BAREKENG: Journal of Mathematics and Its Applications
December 2025 Volume 19 Issue 4 Page 2531–2542
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss4pp2531-2542

IMPROVING SUPPORT VECTOR MACHINE PERFORMANCE WITH BINARY GAUSSIAN IMPROVED WHALE OPTIMIZATION ALGORITHM: A CASE STUDY ON DIABETES DATA

Haidar Ahmad Fajri[1], **Safrizal Ardana Ardiyansa**[2], **Syaiful Anam**[3*], **Natasha Clarrisa Maharani**[4], **Eric Julianto**[5]

[1,2,3,4]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya*
Jln. Veteran, Malang, 65145, Indonesia

[1,2,4,5]*Braincore Indonesia*
Jln. Letjen S. Parman, No. 28, Jakarta, 11470, Indonesia

*Corresponding author's e-mail: *syaiful@ub.ac.id*

Article History:

Received: 21st January 2025
Revised: 29th April 2025
Accepted: 7th June 2025
Available online: 1st September 2025

Keywords:

Diabetes Mellitus;
Feature selection;
Gaussian Improved WOA;
Support Vector Machine.

ABSTRACT

Diabetes mellitus is a chronic condition with high blood sugar that can cause severe organ damage, affecting all ages globally. Early diagnosis is crucial for improving patients' quality of life, and machine learning offers a promising approach. The Support Vector Machine (SVM) is effective for classification, but feature selection is essential to enhance the relevance of features. The Whale Optimization Algorithm (WOA) is an optimal method for global feature selection, but it has a drawback-premature convergence, which can lead to suboptimal results. This issue should be addressed by modifying mutation operations, convergence factors, and population initialization, resulting in Binary Gaussian IWOA (BGIWOA). This research focuses on feature selection using BGIWOA, comparing it with Variance Inflation Factor (VIF) using SVM. The result show that BGIWOA is better than VIF and the best configuration BGIWOA's parameter is $B = 1$, $N = 50$ with linear kernel. This configuration produces the best accuracy of 95.00%. BGIWOA-SVM demonstrates better accuracy with stable consistency compared to VIF-SVM. The best SVM model achieves average accuracy of 95.62% for training data and 95.58% for validation data, with an accuracy of 93.85% for the test data. This model also yields an average precision of 94.00%, a recall of 91.00%, and an f_1-score of 92.00%. The model was also better than SVM without optimization, which only achieved a training accuracy of 84.25% and a testing accuracy of 81.30%. This model can assist in diagnosing diabetes with accurate and consistent predictions for new data. The results are specific to the diabetes dataset used in this research, so further testing on other binary datasets is necessary to confirm the model's effectiveness and generalizability across different domains and types of data.

How to cite this article:

H. A. Fajri, S. A. Ardiyansa, S. Anam, N. C. Maharani, and E. Julianto, "IMPROVING SUPPORT VECTOR MACHINE PERFORMANCE WITH BINARY GAUSSIAN IMPROVED WHALE OPTIMIZATION ALGORITHM: A CASE STUDY ON DIABETES DATA", *BAREKENG: J. Math. & App.*, vol. 19, no. 4, pp. 2531-2542, December, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Diabetes mellitus is a global health challenge affecting millions worldwide [1]. It is characterized by chronic high blood sugar levels that, if untreated, can lead to severe complications such as heart disease, kidney failure, and blindness [2][3]. The two primary types of diabetes are Type 1 Diabetes Mellitus (T1DM), caused by an autoimmune attack on insulin-producing beta cells [4], and Type 2 Diabetes Mellitus (T2DM), characterized by insulin resistance [5]. Among these, T2DM accounts for the vast majority of cases globally [6][7]. The World Health Organization (WHO) reports a significant rise in diabetes prevalence, with the number of affected individuals increasing from 108 million in 1980 to 422 million in 2014 [7][8]. For example, in Indonesia has sixth rank with 10.3 million diabetes patients in 2017. This number is expected to keep increasing to 16.7 million by 2045 [9]. This awareness is essential to reduce diabetes-related mortality by enabling early diagnosis and preventing long-term complications [1].

The early detection of diabetes is crucial for managing the disease and preventing complications [10]. However, diagnosing diabetes is challenging [11], especially in the early stages when symptoms may not be apparent [12]. Traditional diagnostic methods often involve evaluating test results [13], which can be complex and subject to human error [14]. This highlights the need for advanced, automated systems that can support healthcare professionals in making accurate diagnoses [15][16].

Machine learning models can detect patterns and risk factors in diabetes datasets, aiding diagnosis, [17], and have effectively solved real-world problems like product recommendations and object recognition [18]. In healthcare, it successfully predicts heart attacks [19], optimizes costs and delivery routes in industry [20], and even provides insights in politics [21]. Several machine learning models are commonly used to classify in diagnosis process, which are K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), and Support Vector Machines (SVM) [22]. KNN faces challenges with distance measurement and computational complexity [23], while ANN is robust but requires large amounts of data. However, SVM is a reliable and widely used classifier with higher accuracy, stability, and training speed [24].

The results of previous research on SVM produced an excellent accuracy rate of 99.2% compared to 97.2% for KNN and 92% for ANN [14][25][26]. The SVM performs well in classifying complex datasets, making it advantageous for detecting diabetes [14]. However, patient data usually presents the challenge of containing unnecessary or redundant features [16]. Redundant features become an obstacle in training process and will affect the prediction accuracy. the effectiveness of SVM is not solely determined by the algorithm itself, but also by the quality of features used during training. Real-world medical data often include noisy, irrelevant, or redundant features. Therefore, selecting features carefully is important [27]. Proper feature selection can significantly reduce training time, enhance generalization, and increase prediction accuracy [28]. To address this, hybrid approaches combining SVM with optimization-based feature selection methods have gained traction. These approaches aim to improve classification performance by selecting the most informative features while preserving or enhancing SVM's strengths [27][28].

In recent years, there has been growing interest in hybrid approaches combining SVM with optimization for feature selection techniques [16]. These hybrid models aim to achieve superior performance. The primary goal of these approaches is to automatically identify and select the most relevant features from the original dataset [27], while preserving or enhancing the advantages of SVM, such as its high accuracy, robustness to overfitting, and efficiency in handling high-dimensional data [27][28]. One hybrid approach combines SVM with metaheuristic optimization techniques and variable correlation analysis with statistics like Variance Inflation Factors (VIF). However, variable correlation analysis has the limitation that it is challenging to capture non-linear relationships and relies on linear assumptions, leading to sub-optimal feature selection on complex data sets [28].

Therefore, metaheuristic algorithm is necessary to solve this issue. Metaheuristic algorithm can be an appropriate solution due to their exploitation and exploration capabilities to search the relevant features [11]. There are several metaheuristic algorithms, such as Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Ant Colony Optimization (ACO), and Whale Optimization Algorithm (WOA). The WOA algorithm is more superior than PSO [29][30], ACO, and GA because it has a faster convergence speed and higher search accuracy at the same iteration [31]. The main advantage of WOA is its ability to perform a global search in the search space. This means that the WOA algorithm can find the optimal subset of features with the most significant impact on the classification goal of diabetes detection [32].

