International Journal of Robotics and Control Systems
Vol. 6, No. 1, 2026, pp. 36-53
ISSN 2775-2658
http://pubs2.ascee.org/index.php/ijrcs

Optimizing K-Nearest Neighbor Using Grey Wolf Optimizer for Breast Cancer Classification

Florentina Yuni Arini [a,1,*], Abas Setiawan [a,2], Shona Chayy Bilqisth [a,3], Khamron Sunat [b,4], Poomin Duankhan [b,5]

[a] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Sekaran, Gunungpati, Semarang, Indonesia 50229
[b] Department of Computer Science, College of Computing, Khon Kaen University, Khon Kaen, Thailand 40002
[1] floyuna@mail.unnes.ac.id; [2] abas.setiawan@mail.unnes.ac.id; [3] shona@mail.unnes.ac.id; [4] skhamron@kku.ac.th; [5] mini@kkumail.com
* Corresponding Author

ARTICLE INFO

Article history
Received October 04, 2025
Revised December 07, 2025
Accepted December 12, 2025

Keywords
K-Nearest Neighbor;
Grey Wolf Optimizer;
Principal Component Analysis;
Hyperparameter Tuning;
Breast Cancer Classification

ABSTRACT

Breast cancer continues to be a significant global health challenge, emphasizing the need for precise methods that support early detection. This research introduces an enhanced classification framework that integrates the K-Nearest Neighbors (KNN) algorithm with the Grey Wolf Optimizer (GWO). In this approach, GWO autonomously identifies the most informative features and determines the optimal KNN parameter settings, contributing to improved model performance. The Wisconsin Diagnostic Breast Cancer dataset was utilized, and an initial exploratory analysis was conducted to better understand feature patterns and class distributions. To examine the benefit of optimization, the proposed KNN-GWO model was compared with a Principal Component Analysis (PCA) based KNN model that reduces data dimensionality. Experimental findings show that the KNN-GWO approach achieved an accuracy of 97.07%, surpassing the KNN-PCA model's accuracy of 95.47%. The optimized model also delivered higher sensitivity and reduced false-positive predictions, both of which are crucial for clinical assessment. These results demonstrate that GWO effectively strengthens the performance of KNN while preserving the model's interpretability and computational simplicity. Overall, this research highlights the promise of optimization-enhanced KNN techniques as dependable and transparent tools for detecting breast cancer at an early stage.

1. Introduction

Breast cancer is one of the most common types of cancer experienced by women with an increasing incidence [1], [2]. More than 8 million people are diagnosed with cancer each year worldwide, and about 1 million of them are breast cancer cases [3], [4]. Based on data from the World Health Organization (WHO), it shows that approximately 1 in 8 to 10 women will experience breast cancer [5]. Although rarely diagnosed in younger women, especially under the age of 25–30 years, the risk increases sharply in the 40–49 age group where around 25% of breast cancer deaths are recorded [6]. Various factors contribute to the development of breast cancer including family history,

aging, hormonal factors, obesity, and postmenopausal hormonal imbalance [7]. Researchers show that more than 5%-6% of breast cancer patients have genetic mutations that are passed down in the family [8], [9]. This disease arises from the growth of abnormal cells which spread over time and attack the lymph nodes to form tumors that can be benign or malignant [10]. Benign present harmless symptom, however malignant may cause a risk of patient life [11]. This classification encompasses tumor effect of the host's health [12].

Today, the use of machine learning technology, especially nature inspired optimization (NIO) algorithms, are utilized to enhance the performance of various cases [13]–[17]. Thus, NIO algorithms also assist to improve the accuracy of breast cancer detection, providing more precise and reliable results [18]–[20]. There has been much research on the classification of breast cancer with various machine-learning models [21]–[23]. Assegie [24] reveals that the percentage accuracy of the KNN algorithm on the breast cancer detection test dataset is 94.35 and KNN with hyper-parameter produces 90.10. In 2025, Nataliani *et al* [25] achieved accuracy of 93.7% in multi-objective KNN. Meanwhile, Rizki *et al* [26] employing Euclidean (73,61%) and Manhattan Distance (75%) measurement for optimizing KNN hyperparameters with PSO. PSO is one of nature-inspired optimization algorithm [27].

