard Deviation	$\sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(S_i^* - \text{Mean})^2}$

Table. 2 summarizes the evaluation results before feature selection for different classifiers. Metrics of Positive Predictive Value (PPV), Negative Predictive Value (NPV), F1 Score, Accuracy, and other parameters such as True Positive Rate (TRP) and True Negative Rate (TNP).

Table 2: Models' Evaluation Results

	PPV	NPV	FScore	Accuracy	TRP	TNP
DT	0.79337	0.76790	0.61569	0.71756	0.74408	0.67549
KNN	0.76305	0.80045	0.66256	0.73089	0.84191	0.55480
SVC	0.79862	0.87658	0.87540	0.81756	0.92343	0.57204
RF	0.87799	0.88702	0.80542	0.85089	0.89626	0.77894
MLP	0.94665	0.87967	0.94665	0.92218	0.94665	0.87967

The outcome of feature reduction via different algorithms is given in Table. 3, which includes the methods of bGGO, bBER, bDTO, bPSO, bWAO, bGWO, and bFA. Among other metrics, average error, average select size, average fitness, best fitness, worst fitness, and standard deviation fitness are also reported.

Table 3: Feature Selection Results

	bGGO	bBER	bDTO	bPSO	bWAO	bGWO	bFA
Average error	0.565	0.602	0.616	0.636	0.636	0.622	0.634
Average Select size	0.538	0.738	0.680	0.738	0.901	0.660	0.772
Average Fitness	0.648	0.664	0.676	0.663	0.670	0.670	0.715
Best Fitness	0.550	0.585	0.579	0.643	0.635	0.648	0.633
Worst Fitness	0.648	0.651	0.694	0.711	0.711	0.724	0.731
Standard deviation Fitness	0.470	0.475	0.477	0.474	0.477	0.476	0.511

Table. 4 presents the performance metrics of classification models after feature selection. For instance, Decision Trees achieved an Accuracy of 0.76772, k-Nearest Neighbors reached an Accuracy of 0.82826, Support Vector Classifier attained an Accuracy of 0.88941, Random Forests demonstrated an Accuracy of 0.91501, and Multilayer Perceptron showcased an Accuracy of 0.95985.

Table 4: Models Evaluation Results After Feature Selection

	PPV	NPV	FScore	Accuracy	TRP	TNP
DT	0.84238	0.87967	0.81678	0.76772	0.94665	0.69389
KNN	0.80679	0.89470	0.88502	0.82826	0.98096	0.58301
SVC	0.89470	0.87967	0.91991	0.88941	0.94665	0.79946
RF	0.91862	0.94703	0.92645	0.91501	0.96779	0.81756
MLP	0.94675	0.91259	0.96646	0.95985	0.96675	0.92797

The experimental results demonstrate the effectiveness of feature selection in enhancing the performance of various classification models for diabetes prediction. As shown in Table 1, the evaluation criteria for feature

selection quality include best fitness, worst fitness, mean error, mean fitness, and standard deviation. These metrics comprehensively assess the optimization algorithms' ability to refine feature subsets while maintaining high classification accuracy.

Before feature selection, the models exhibited varying levels of predictive performance, as presented in Table 2. The Multilayer Perceptron (MLP) achieved the highest accuracy (0.92218), followed by the Random Forest (0.85089) and Support Vector Classifier (SVC) (0.81756). In contrast, Decision Trees (DT) and k-nearest Neighbors (KNN) exhibited lower accuracy levels, with values of 0.71756 and 0.73089, respectively. The True Positive Rate (TPR) and True Negative Rate (TNR) values indicate that specific models, such as MLP and RF, effectively classified diabetic and non-diabetic cases. However, some classifiers, including KNN, showed a lower TNR (0.55480), suggesting a higher misclassification rate for negative cases.

Significant improvements were observed in classification accuracy and predictive metrics following feature selection, as outlined in Table 4. The MLP classifier continued to achieve the highest accuracy (0.95985), followed by RF (0.91501), SVC (0.88941), and KNN (0.82826). Notably, the Decision Tree model also demonstrated an improved accuracy of 0.76772, underscoring the positive impact of feature selection in refining model performance. Additionally, the observed increases in Positive Predictive Value (PPV) and Negative Predictive Value (NPV) further confirm the enhanced reliability of classification models after feature selection.

Table 3 presents the feature selection results across multiple binary optimization techniques, including bGGO, bBER, bDTO, bPSO, bWAO, bGWO, and bFA. Among these, bGGO achieved the lowest average error (0.565) and best fitness (0.550), demonstrating its superior capability in optimizing feature selection. Conversely, bFA exhibited the highest average fitness (0.715) but slightly higher average error (0.634). The standard deviation values remained relatively stable across the optimization algorithms, indicating consistency and robustness in the optimization process.