WOA have a fast convergence capability in some cases, which can save the computational time required to search for the best combination of features, resulting in reliable accuracy above 90% [33]. However, the standard WOA relies on a convergence factor parameter or ($\mathcal{C}$) that will decrease linearly from two to zero for each iteration. It makes WOA can easily trap into a local optimum. It causes the combination of features obtained by WOA is not the best combination. Therefore, an improvement from WOA should be applied to address this problem. The improvement of WOA can be obtained by modifying convergence factor to non-linear and improvement of mutation with Gaussian distribution. This improvement algorithm is called Gaussian Improved WOA or GIWOA that was proposed by Ning and Cao [34].

Previous research on GIWOA showed an excellent performance in finding the optimum value in solving continuous functions, especially on eight benchmark functions tested [34]. The algorithm has proven effective in enhancing global exploration and local exploitation capabilities. These are two essential aspects to ensure the solution not only explores the solution space widely but also consolidates the best solution in the local region. The superior performance of GIWOA is seen in comparison with other methods such as Independent Principal Element Search (IPES), Iterative First-Order Algorithm (IFOA), combinatorial decomposition, and Independent Component Analysis (ICA), which are often used in optimization and data analysis.

However, the problem in the feature selection process is in binary space, so GIWOA must be modified to a binary version. GIWOA has excellent potential to be adapted to binary form through one of the transfer functions. GIWOA, which has been adapted to binary, is expected to be a competitive algorithm with excellent performance in the feature selection process, especially in diagnosing diseases such as diabetes mellitus [34]. This prompts this article to modify the conventional GIWOA into a Binary Gaussian Improved Whale Optimization Algorithm (BGIWOA). This research will assist the medical team in making better decisions for diabetes mellitus patients and potentially reduce the mortality rate from this disease.

This article will discuss research methodology, results, and conclusions. The methodological research section will contain the research flow and dataset used in this research. In the second section, the preliminaries of this article are also explained, which briefly include concepts about SVM, GIWOA, BGIWOA, and VIF. The performance comparison results of feature selection using BGIWOA and VIF from this article will be explained in detail in the results and discussion section. Meanwhile, the last section is the conclusion and suggestions for further research.

2. RESEARCH METHODS

The research procedure detailed in **Figure 1**, involves several critical stages to refine diabetes diagnosis through advanced machine learning methods. It begins with importing the dataset into Google Colab, which undergoes preprocessing, including cleaning and normalization, before being split into training, testing, and validation sets. This ensures that the model is trained effectively and evaluated on different data subsets to prevent overfitting. The next stage involves feature selection using the BGIWOA. During this phase, specific parameters for BGIWOA are selected, and the SVM kernel is also chosen. The feature selection results from BGIWOA are compared with those obtained from the traditional VIF method to assess their relative effectiveness. VIF is used as a comparison because it is a well-known, straightforward method for identifying correlated features, helping reduce data redundancy [24]. Comparing it with BGIWOA allows for the evaluation of whether the optimization-based approach results in better feature selection and model performance. Once feature selection is complete, the selected features are used to train the SVM model, which incorporates various kernel types. An iterative process is employed to explore different SVM kernels and BGIWOA parameters, explicitly focusing on finding the optimal kernel ($\boldsymbol{k}$) for SVM, whale population parameter ($\boldsymbol{\mathcal{N}}$), and wave spiral-shaped parameter ($\boldsymbol{\mathcal{B}}$) of BGIWOA. This helps fine-tune the model for better performance [32][34].

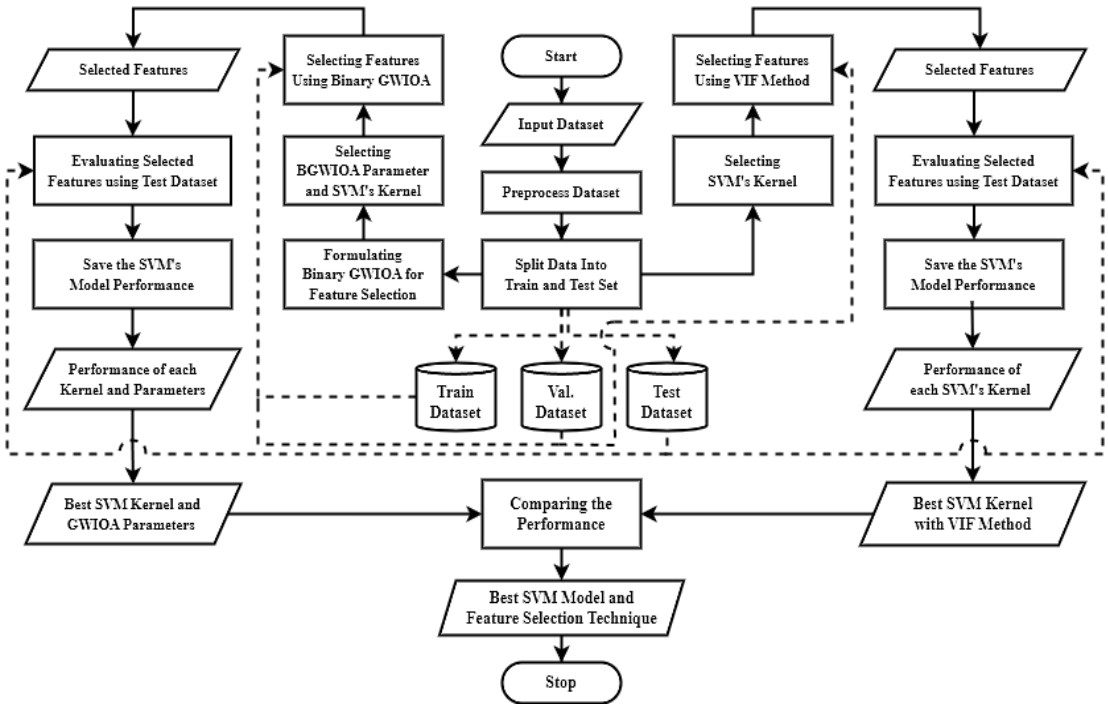


Figure 1. Research Step

The final stage evaluates the SVM model's performance using key metrics, including the confusion matrix, classification report, and average accuracy for training, testing, and validation datasets. The confusion matrix shows true positives, false positives, and false negatives, while the classification report provides precision, recall, and f_1-score to assess performance across different classes. Average accuracy reflects the overall effectiveness of the model. These metrics are vital for comparing configurations and hyperparameters, helping fine-tune the model for optimal performance. This process ensures the SVM model is reliable and accurate in diagnosing diabetes in real-world scenarios [27].