Many nature-inspired optimization algorithms succeed in enhancing classification cases for KNN. Anvari et al [28] demonstrates the effectiveness of KNN hyper parameter with the whale optimization algorithm. While Ye *et al* [29] proposed a prediction model using Harris Hawks optimization to optimize Fuzzy KNN. Nejad *et al* [30] produce promising performance salp swarm optimization elevating KNN.

Moreover, many researchers utilize the grey wolf optimizer (GWO) to enrich various classification purposes in feature selection [31]–[34]. Therefore, Prokop [35] combine the Grey wolf optimizer with KNN to solve the clustering issue. Aswani *et al* [36] have successful in calculating outliers based on a given threshold using modified grey wolf optimization and KNN. Rajammal *et al* [37] improve the classification performance of Binary Improved Grey Wolf Optimizer (BIGWO) to optimize Adaptive kNN (AkNN) to categorize Parkinson's disease. Tahraoui *et al* [38] developing KNN using an improved grey wolf optimization algorithm to predict water quality.

While various methods such as Particle Swarm Optimization (PSO) and Harris Hawks Optimization (HHO) have been applied to feature selection, GWO is selected for this study due to its superior balance between exploration and exploitation phases and its requirement for fewer control parameters compared to PSO [39]. A critical gap in the current literature is the lack of direct comparison between nature-inspired wrapper methods (such as GWO) and statistical dimensionality reduction (such as PCA) within the specific context of the Wisconsin Breast Cancer dataset [40], [41].

It is also important to note that the Wisconsin Breast Cancer dataset exhibits a class imbalance. This imbalance could potentially affect evaluation metrics such as accuracy, as a model can achieve high overall accuracy by favoring the majority class while performing poorly on the minority class [42]. PCA efficiently reduces data dimensionality, transforms original features into principal components, which can obscure clinically meaningful information and potentially underrepresent patterns associated with the minority class [43]. In contrast, GWO preserves the original features, directly selecting the most informative attributes directly [44]. This allows the classifier to retain clinically relevant information and better distinguish malignant cases, making GWO particularly suitable for imbalanced datasets where accurate detection of the minority class is critical [45].

Therefore, we proposed the performance enhancement of KNN and GWO for breast cancer classification based on the feature selection approach to select the fittest feature and to produce the recommended accuracy. For comparison, in this research we also examine the selection of features using principal component analysis (PCA) [46], [47] to optimize the cooperative performance KNN and GWO on the breast cancer dataset by reducing the dimension of the data. The research contributions: (a) the role of GWO is to successfully control and optimize the KNN hyperparameter to select the optimal number of neighbors (K); and (b) enhanced KNN combined with GWO achieves

higher diagnostic accuracy for breast cancer classification. The manuscript structure present in the following section is Introduction, Method, Research Result and Discussion, and Conclusions.

2. Research Methods

This research was carried out in several stages to build an optimal classification model for breast cancer detection. Starting with the selection of a dataset that contains various relevant attributes, each stage of the investigation was arranged in stages to process the data, optimize the parameters, and evaluate the resulting classification model. The flowchart of the research can be seen in Fig. 1.