Overall, these results highlight the significance of feature selection in improving classification performance. The optimization techniques effectively reduced feature redundancy while preserving critical information, improving accuracy across all classification models. These findings suggest that feature selection mitigates the curse of dimensionality, enhances computational efficiency, and improves model interpretability. Future work could explore hybrid optimization approaches that integrate multiple optimization strategies to refine feature selection further and enhance classification performance.

5 conclusion

In conclusion, we can state that the employment of feature selection tools has enhanced the precision levels and reliability of diabetes diagnosis models. In this regard, several experiments have been performed, and the significance of selected features with the help of learning algorithms like bGGO and bBER has been demonstrated. In the next phase, we observed that classification models exhibited significant improvements in the core performance indicators such as PPV, NPV, F1 Score, and Accuracy. Such outcomes emphasize the role of feature selection in optimizing the model performance, which can be achieved by fighting the curse of dimensionality and the enhancement of interpretability of the diabetes classification models. Moving ahead, future research can delve into the more advanced feature selection algorithms to check their resilience with diverse patient populations. As a result, more proper risk assessment and personalized intervention strategies can be served to individuals with diabetes.

References

[1] S. Bhandari, S. Pathak, and S. A. Jain. A literature review of early-stage diabetic retinopathy detection using deep learning and evolutionary computing techniques. *Archives of Computational Methods in Engineering*, 30(2):799–810, 2023.

[2] M. Gollapalli, A. Alansari, H. Alkhorasani, M. Alsubaii, R. Sakloua, R. Alzahrani, M. Al-Hariri, M. Alfares, D. AlKhafaji, R. Al Argan, and W. Albaker. A novel stacking ensemble for detecting three types of diabetes mellitus using a saudi arabian dataset: Pre-diabetes, t1dm, and t2dm. *Computers in Biology and Medicine*, 147:105757, 2022.

[3] M. Allam and M. Nandhini. Optimal feature selection using binary teaching learning based optimization algorithm. *Journal of King Saud University - Computer and Information Sciences*, 34(2):329–341, 2022.

[4] F. Barbetti, N. Rapini, R. Schiaffini, C. Bizzarri, and S. Cianfarani. The application of precision medicine in monogenic diabetes. *Expert Review of Endocrinology Metabolism*, 17(2):111–129, 2022.

[5] F. G. Preston, Y. Meng, J. Burgess, M. Ferdousi, S. Azmi, I. N. Petropoulos, S. Kaye, R. A. Malik, Y. Zheng, and U. Alam. Artificial intelligence utilising corneal confocal microscopy for the diagnosis of peripheral neuropathy in diabetes mellitus and prediabetes. *Diabetologia*, 65(3):457–466, 2022.

[6] B. Kurt, B. Gürlek, S. Keskin, S.Özdemir,Ö. Karadeniz, İ. B. Kırkbir, T. Kurt, S.Ünsal, C. Kart, N. Baki, and K. Turhan. Prediction of gestational diabetes using deep learning and bayesian optimization and traditional machine learning techniques. *Medical & Biological Engineering & Computing*, 61(7):1649–1660, 2023.

[7] L. Ismail, H. Materwala, M. Tayefi, P. Ngo, and A. P. Karduck. Type 2 diabetes with artificial intelligence machine learning: Methods and evaluation. *Archives of Computational Methods in Engineering*, 29(1):313–333, 2022.

[8] H. A. Aliyu, I. O. Muritala, H. Bello-Salau, S. Mohammed, A. J. Onumanyi, and O.-O. Ajayi. Optimizing machine learning algorithms for diabetes data: A metaheuristic approach to balancing and tuning classifiers parameters. *Franklin Open*, 8:100153, 2024.

[9] U. Ahmed, G. F. Issa, M. A. Khan, S. Aftab, M. F. Khan, R. A. T. Said, T. M. Ghazal, and M. Ahmad. Prediction of diabetes empowered with fused machine learning. *IEEE Access*, 10:8529–8538, 2022.

[10] H. Shao, X. Liu, D. Zong, and Q. Song. Optimization of diabetes prediction methods based on combinatorial balancing algorithm. *Nutrition & Diabetes*, 14(1):1–13, 2024.

[11] C. Burchill L. C. A. Rosella, D. G. Manuel and T. A. Stukel. A population-based risk algorithm for the development of diabetes: development and validation of the diabetes population risk tool (dport). *J. Epidemiol. Community Health*, (7):613–620, 2011.