2.1 Dataset and Preprocessing Methods

The binary classification task in this research focuses on diabetes due to its high prevalence and clinical significance, making it a relevant and impactful case for predictive modeling. The dataset was obtained from Kaggle with the samples in **Table 1**, because it is openly accessible, widely used in research, and offers comprehensive health-related features not readily available from other sources. Although the classes are somewhat imbalanced, no extra steps were taken to balance them because SVM can handle this issue well, keeping the model reliable [27]. The dataset used originates from 4303 patients diagnosed with diabetes. The dataset used is including seventeen features that potentially influence health outcomes related to diabetes. These features encompass demographic features such as age and gender of patient, along with clinical measurements like Body Mass Index (BMI), which provides insight into obesity levels [35]. Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) also the features which are important for assessing diabetes. Metabolic indicators such as Fasting Plasma Glucose (FPG), cholesterol levels (Chol), and triglycerides (Tri) offer a detailed view of the patient's glucose and lipid metabolism. The dataset includes High Density Lipoprotein (HDL) and Low Density Lipoprotein (LDL), which are essential for evaluating cardiovascular health [36]. Liver function is assessed through Alanine Level Transaminase (ALT). Additionally, renal function is represented by Blood Urea Nitrogen (BUN) and Creatinine Clearance Rate (CCR). Moreover, lifestyle factors such as smoking and drinking habits, as well as a family history of diabetes, are incorporated to provide a holistic view of the patient's health profiles.

Table 1. Sample of the Dataset

Age	Gender	BMI	SBP	DBP	FPG	Chol	Tri	HDL	LDL	ALT	BUN	…	Diabetic
28	1	20.1	119	81	5.80	4.36	0.86	0.90	2.43	12.0	5.40	…	0
40	1	17.7	97	54	4.60	3.70	1.02	1.50	2.04	9.2	3.70	…	0
40	2	19.7	85	53	5.30	5.87	1.29	1.75	3.37	10.1	4.10	…	0
43	1	23.1	111	71	4.50	4.05	0.74	1.27	2.60	36.5	4.38	…	0
36	1	26.5	130	82	5.54	6.69	3.49	0.91	3.64	69.3	3.86	…	0

***Data source**: Kaggle*

2.2 Gaussian Improved Whale Optimization Algorithm

Humpback whales are large, social animals known for hunting in groups using a unique method called bubble-net feeding. Goldbogen et al. studied this behavior by tagging nine whales and recording 300 feeding events. They identified two main maneuvers: upward spirals and double loops. In the upward-spiral, whales dive deep and release bubbles in a spiral as they rise. The double-loop has three stages: the coral loop, the lobtail, and the capture loop **[37]**.

Mirjalili and Lewis developed a mathematical model of whale behaviour to solve benchmark optimization problems. Their method outperformed several other algorithms, including Particle Swarm Optimization (PSO), Particle Swarm Optimization with Passive Congregation (PSOPC), Hybrid Particle Swarm Optimization (HPSO), and Discrete Hybrid Particle Swarm Optimization with Ant Colony Optimization (DHPSACO), in terms of performance and convergence. The Whale Optimization Algorithm (WOA) uses a linearly decreasing convergence factor from 2 to 0, enhancing its early-stage global search ability. However, because convergence is slow initially and faster later, WOA can get trapped in local optima, especially for problems with multiple peaks. This linear decrease does not fully reflect the complexity of the search process **[29]**.

Gui-Ying Ning and Dun-Qian Cao introduced the updated WOA algorithm as an update of the standard WOA that improves the convergence of global explorability and local exploitability **[34]**. In finding the position of the whale index vector, the WOA is divided into two stages: exploration and exploitation. The convergence factor determines exploration and exploitation stages, and the WOA in **Equation (1)** updates the convergence factor to accelerate global convergence and avoid convergence to the local optimum.

$$\boldsymbol{a} = \begin{cases} \mathbf{2}^{(\mathbf{1}-\boldsymbol{t}\boldsymbol{T}^{-\mathbf{1}})}, & \boldsymbol{t} < \mathbf{0.7}\boldsymbol{T} \\ \mathbf{2} - \mathbf{2}\boldsymbol{t}\boldsymbol{T}^{-\mathbf{1}}, & \boldsymbol{t} \geq \mathbf{0.7}\boldsymbol{T} \end{cases} \tag{1}$$

where $\boldsymbol{t}$ is the current iteration and $\boldsymbol{T}$ is the maximum iteration.

Defined the generalized absolute value operation $|\cdot|: \mathbb{R}^{\mathcal{D}} \to \mathbb{R}^{\mathcal{D}}$, with $|\vec{\boldsymbol{v}}| = |(\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots \boldsymbol{v_{\mathcal{D}}})| = (|\boldsymbol{v_1}|, |\boldsymbol{v_2}|, \ldots |\boldsymbol{v_{\mathcal{D}}}|)$, where $\vec{\boldsymbol{v}} = (\boldsymbol{v_1}, \boldsymbol{v_2}, .. \boldsymbol{v_{\mathcal{D}}}) \in \mathbb{R}^{\mathcal{D}}$. In the GIWOA operation, the whale position is defined as $\vec{\boldsymbol{Z}}_{\boldsymbol{i}} \in \mathbb{R}^{\mathcal{D}}$. The optimal whale position is the position of the whale on its prey for $\mathcal{N}$ total whale populations determined at a random value $\boldsymbol{p} \in [\mathbf{0}, \mathbf{1}]$. Position updated by looking at $\boldsymbol{p} = \mathbf{0.5}$, if $\boldsymbol{p} < \mathbf{0.5}$ will look at the value of $|\boldsymbol{A}|$. If $|\boldsymbol{A}| \geq \mathbf{1}$, it means that the prey attack stage occurs with the whale's position away from the prey so that the whale's position is updated by looking at the position of the prey $\vec{\boldsymbol{Z}}(\boldsymbol{t})_{\boldsymbol{i}}^{*}$ as best position or prey position and $\vec{\boldsymbol{Z}}(\boldsymbol{t})_{\boldsymbol{i}}$ is consider position in **Equation (2)** as follows:

$$\begin{aligned} \vec{\boldsymbol{D}} &= \left|\boldsymbol{C}\vec{\boldsymbol{Z}}(\boldsymbol{t})_{\boldsymbol{i}}^{*} - \vec{\boldsymbol{Z}}(\boldsymbol{t})_{\boldsymbol{i}}\right| \\ \vec{\boldsymbol{Z}}(\boldsymbol{t}+\mathbf{1})_{\boldsymbol{i}} &= \left(\mathbf{1} - \boldsymbol{t}\boldsymbol{T}^{-\mathbf{1}}\right)\vec{\boldsymbol{Z}}(\boldsymbol{t})_{\boldsymbol{i}_{\boldsymbol{rand}}} - \boldsymbol{A}\vec{\boldsymbol{D}} \\ \boldsymbol{A} &= \mathbf{2}\boldsymbol{a}\boldsymbol{r} - \boldsymbol{a}, \boldsymbol{C} = \mathbf{2}\boldsymbol{r} \end{aligned} \tag{2}$$

where $\boldsymbol{A}\vec{\boldsymbol{D}}$ is the step size of the movement, and $\boldsymbol{r}$ is a random number at $[\mathbf{0}, \mathbf{1}]$. If $|\boldsymbol{A}| < \mathbf{1}$ which means it is close to the prey position, then the whale position is updated in **Equation (3)**, namely:

$$\vec{\boldsymbol{Z}}(\boldsymbol{t}+\mathbf{1})_{\boldsymbol{i}} = \boldsymbol{t}\boldsymbol{T}^{-\mathbf{1}}\vec{\boldsymbol{Z}}(\boldsymbol{t})_{\boldsymbol{i}}^{*} - \boldsymbol{A}\vec{\boldsymbol{D}} \tag{3}$$

If the random value $\boldsymbol{p} \geq \mathbf{0.5}$ then it indicates a bubble-net attack, which is represented by **Equation (4)**,

$$\vec{Z}(t+1)_i = e^{\mathcal{B}l}\cos(2\pi l)\left|\vec{Z}(t)_i^* - \vec{Z}(t)_i\right| + \vec{Z}(t)_i^* \quad (4)$$

where $\mathcal{B}$ is a wave spiral-shaped constant and l is random number at $[-1, 1]$. The process of obtaining the optimal position is depicted in **Figure 2**.