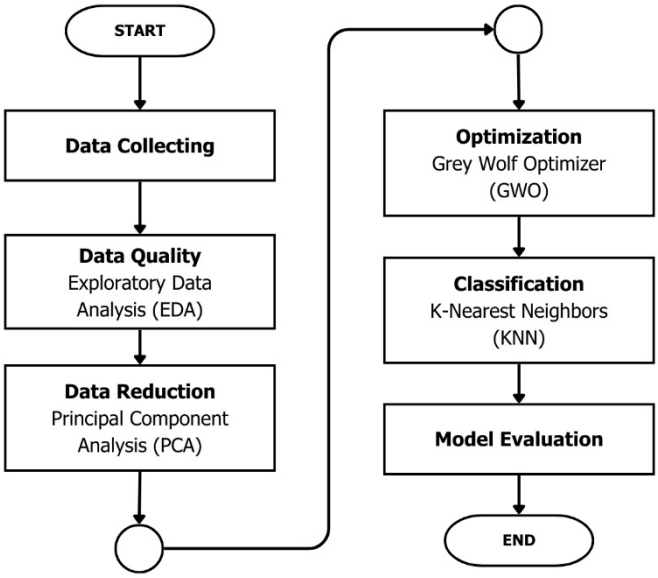


Fig. 1. Research method flowchart

Based on Fig. 1, first collect the data from "Breast Cancer Dataset" and then analyze the feature of the dataset. Second, prepare for the preprocessing data, then continue for data exploration. Based on data exploration, employ PCA for data reduction. In phase optimization, find the best k for the KNN parameter.

The dataset used in this research is publicly accessible through the Kaggle Wisconsin Breast Cancer Dataset (https://www.kaggle.com/datasets/moezkloula/breast-cancer-wisconsin-diagnostic-data-set) and has been utilized by [48]–[50]. Based on this dataset, the research methodology is presented in several sections. Section 2.1 Exploratory Data Analysis (EDA) dataset. Section 2.2 captures the most vital information with Principal Component Analysis (PCA) to reduce the dimensionality of the dataset. Section 2.3 demonstrates how K-Nearest Neighbors (KNN) classifies the dataset. Finally, Section 2.4 presents the Grey Wolf Optimizer (GWO) algorithm and its role in selecting the best features and optimizing the neighbors for the K-Nearest Neighbors (KNN).

2.1. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) [51] is the initial step in data analysis that focuses on exploring, understanding, and cleaning data before proceeding to modeling or advanced statistical analysis. This process aims to identify patterns, analyze the characteristics of the data, find outliers, and understand the relationships between variables. Thus, stages of analysis are conducted. First, column distribution facilitates selection producing the fittest modeling technique. Second, the correlation features dataset indicates similarities in the information provided by the features to avoid multicollinearity so that the model is not confused by variables that have similar information. Third, clustered heatmap numeric features are useful for analyzing redundant variables and providing guidance for more efficient feature selection.

2.2. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) [52] is defined as a measurable strategy utilized to diminish the dimensionality of information while maintaining as much change as possible. At this stage, PCA is connected to converting the first highlights into a smaller number of vital components, often two or three, to encourage examination and visualization. The first step in PCA is to calculate the covariance matrix [53], which is stated in (1).

$$Cov(X_i, X_j) = \frac{1}{n-1}\sum_{z=1}^{n}(X_{i,z} - \mu_i)(X_{j,z} - \mu_j), \tag{1}$$

where X is a data matrix of size $m \times n$, where m is the number of observations and n is the number of features where n number of points, $X_{i,z}$ and $X_{j,z}$ present data point variables i-th and j-th, μ_i and μ_j present data point variables i-th and j-th. The divider $n-1$ ensures an unbiased estimate of the covariance [54], [55]. The process begins by calculating the average for each feature. After the average is found, each value in the feature is subtracted from its average, thus obtaining standardized data. This process is done to ensure that all features have the same scale. Then, calculate the characteristic by finding the eigenvalue using (2) and eigenvectors [56] for matrix X, where X square matrix and I identity matrix which assist matrix X by λ (eigenvalue).

$$\det(X - \lambda I) = 0. \tag{2}$$

To obtain the eigenvalues λ, calculate the determinant of the matrix $X - \lambda I$, where I is the identity matrix. This process produces a characteristic polynomial that can be solved to find the eigenvalues. Once the eigenvalues are obtained, substitute them into (2) to obtain the eigenvectors V. These eigenvectors indicate the direction in which the data has the maximum variance, which is vital information in data analysis. Furthermore, to determine the number of principal components to be selected [21], the criteria used are as in (3).