[12] J. A. L. Marques, F. N. B. Gois, J. P. do V. Madeiro, T. Li, and S. J. Fong. Artificial neural network-based approaches for computer-aided disease diagnosis and treatment. In A. K. Bhoi, V. H. C. de Albuquerque, P. N. Srinivasu, and G. Marques, editors, *Cognitive and Soft Computing Techniques for the Analysis of Healthcare Data*, pages 79–99. Academic Press, 2022.

[13] A. Dutta, M. K. Hasan, M. Ahmad, M. A. Awal, M. A. Islam, M. Masud, and H. Meshref. Early prediction of diabetes using an ensemble of machine learning models. *International Journal of Environmental Research and Public Health*, 19(19):Article 19, 2022.

[14] Bilal, A., Sun, G., Mazhar, S., Imran, A., & Latif, J. (2022). A Transfer Learning and U-Net-based automatic detection of diabetic retinopathy from fundus images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 10(6), 663-674.

[15] C. C. Olisah, L. Smith, and M. Smith. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220:106773, 2022.

[16] M. S. Ali, M. K. Islam, A. A. Das, D. U. S. Duranta, Mst. F. Haque, and M. H. Rahman. A novel approach for best parameters selection and feature engineering to analyze and detect diabetes: Machine learning insights. *BioMed Research International*, 2023(1):8583210, 2023.

[17] M. O'Neill T. Piggott J. D. Morgenstern, E. Buajitti and V. Goel. Predicting population health with machine learning: a scoping review. *BMJ Open*, 10(10):e037860, 2020.

[18] W. Xu, Z. Zhang, K. Hu, P. Fang, R. Li, D. Kong, M. Xuan, Y. Yue, D. She, and Y. Xue. Identifying metabolic syndrome easily and cost effectively using non-invasive methods with machine learning models. *Diabetes, Metabolic Syndrome and Obesity*, 16:2141–2151, 2023.

[19] S S S, J Surendiran, N Yuvaraj, M Ramkumar, CN Ravi, and RG Vidhya. Classification of diabetes using multilayer perceptron. In *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, pages 1–5, 2022.

[20] I. D. Dinov. Data science and predictive analytics: Biomedical and health applications using r. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2023.

[21] TM Le, TM Vo, TN Pham, and SVT Dao. A novel wrapper–based feature selection for early diabetes prediction enhanced with a metaheuristic. *IEEE Access*, 9:7869–7884, 2021.

[22] F Khademi, M Rabbani, H Motameni, and E Akbari. A weighted ensemble classifier based on woa for classification of diabetes. *Neural Computing and Applications*, 34(2):1613–1621, 2022.

[23] C Mallika and S Selvamuthukumaran. A hybrid crow search and grey wolf optimization technique for enhanced medical data classification in diabetes diagnosis system. *International Journal of Computational Intelligence Systems*, 14(1):157, 2021.

[24] R. Ahuja, P. Dixit, A. Banga, and S. C. Sharma. Classification algorithms for predicting diabetes mellitus: A comparative analysis. In M. S. Husain, M. H. B. M. Adnan, M. Z. Khan, S. Shukla, and F. U. Khan, editors, *Pervasive Healthcare: A Compendium of Critical Factors for Success*, pages 233–253. Springer International Publishing, 2022.

[25] AA Abdelhamid, SK Towfek, N Khodadadi, AA Alhussan, DS Khafaga, MM Eid, and A Ibrahim. Waterwheel plant algorithm: A novel metaheuristic optimization method. *Processes*, 11(5):Article 5, 2023.

[26] G. Rajarajeshwari and G. C. Selvi. Application of artificial intelligence for classification, segmentation, early detection, early diagnosis, and grading of diabetic retinopathy from fundus retinal images: A comprehensive review. *IEEE Access*, 12:172499–172536, 2024.

[27] Shi, Y., Fang, J., Li, J., Yu, K., Zhu, J., & Lu, Y. (2024). Fracture risk prediction in diabetes patients based on Lasso feature selection and Machine Learning. *Computer Methods in Biomechanics and Biomedical Engineering*, 1-17.

[28] AH Alharbi, SK Towfek, AA Abdelhamid, A Ibrahim, MM Eid, DS Khafaga, N Khodadadi, L Abualigah, and M Saber. Diagnosis of monkeypox disease using transfer learning and binary advanced dipper throated optimization algorithm. *Biomimetics*, 8(3):Article 3, 2023.

[29] OY Dweekat and SS Lam. Optimized design of hybrid genetic algorithm with multilayer perceptron to predict patients with diabetes. *Soft Computing*, 27(10):6205–6222, 2023.

[30] Diabetes dataset. [dataset]. Retrieved March 13, 2024, from https://www.kaggle.com/datasets/mathchi/diabetes-data-set.