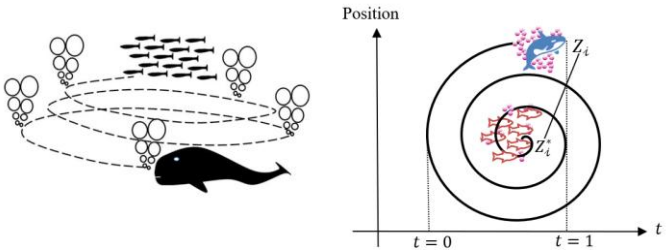


Figure 2. Process of Obtaining the Optimal Position

A Gaussian mutation distribution is implemented on the updated best whale positions by considering the distance and midpoint between the current and optimal positions. This involves usage of a randomly chosen variance distribution formula to prevent premature convergence to a particular solution by introducing random variation. This maintains diversity in the population to find a globally superior optimal solution. The variance distribution formula is formulated in **Equation (5)** as follows:

$$\vec{Z}(t+1)_i \sim \mathcal{N}\left(2^{-1}\left(\vec{Z}(t)_i^* + \vec{Z}(t)_i\right), \left|\vec{Z}(t)_i^* - \vec{Z}(t)_i\right|\right) \quad (5)$$

The values in the Gaussian operation are constrained with sigmoid clip to ensure that the mutated individuals remain within the valid bounds of the initial domain and avoid premature convergence to a sub-optimal solution **[34]**.

2.3 Binary Gaussian Improved Whale Optimization Algorithm

This approach builds on the method introduced in **[34]**, where the Gaussian update operation was used for continuous optimization problems. In this research, continuous modeling is adapted to binary data through a transfer function, which discretizes the solution space. This adaptation enables the algorithm to perform binary feature selection, making it more suitable for tasks like feature selection, which was not addressed in the original WOA method **[34]**. The best whale index position updated with the Gauss operation is then converted to binary space using a transfer function. In binary space, the desired solution is $\vec{Z}_{\mathbb{B}}(t) = (Z(t)^1, Z(t)^2, ..., Z(t)^j, ..., Z(t)^{\mathcal{D}})$ with $Z(t)^j = 1$ or $Z(t)^j = 0$ for $j = 1, 2, ..., \mathcal{D}$. The value $Z(t)^j = 1$ states that the j-th feature will be selected, while the value $Z(t)^j = 0$ states that the j-th feature is not selected. The sigmoid function converts a continuous function to a binary space used as a selection feature **[38]**. This feature selection has a threshold that depends on the random value $R \sim U(0, 1)$, so the threshold limit of $R = 0.5$ was chosen. The sigmoid function is defined as in **Equation (6)**,

$$\vec{Z}(t)_i^j = \begin{cases} 1, & if \left[1 + exp\left(-\vec{Z}(t)_i^j\right)\right]^{-1} > 0.5 \\ 0, & if \left[1 + exp\left(-\vec{Z}(t)_i^j\right)\right]^{-1} \leq 0.5 \end{cases} \quad (6)$$

There are three stages in BGIWOA, namely initialization, exploration, and exploitation. The initialization stage consists of initializing the whale population ($\mathcal{N}$), whale position ($\vec{Z}(t)$) and several spiral-shaped waves ($\mathcal{B}$) of each whale randomly. After the initialization stage, the convergence factor value will be calculated using **Equation (2)**, and each whale will calculate the fitness value to get the best solution $\left(\vec{Z}(t)_i^*\right)$. The exploration process in BGIWOA is carried out after updating the convergence factor and the best solution using **Equation (3)**, which is then carried out in the exploitation stage using **Equation (4)** by looking at the random value $p \in [0, 1]$ and the value of $|A|$. If the best solution is obtained, the whale position will be updated using the Gaussian operation to check whether it has reached the global optimum in **Equation (5)**. Whale position update can be done if the whale position has the best fitness value and fulfils the conditions in **Equation (6)**. This process will be repeated until it reaches the maximum iteration. The overall stages in BGIWOA are shown in the flowchart in **Figure 3**.

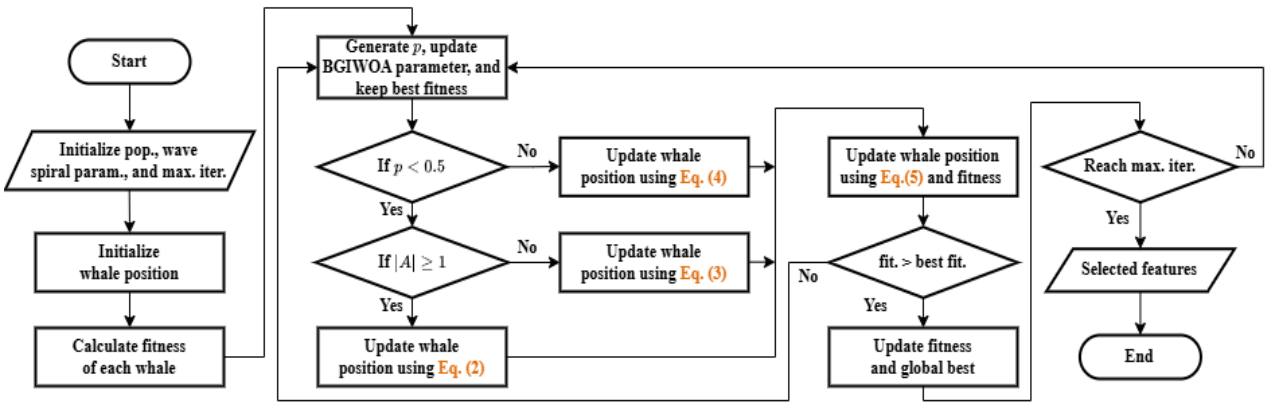


Figure 3. BGIWOA steps

2.4 Variance Inflation Factors

VIF serves as a metric for gauging and quantifying the inflation of variance [39]. VIF values are often used in regression analysis to detect multicollinearity by showing how much a variable's variance is inflated. A VIF of 1 indicates no correlation, $1-5$ suggests moderate correlation, and values above five show high multicollinearity that may impact model stability. The formula for calculating VIF is in **Equation (7)**,

$$\boldsymbol{VIF = \left(1 - R_i^2\right)^{-1}} \tag{7}$$

Calculation of the VIF by regressing each predictor against all the other predictors. The resulting R^2 or R-squared value from this regression indicates how much multicollinearity exists for that predictor. A higher R-squared means greater multicollinearity [40]. The coefficient of determination (R^2) is calculated using the formula in **Equation (8)**,

$$\boldsymbol{R_i^2 = 1 - \left[\sum_{i=1}^{m}(P_i - Y_i)^2\right]\left[\sum_{i=1}^{m}(\bar{Y} - Y_i)^2\right]^{-1}} \tag{8}$$

where m is amount of data, P_i is the predicted i-th value, while Y_i is the actual i-th value with $\bar{Y}$ as the mean of Y_i.