$$\sum_{j=1}^{k}\lambda_j \geq 95\% \tag{3}$$

The criterion in (3) ensures that the total variance is extracted from the principal components 39 sat least 95% of the total variance in the dataset. The value of 95% is used as the threshold for total variance because it captures most of the information in the data while significantly reducing the dimensionality. The choice of 95% is also a widely accepted standard of practice because it ensures the efficiency of the analysis without losing too much valuable information. This process is followed by the formation of an eigenvector matrix, where the eigenvectors associated with the largest eigenvalues are arranged in a matrix V with K rows (number of principal components chosen) and P columns (number of original features), indicating a new direction in the data space that shows maximum variability [22]. After dimensionality reduction, the matrix is calculated by the formula in (4).

$$Y = VX \tag{4}$$

where V is a matrix containing the selected eigenvectors. This process involves multiplying the matrix V by the original data X that has been transposed. The result is a matrix Y that stores the data in a lower dimension, where each feature in Y is a linear combination of the features in X. With this transformation, data analysis can be done more easily, and visualizations can be made clearer without losing valuable information.

2.3. K-Nearest Neighbor (KNN)

The K-Nearest Neighbors (KNN) algorithm determines classify unlabeled data observations by positioning those data to the most similar labeled class sampled [57]. In KNN, the classification

decision is based on the majority of the classes of the k nearest neighbors [58] present in Algorithm 1. The class assigned to a test sample is determined by a majority vote of its k nearest neighbors, where the class with the most members among the k neighbors will be the predicted class. The distance calculation used to calculate proximity in KNN is the straight-line distance, commonly known as the Euclidean distance [59] as demote in (5).

$$d(x,X) = \sqrt{\sum_{i=1}^{n}(x_i - X_i)^2} \tag{5}$$

The formula for calculating the distance between each test sample x and training data point X with elements $x = (x_1, x_2, \ldots, x_n)$ and $X_i = (X_1, X_2, \ldots, X_n)$ i-th training data point present in $d(x,X)$.

Algorithm 1. K-Nearest Neighbor

Input:
- Dataset with n data points (x_1, x_2, ..., x_n)
- Label classes (C_1, C_2, ..., C_n)
- x: data point to be classified
- k: number of nearest neighbors

Output:
- Class of data point x

Begin
1. calculate the Euclidean distance $d(x,X)$
2. calculate n Euclidean distances.
3. let d be the sorted list of distances.
4. find the corresponding k data point for these k distances let $\{x_1, x_2, \ldots, x_k\}$ be the k nearest point.
5. initialize a frequency counter for each class:
   let class_count = $\{C_1: 0, C_2: 0, \ldots, C_m: 0\}$ //where m is the number of classes
6. **for** each of the k nearest **do**
   increment the count for the class of point x_j
   end for
7. find the class with the highest frequency.
8. return the class of data point x as the classification result.

End

2.4. Grey Wolf Optimizer (GWO)

Grey Wolf Optimizer (GWO) algorithm inspired by the hunting behavior of gray wolves in nature [60]. GWO is employed to determine the most suitable k value for the K-Nearest Neighbors (KNN) algorithm, thereby improving its classification performance. Selecting the optimal k is critical: a small k might cause the model overly sensitive to noise, while a large k is likely to produce smooth out important class distinctions, reducing accuracy [61].

In the Grey Wolf Optimizer (GWO), each candidate solution is presented as a "wolf." Each wolf represents a potential value of the nearest neighbor k for KNN within a predefined range of $[k_{min}, k_{max}]$. This mapping allows the optimization process to treat each possible k value as a position in the search space. During the iterations of GWO, the wolves adjust their positions, exploring different k values and gradually moving toward the one that maximizes KNN's classification accuracy or minimizes its error.