3. RESULTS AND DISCUSSION

3.1 Importance Feature based on VIF and BGIWOA Technique

The feature selection process was conducted to ensure that each feature used in the model did not experience multicollinearity problems and contributed informatively to the classification, as shown in **Table 2**. The analysis was conducted using two approaches: statistical evaluation through VIF and selection based on the BGIWOA metaheuristic algorithm. VIF is used to detect multicollinearity in numerical features, with higher VIF values indicating potential overlapping information between features. Meanwhile, BGIWOA was used to select features based on their optimal contribution to the performance of the classification model, with the whale index as an indicator of feature selection. BGIWOA determines feature selection based on the whale index value with a threshold of 0.5, and demonstrates the ability to select relevant features for classification models. In this experiment, manual hyperparameter tuning was conducted through sensitivity analysis by testing different population size values $\mathcal{N} = 30, 35, 40, 45, 50$ to evaluate the impact of search agent quantity on the exploration capability and convergence performance of BGIWOA. These values aim to balance solution diversity. The number of spiral waves $\mathcal{B} = 1, 2, 3$ was tested to analyze the influence of the bubble-net mechanism on the algorithm's search behavior. Although $\mathcal{B} = 1$ is commonly used as the default in previous studies, variations in $\mathcal{B}$ were explored to examine potential improvements in avoiding local optima. The SVM cost parameter $c = 10$ was chosen based on preliminary experiments that showed it provided a good balance between margin maximization and classification error [41]. Both approaches use the

same parameter to ensure consistency in the feature selection process and improve the accuracy of the classification model.

Table 2. VIF Value and BGIWOA of Each Feature

Feature	VIF Value	BGIWOA	Feature	VIF Value	BGIWOA
Age	1.72171	non-selected	Chol	1.97521	selected
Gender	2.14205	non-selected	Tri	1.45143	non-selected
BMI	1.46936	non-selected	HDL	1.31472	selected
SBP	1.83487	selected	LDL	1.96127	non-selected
DBP	1.66322	selected	ALT	1.29504	selected
FPG	1.61670	selected	BUN	1.17796	selected
CCR	1.99560	non-selected	Smoke	> 5	selected
FFPG	1.59270	selected	Drink	> 5	selected
Fam. His.	> 5	selected			

Table 2 shows that numerical features such as Age, BMI, SBP, Chol, Tri, HDL, LDL, ALT, BUN, FPG, CCR, and DBP have VIF values below 2, indicating no significant multicollinearity among these features. Categorical features such as smoking, drinking, and family history had higher VIF values, but this did not affect the analysis due to using one-hot encoding on categorical data. Correlations between categorical features were tested using Cramer's V, which showed moderate correlations (0.60) between smoking and drinking, and high correlations with the target variables (0.77 and 0.78). These features remained selected in the model due to their important contribution to prediction. Unlike VIF, BGIWOA can handle categorical data that has been converted to numerical through the one-hot encoding technique without affecting the calculation of the optimal solution, achieving a highest fitness value of 95%. BGIWOA identified features such as SBP, DBP, FPG, Chol, HDL, ALT, BUN, FFPG, smoking, drinking, family history. history as important features based on fitness value. These feature selection results were consistent with the correlation analysis that showed good relationships between the selected features, such as a correlation of 0.66 between HDL and LDL, and 0.65 between SBP-DBP and Chol-LDL. The features selected by BGIWOA showed a strong correlation, making them highly relevant for feature selection. When used in the SVM model training, these features improved the model's performance and accuracy, highlighting the importance of proper feature selection.

3.2 Comparison between VIF and BGIWOA Technique

The performance of the VIF and BGIWOA algorithm is thoroughly evaluated through a cross-validation process with $k = 10$ folds and by fine-tuning a SVM model. This evaluation is conducted using the mean accuracy across the training, validation, and test sets as the primary metric for assessing the model's effectiveness. Following this, the performance of the BGIWOA-enhanced SVM model is systematically compared against a standard SVM model trained on features selected based on the VIF, as detailed in **Table 3**. This comparison aims to identify which model configuration yields the highest predictive accuracy, thereby determining the most robust approach for optimizing the SVM's classification capabilities.

Table 3. SVM Performance Based on VIF and BGIWOA Technique for Feature Selection

BGIWOA Parameter		Mean Accuracy Train			Mean Accuracy Validation		
$\mathcal{N}$	$\mathcal{B}$	Linear	Polynom.	RBF	Linear	Polynom.	RBF
30	1	85.58%	90.43%	93.68%	90.66%	88.92%	89.96%
	2	86.05%	91.28%	94.88%	91.01%	88.57%	89.26%
	3	85.89%	88.02%	92.09%	90.43%	89.15%	90.54%
35	1	85.93%	89.07%	92.67%	91.66%	90.38%	92.24%
	2	85.93%	91.12%	94.92%	91.66%	88.99%	90.03%
	3	84.57%	87.83%	92.17%	91.31%	91.19%	91.43%
40	1	86.18%	90.08%	94.78%	93.26%	91.18%	92.90%
	2	85.42%	89.95%	93.78%	92.26%	91.61%	92.96%
	3	86.05%	92.21%	96.04%	92.54%	91.38%	92.31%
45	1	90.66%	91.90%	95.04%	93.31%	91.10%	92.50%
	2	90.78%	91.32%	95.11%	94.12%	91.45%	92.03%
	3	90.78%	90.43%	93.76%	93.66%	91.22%	92.38%
50	1	**95.62%**	**97.39%**	**97.98%**	**95.58%**	94.49%	**95.50%**

BGIWOA Parameter		Mean Accuracy Train			Mean Accuracy Validation		
$\mathcal{N}$	$\mathcal{B}$	Linear	Polynom.	RBF	Linear	Polynom.	RBF
	2	95.37%	96.30%	96.59%	93.75%	**94.69%**	92.61%
	3	94.71%	93.60%	92.01%	93.67%	93.76%	93.61%
VIF		86.70%	84.19%	82.89%	85.70%	84.19%	82.68%

Unlike BGIWOA, which was tested with various parameter settings, the VIF method selects features based solely on a fixed threshold (VIF < 5) because VIF does not have parameters to adjust for the optimization process. VIF is a statistical test that detects multicollinearity for feature selection. Both methods were tested using the same SVM kernels and validation scheme. The results show that most BGIWOA configurations outperformed VIF in validation and test accuracy. VIF's lower performance comes from focusing on multicollinearity removal rather than selecting relevant features for the target class.

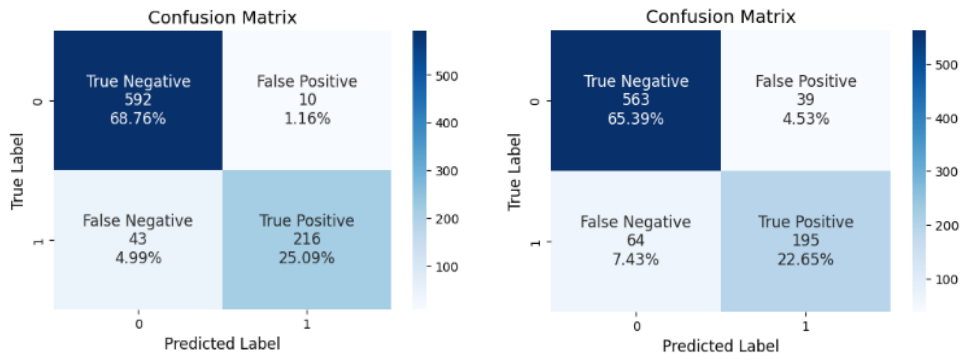
To augment the analysis to obtain a good model, the standard deviation value, the average accuracy of the test data, and a statistical t-test were examined to validate whether the observed improvement in accuracy was significant. The results obtained show that the model has high and stable performance on training data, validation data and test data shown in **Table 4** for comparing the best results of BGIWOA with VIF. The low standard deviation shows minimal variability, indicating the model is reliable and consistent. Based on an independent t-test using the mean and standard deviation values from cross-validation, there is a highly significant difference in validation accuracy between BGIWOA and VIF with a very small chance that the difference is due to chance ($t_{\text{val,train}}$-values > 50), p-values < 0.05).

Table 4. Comparison Performance: Std. dev. with VIF and BGIWOA

BGIWOA		Std. Accuracy Train			Std. Accuracy Validation			Accuracy Test		
$\mathcal{N}$	$\mathcal{B}$	Linear	Polynom	RBF	Linear	Polynom	RBF	Linear	Polynom	RBF
50	1	0.14%	0.13%	**0.08%**	**0.89%**	1.29%	1.07%	**93.85%**	93.50%	93.61%
VIF		0.23%	0.27%	0.18%	1.44%	2.04%	2.16%	88.12%	85.21%	84.16%

3.3 Best SVM's Performance

Based on **Table 3**, the SVM model that achieved the best performance when using BGIWOA is with a linear kernel, whale population $\mathcal{N} = 50$ and wave spiral-shaped $\mathcal{B} = 1$ because it produces a test accuracy performance of 93.85% and the accuracy value on the validation data is the largest compared to other data with the accuracy of 95.58%. The accuracy is also better than that of SVM without optimization, which only achieved 84.25% and 81.30% in the train and test data, respectively. These results were determined through a sensitivity analysis, in which various values of $\mathcal{N}$ and $\mathcal{B}$ were evaluated to assess their impact on model performance. Based on **Table 4**, shows that the best parameter of BGIWOA has small standard deviation so that the model produces consistent performance. **Table 3** shows that the greater number of populations, then the accuracy will also be increase because there will be more samples to find the best position, but the value of wave spiral has not significant effect on the accuracy of the model. As shown in **Table 4**, feature selection using BGIWOA yields higher values for true positives and true negatives compared to the SVM with VIF technique.



(a) **(b)**

Figure 4. Best SVM's Performance Based on (a) BGIWOA and (b) VIF Technique

In addition, BGIWOA technique demonstrates notable performance metrics, with an average precision of 94%, recall of 91%, and f_1-score of 92%, as detailed in **Table 5**. These results affirm that BGIWOA effectively enhances the SVM's capabilities, leading to great performance in diabetes diagnosis.

Table 5. Precision, Recall, and f_1-Score on Proposed Model

Label	Meaning	Precision	Recall	f_1-score
0	Patients do not have the risk of diabetes	93%	98%	96%
1	Patients have the risk of diabetes	96%	83%	89%
	Average	94%	91%	92%

4. CONCLUSION

BGIWOA successfully optimized feature selection for SVM in diabetes detection, achieving an accuracy of 95.00%, which outperforms the VIF method with t-values $t > 50$, p-values < 0.05. The best model used a linear kernel with a population size of $N = 50$ and wave spiral-shaped $\mathcal{B} = 1$. Increasing the population size slightly improved accuracy, but the wave spiral-shaped parameter had minimal impact on the model's performance. The model demonstrated an average accuracy of 95.62% for training data, 95.58% for validation data, and 93.85% for testing data, with a precision of 94%, a recall of 91%, and an f_1-score of 92%. These results surpassed the performance of SVM without optimization, which achieved 84.25% accuracy on training data and 81.30% on test data. These findings suggest that BGIWOA-SVM offers a more accurate and reliable approach for diabetes prediction, with improved feature selection and model performance. However, the results are specific to the diabetes dataset used in this study. Further testing on other binary datasets, optimizing hyperparameter selection, data augmentation, stricter cross-validation techniques, and integrating other optimization algorithms are needed to confirm the effectiveness and generalizability of the model across different domains and data types.

AUTHOR CONTRIBUTIONS

Haidar Ahmad Fajri: Conceptualization, Methodology, Writing - Review and Editing, Software, Formal Analysis, Visualization, Investigation. Safrizal Ardana Ardiyansa: Writing - Review and Editing, Data Curation, Formal Analysis, Visualization, Resources, Investigation. Syaiful Anam: Supervision, Funding Acquisition, Project administration, Writing - Review and Editing, Validation. Natasha Clarrisa Maharani: Writing - Original Draft, Resources. Eric Julianto: Funding Acquisition, Project Administration. All authors discussed the results and contributed to the final manuscript.

FUNDING STATEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for profit sectors

ACKNOWLEDGMENT

The authors thank all reviewers for their valuable feedback and Brawijaya University for the support and facilities provided during the completion of this work.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest to report regarding this study.

REFERENCES

[1] American Diabetes Association, "Diagnosis And Classification Of Diabetes Mellitus," *Diabetes Care*, vol. 34, no. Supplement_1, pp. S62–S69, 2011, doi: https://doi.org/10.2337/dc11-S062.

[2] M. Wątroba, A. D. Grabowska, and D. Szukiewicz, "EFFECTS OF DIABETES MELLITUS-RELATED DYSGLYCEMIA ON THE FUNCTIONS OF FLOOD-BRAIN BARRIER AND THE RISK OF DEMENTIA," *International Journal of Molecular Sciences*, vol. 24, no. 12, 2023.doi: https://doi.org/10.3390/ijms241210069

[3] F. Lotti and M. Maggi, "EFFECTS OF DIABETES MELLITUS ON SPERM QUALITY AND FERTILITY OUTCOMES: CLINICAL EVIDENCE," *Andrology*, vol. 11, no. 2, pp. 399–416, 2023, doi: https://doi.org/10.1111/andr.13342.

[4] A. Katsarou, S. Gudbjörnsdottir, A. Rawshani, D. Dabelea, E. Bonifacio, B. J. Anderson, L.M. Jacobsen, D. A. Schatz, and Å. Lernmark, "TYPE 1 DIABETES MELLITUS," *Nature Reviews Disease Primers*, vol. 3, no. 1, pp. 17016, 2017, doi: https://doi.org/10.1038/nrdp.2017.16.