The wolves are guided by the positions of the four best solutions. First, α (Alpha) presents the leader wolf with the best solution, which is closest to the optimal solution. β (Beta) is the second-level wolf with a better solution than the one below it, but not as good as α. δ (Delta) shows the third-level wolf with a helpful solution, but lower than β. Last, ω (Omega) presents the lowest level that functions to support wolves α, β, and δ in surrounding and hunting prey (optimal solution). The wolves at level

ω are tasked with helping wolves α, β, and δ surround and capture prey (optimal solution) following the mathematical formulation in (6) and (7).

$$d = \left|b \times X_{target} - X_c\right|, \tag{6}$$

$$X_{new} = X_{target} - (Z * d), \tag{7}$$

where X_{target} is the position of the prey at the iteration, and X_c and X_{new} denote as current and new positions. Distance between the wolf and the prey present in (6) where coefficient b controls the wolf moves toward the prey. X_{new} influence by distance d and coefficient Z approaching X_{target}.

The coefficient Z decreases with each iteration. Then, the wolf's position is updated towards the prey's position with a distance calculated as in (8) and (9).

$$Z = (2a * rand_1) - a, \tag{8}$$

$$b = 2 * rand_2. \tag{9}$$

Where Z and b are coefficients to present the move of grey wolves approaching their prey. These formulas guide the approach of the fittest of exploration and exploitation. Thus, the updated position of the wolves is presented in (10), (11), and (12).

$$d_\alpha = |b_1 \times X_\alpha - X_c|, \tag{10}$$

$$d_\beta = \left|b_2 \times X_\beta - X_c\right|, \tag{11}$$

$$d_\delta = |b_3 \times X_\delta - X_c|, \tag{12}$$

Moreover, those equations present distance calculation between the wolf position ω and the wolf positions α, β, and δ. Meanwhile, the updated wolves based on the previous distance exhibit in (13), (14), and (15).

$$X_{\alpha n} = X_\alpha - a_1 \times d_\alpha, \tag{13}$$

$$X_{\beta n} = X_\beta - a_2 \times d_\beta, \tag{14}$$

$$X_{\delta n} = X_\delta - a_2 \times d_\delta, \tag{15}$$

$$X_{np} = \frac{X_{\alpha n} + X_{\beta n} + X_{\delta n}}{3}, \tag{16}$$

where $X_{\alpha n}$, $X_{\beta n}$, and $X_{\delta n}$ present update wolf position, while X_{np} integrating information from the three main positions, thus increasing the efficiency of the solution search based on $X_{\alpha n}$, $X_{\beta n}$, and $X_{\delta n}$ by 3 represent in (16).

3. Results and Discussion

The result and discussion of the K-Nearest Neighbor (KNN) classifier for breast cancer detection using the Grey Wolf Optimizer (GWO) are described in the following section. Section 3.1 Wisconsin Breast Cancer Dataset Analysis. Section 3.2 Experiment Discussion.

3.1. Analysis of Wisconsin Breast Cancer Dataset

Wisconsin Breast Cancer Dataset consists of 30 numerical features measured using standard error (SE) and worst case (worst value). This Wisconsin dataset [62], [63] is declared clean and usable because there are no missing values. Furthermore, detail feature description in each column present as follows *id* (unique identification for each tumor sample), *diagnosis* (tumor classification), *M*

(malignant), *B* (benign), *radius_mean* (tumor average radius), *texture_mean* (tumor average texture), *perimeter_mean* (tumor average perimeter), *area_mean* (tumor average area), *smoothness_mean* (average of tumor surface smoothness), *compactness_mean* (tumor average density), *concavity_mean* (average depth of concavity in the tumor), *concave points_mean* (average number of concave points in the tumor), *symmetry_mean* (tumor average symmetry), and *fractal_dimension_mean* (tumor average fractal dimension).

The stage of EDA is essential to guide the preprocessing steps and to select meaningful features for further analysis. The EDA process identifies relevant features and uncovers the characteristics, structure, and patterns of the data as an initial step in the analysis. The EDA data distribution shown in Fig. 2 assists data examination by recognizing distribution pattern for each feature in the Wisconsin Breast Cancer Dataset.

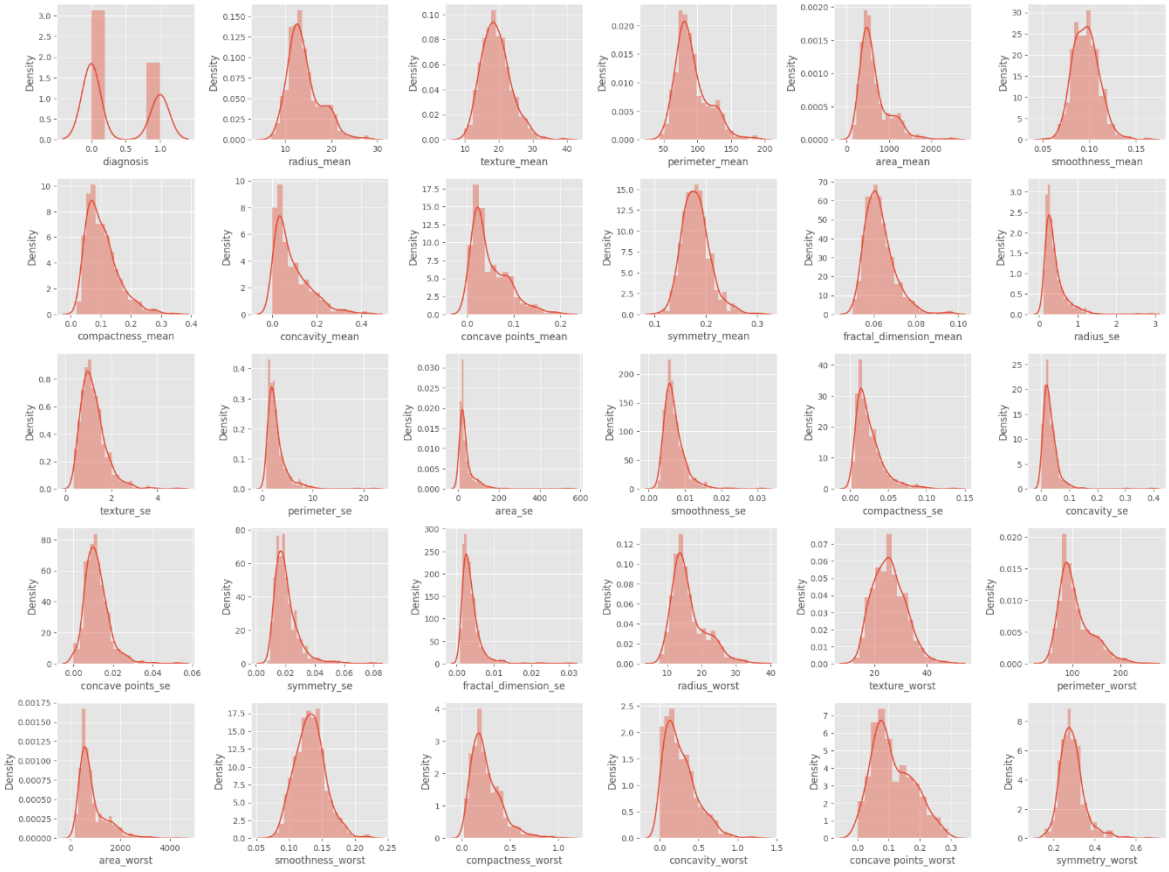


Fig. 2. Data distribution in the EDA stage

Each plot combines a histogram (bar chart) that shows frequency counts and a smooth curve indicating data density. The height of the bars represents the frequency of specific values, while the curve highlights areas where data clusters or spreads. Peaks in the curve indicate higher data concentrations in specific regions.