[5] K. Yang, X. Yang, C. Jin, S. Ding, T. Liu, B. Ma, H. Sun, J. Zhang, and Y. Li, "GLOBAL BURDEN OF TYPE 1 DIABETES IN ADULTS AGED 65 YEARS AND OLDER, 1990-2019: POPULATION BASED STUDY," *BMJ*, pp. e078432, 2024, doi: https://doi.org/10.1136/bmj-2023-078432.

[6] H. Yaribeygi, T. Sathyapalan, S. L. Atkin, and A. Sahebkar, "MOLECULAR MECHANISMS LINKING OXIDATIVE STRESS AND DIABETES MELLITUS," *Oxidative Medicine and Cellular Longevity*, vol. 2020, pp. 1–13, 2020, doi: https://doi.org/10.1155/2020/8609213.

[7] WHO, "Diabetes." Accessed: Aug. 25, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[8] L. Pujiati, "PENGARUH KONTROL KADAR GULA DARAH DAN PEMILIHAN BALUTAN TERHADAP LAMA PENYEMBUHAN LUKA PADA PASIEN ULKUS DIABETIKUM," vol. 13, no. 2, 2020.

[9] N. H. Cho, J. E. Shaw, S. Karuranga, J. D. D. R. Fernandes, A. W. Ohlrogge, and B. Malanda, "IDF DIABETES ATLAS: GLOBAL ESTIMATES OF DIABETES PREVALENCE FOR 2017 AND PROJECTIONS FOR 2045," *Diabetes Research and Clinical Practice*, vol. 138, pp. 271–281, 2018, doi: https://doi.org/10.1016/j.diabres.2018.02.023.

[10] S. Syamsurizal, "TYPE-2 DIABETES MELLITUS OF DEGENERATIVE DISEASE," *Bioscience*, vol. 2, no. 1, pp. 34, 2018, doi: https://doi.org/10.24036/02018219980-0-00.

[11] R. Ambady and S. Chamukuttan, "EARLY DIAGNOSIS AND PREVENTION OF DIABETES IN DEVELOPING COUNTRIES," *Reviews in Endocrine and Metabolic Disorders*, vol. 9, no. 3, pp. 193, 2008, doi: https://doi.org/10.1007/s11154-008-9079-z.

[12] J. Hippisley-Cox and C. Coupland, "DIABETES TREATMENTS AND RISK OF AMPUTATION, BLINDNESS, SEVERE KIDNEY FAILURE, HYPERGLYCAEMIA, AND HYPOGLYCAEMIA: OPEN COHORT STUDY IN PRIMARY CARE," *BMJ*, pp. i1450, 2016, doi: https://doi.org/10.1136/bmj.i1450.

[13] A. Viloria, J.-P. Lis-Gutiérrez, M. Gaitán-Angulo, A. R. M. Godoy, G. C. Moreno, and S. J. Kamatkar, "METHODOLOGY FOR THE DESIGN OF A STUDENT PATTERN RECOGNITION ROOL TO FACILITATE THE TEACHING - LEARNING PROCESS THROUGH KNOWLEDGE DATA DISCOVERY (BIG DATA)," in *Data Mining and Big Data*, vol. 10943, Y. Tan, Y. Shi, and Q. Tang, Eds., in Lecture Notes in Computer Science, vol. 10943. , Cham: Springer International Publishing, 2018, pp. 670–679. doi: https://doi.org/10.1007/978-3-319-93803-5_63.

[14] A. Viloria, Y. Herazo-Beltran, D. Cabrera, and O. B. Pineda, "DIABETES DIAGNOSTIC PREDICTION USING VECTOR SUPPORT MACHINES," *Procedia Computer Science*, vol. 170, pp. 376–381, 2020, doi: https://doi.org/10.1016/j.procs.2020.03.065.

[15] S. Li, H. Zhao, Z. Ru, and Q. Sun, "PROBABILISTIC BACK ANALYSIS BASED ON BAYESIAN AND MULTI-OUTPUT SUPPORT VECTOR MACHINE FOR A HIGH CUT ROCK SLOPE," *Engineering Geology*, vol. 203, pp. 178–190, 2016, doi: https://doi.org/10.1016/j.enggeo.2015.11.004.

[16] Md. R. Islam, S. Banik, K. N. Rahman, and M. M. Rahman, "A Comparative Approach To Alleviating The Prevalence Of Diabetes Mellitus Using Machine Learning," *Computer Methods and Programs in Biomedicine Update*, vol. 4, pp. 100113, 2023, doi: https://doi.org/10.1016/j.cmpbup.2023.100113.

[17] H. Kaur and V. Kumari, "PREDICTIVE MODELLING AND ANALYTICS FOR DIABETES USING A MACHINE LEARNING APPROACH," *Applied Computing and Informatics*, vol. 18, no. 1/2, pp. 90–100, 2022, doi: https://doi.org/10.1016/j.aci.2018.12.004.

[18] V. Chaoji, R. Rastogi, and G. Roy, "Machine Learning In The Real World" in *Proceedings of the VLDB Endowment*, vol. 9, no. 13, 2016.doi: https://doi.org/10.14778/3007263.3007318

[19] S. A. Ardiyansa, N. C. Maharani, S. Anam, and E. Julianto, "OPTIMIZING HEART ATTACK DIAGNOSIS USING RANDOM FOREST WITH BAT ALGORITHM AND GREEDY CROSSOVER TECHNIQUE," *BAREKENG: Jurnal Ilmu Matematika dan Terapan,.*, vol. 18, no. 2, pp. 1053–1066, 2024, doi: https://doi.org/10.30598/barekengvol18iss2pp1053-1066.

[20] C. Yaiprasert and A. N. Hidayanto, "AI-Powered Ensemble Machine Learning To Optimize Cost Strategies In Logistics Business," *International Journal of Information Management Data Insights*, vol. 4, no. 1, pp. 100209, 2024, doi: https://doi.org/10.1016/j.jjimei.2023.100209.

[21] R. F. Khoiroh, E. Julianto, S. A. Adiyansa, H. A. Fajri, A. A. R. Yasa, and B. Sangapta, "IMPLEMENTASI SPEECH RECOGNITION WHISPER PADA DEBAT CALON WAKIL PRESIDEN REPUBLIK INDONESIA," *EXPLORE: Jurnal Informatika dan Komputer*, vol. 14, no. 2, 2024, doi: https://doi.org/10.35200/ex.v14i2.115.

[22] M. Bansal, A. Goyal, and A. Choudhary, "A COMPARATIVE ANALYSIS OF K-NEAREST NEIGHBOR, GENETIC, SUPPORT VECTOR MACHINE, DECISION TREE, AND LONG SHORT TERM MEMORY ALGORITHMS IN MACHINE LEARNING," *Decision Analytics Journal*, vol. 3, pp. 100071, 2022, doi: https://doi.org/10.1016/j.dajour.2022.100071.

[23] T. Liao, Z. Lei, T. Zhu, S. Zeng, Y. Li, and C. Yuan, "DEEP METRIC LEARNING FOR K-NEAREST NEIGHBOR CLASSICATION," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021, doi: https://doi.org/10.1109/TKDE.2021.3090275.

[24] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A COMPREHENSIVE SURVEY ON SUPPORT VECTOR MACHINE CLASSIFICATION: APPLICATIONS, CHALLENGES AND TRENDS," *Neurocomputing*, vol. 408, pp. 189–215, 2020, doi: https://doi.org/10.1016/j.neucom.2019.10.118.

[25] S. Srivastava, L. Sharma, V. Sharma, A. Kumar, and H. Darbari, "PREDICTION OF DIABETES USING ARTIFICIAL NEURAL NETWORK APPROACH," in *Engineering Vibration, Communication and Information Processing*, vol. 478, K. Ray, S. N. Sharan, S. Rawat, S. K. Jain, S. Srivastava, and A. Bandyopadhyay, Eds., in Lecture Notes in Electrical Engineering, vol. 478 , Singapore: Springer Singapore, 2019, pp. 679–687. doi: https://doi.org/10.1007/978-981-13-1642-5_59.

[26] A. Gunawan and I. Fenriana, "DESIGN OF DIABETES PREDICTION APPLICATION USING K-NEAREST NEIGHBOR ALGORITHM," *Bit-Tech*, vol. 6, no. 2, pp. 110–117, 2023, doi: https://doi.org/10.32877/bt.v6i2.939.

[27] Y. Zhang, G. Wang, F. Chung, and S. Wang, "SUPPORT VECTOR MACHINES WITH THE KNOWN FEATURE-EVOLUTION PRIORS," *Knowledge-Based Systems*, vol. 223, pp. 107048, 2021, doi: https://doi.org/10.1016/j.knosys.2021.107048.

[28] M. Afshar and H. Usefi, "OPTIMIZING FEATURE SELECTION METHODS BY REMOVING IRRELEVANT FEATURES USING SPARSE LEAST SQUARES," *Expert Systems with Applications*, vol. 200, pp. 116928, 2022, doi: https://doi.org/10.1016/j.eswa.2022.116928.

[29] S. Mirjalili and A. Lewis, "THE WHALE OPTIMIZATION ALGORITHM," *Advances in Engineering Software*, vol. 95, pp. 51–67, 2016, doi: https://doi.org/10.1016/j.advengsoft.2016.01.008.

[30] W. Xu and J. Hu, "A COMPARISON OF PARAMETERS OPTIMIZED-TYPE VMD METHODS USED IN BEARING FAULT DIAGNOSIS," *Journal of Physics: Conference Series*, vol. 2029, no. 1, pp. 012131, 2021, doi: https://doi.org/10.1088/1742-6596/2029/1/012131.

[31] L. Liu, J. Liang, K. Guo, C. Ke, D. He, and J. Chen, "DYNAMIC PATH PLANNING OF MOBILE ROBOT BASED ON IMPROVED SPARROW SEARCH ALGORITHM," *Biomimetics*, vol. 8, no. 2, pp. 182, 2023, doi: https://doi.org/10.3390/biomimetics8020182.

[32] H. Nematzadeh, R. Enayatifar, M. Mahmud, and E. Akbari, "FREQUENCY BASED FEATURE SELECTION METHOD USING WHALE ALGORITHM," *Genomics*, vol. 111, no. 6, pp. 1946–1955, 2019, doi: https://doi.org/10.1016/j.ygeno.2019.01.006.

[33] Mila Desi Anasanti, Khairunisa Hilyati, and Annisa Novtariany, "THE EXPLORING FEATURE SELECTION TECHNIQUES ON CLASSIFICATION ALGORITHMS FOR PREDICTING TYPE 2 DIABETES AT EARLY STAGE," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 5, pp. 832–839, 2022, doi: https://doi.org/10.29207/resti.v6i5.4419.

[34] G.-Y. Ning and D.-Q. Cao, "IMPROVED WHALE OPTIMIZATION ALGORITHM FOR SOLVING CONSTRAINED OPTIMIZATION PROBLEMS," *Discrete Dynamics in Nature and Society*, vol. 2021, pp. 1–13, 2021, doi: https://doi.org/10.1155/2021/8832251.

[35] D. Mohajan and H. K. Mohajan, "BODY MASS INDEX (BMI) IS A POPULAR ANTHROPOMETRIC TOOL TO MEASURE OBESITY AMONG ADULTS," *Journal of Innovations in Medical Research*, vol. 2, no. 4, pp. 25–33, 2023, doi: https://doi.org/10.56397/JIMR/2023.04.06.

[36] D. P. Wilson *et al.*, "USE OF LIPOPROTEIN(A) IN CLINICAL PRACTICE: A BIOMARKER WHOSE TIME HAS COME. A SCIENTIFIC STATEMENT FROM THE NATIONAL LIPID ASSOCIATION," *Journal of Clinical Lipidology*, vol. 16, no. 5, pp. e77–e95, Sep. 2022, doi: https://doi.org/10.1016/j.jacl.2022.08.007.

[37] J. A. Goldbogen, A. S. Friedlaender, J. Calambokidis, M. F. McKenna, M. Simon, and D. P. Nowacek, "INTEGRATIVE APPROACHES TO THE STUDY OF BALEEN WHALE DIVING BEHAVIOR, FEEDING PERFORMANCE, AND FORAGING ECOLOGY," *BioScience*, vol. 63, no. 2, pp. 90–100, 2013, doi: https://doi.org/10.1525/bio.2013.63.2.5.

[38] M. Barhoush, B. H. Abed-alguni, and N. E. A. Al-qudah, "IMPROVED DISCRETE SALP SWARM ALGORITHM USING EXPLORATION AND EXPLOITATION TECHNIQUES FOR FEATURE SELECTION IN INTRUSION DETECTION SYSTEMS," *The Journal of Supercomputing*, vol. 79, no. 18, pp. 21265–21309, 2023, doi: https://doi.org/10.1007/s11227-023-05444-4

[39] J. I. Daoud, "MULTICOLLINEARITY AND REGRESSION ANALYSIS," *Journal of Physics: Conference Series*, vol. 949, no. 1, pp. 012009, 2017, doi: https://doi.org/10.1088/1742-6596/949/1/012009.

[40] D. Chicco, M. J. Warrens, and G. Jurman, "THE COEFFICIENT OF DETERMINATION R-SQUARED IS MORE INFORMATIVE THAN SMAPE, MAE, MAPE, MSE AND RMSE IN REGRESSION ANALYSIS EVALUATION," *PeerJ Computer Science*, vol. 7, pp. e623, 2021, doi: https://doi.org/10.7717/peerj-cs.623.

[41] S. D. Wahyuni and R. H. Kusumodestoni, "OPTIMALISASI ALGORITMA SUPPORT VECTOR MACHINE (SVM) DALAM KLASIFIKASI KEJADIAN DATA STUNTING," *Bulletin of Information Technology (BIT)*, vol. 5, no. 2, pp. 56–64, Jun. 2024, doi: https://doi.org/10.47065/bit.v5i2.1247.