By combining histograms with smooth density curves, the figure makes it easier to observe the general shape and behavior of the data. Most features exhibit a noticeable right-skewed pattern, where lower values appear frequently, while higher values are less common. These patterns suggest that most cells are relatively small or within normal size ranges, while a smaller portion displays unusually large measurements, which are often linked to malignant cases. Standard error (SE) features show sharp peaks near zero, meaning that most measurements exhibit low variability, although some malignant cases present higher instability. The "worst" features, representing the most abnormal cell regions, tend to have wider and more spread-out curves, capturing the extreme values frequently observed in cancerous tissue. Overall, these density plots highlight the features that are stable, those that are more

variable, and those that provide the strongest signals to distinguish between benign and malignant samples.

The comparison bar chart of Malignant (M) and Benign (B) Wisconsin Breast Cancer Dataset bar chart in Fig. 3 presents values around 200 (37.3%) and 350 (62.7%), respectively. The red bars represent malignant cases, which refer to tumors with the potential to spread to other parts of the body. In contrast, the blue bars indicate benign cases, representing tumors that are generally less harmful and unlikely to spread. Then, the dataset splits into two main components: features (X) and labels (y). The features (X) include all independent variable columns that describe cancer characteristics, while the label (y) corresponds to the diagnosis column, serving as the dependent variable for prediction. Subsequently, the data is split into training and testing sets using the *train_test_split* method. The split 80% of the data is allocated for training and 20% for testing based on [64], [65].

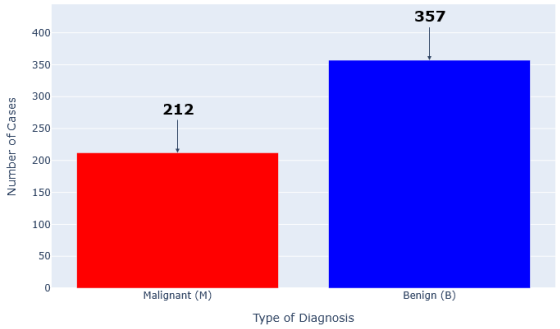


Fig. 3. Number of malignant and benign cancer diagnoses

Then, the correlation matrix in Fig. 4 shows that many tumor features are very closely related to each other. Measurements that describe tumor size such as radius, perimeter, and area, for both average and worst cases are almost perfectly related ($r \approx 0.94$–0.99). This makes sense because if the tumor gets bigger in one way, the other size measurements naturally increase too. The features that describe how irregular or bumpy edges of the tumor are like concave points and concavity also show strong connections ($r > 0.85$). All these features all capture similar information on how uneven the tumor's surface is, which is often seen in more aggressive cancers. The diagnosis variable is most strongly related to both the size and irregularity features. This means that malignant tumors tend to be larger and have more uneven edges. Among these, the worst of the concave points is one of the best indicators ($r \approx 0.79$). In general, the heatmap suggests that tumor growth and aggressiveness can be identified through a group of strongly linked size and shape features.

3.2. Experiment Discussion

During data processing preparation, the contribution data for each column is determined specifically. First, eliminate irrelevant columns to achieve an efficient and accurate model. Second, the diagnosis column is transformed into a numeric format, where benign cases are represented as "0" and malignant cases as "1". The experiment compared three classification strategies: (a) GWO-optimized KNN, (b) KNN with PCA, and (c) a hybrid KNN-GWO-PCA model. Euclidean distance was used for robust measurement, as dataset features are continuous physical variables. GWO feature selection employed 12 wolves over 35 iterations to balance search power and efficiency. Odd K-values (3–13) helped avoid voting ties. The models were assessed using confusion matrices, precision, recall, and F1-scores for a comprehensive evaluation.

3.2.1. Encoding and Fitness Function

Each candidate solution (wolf) in the GWO population is encoded as a real-valued vector X, representing the potential value of k. The continuous value is discretized denoted in (17).

$$k = round(X_i) \;\; s.t. \;\; k \in \{odd\ integers\} \tag{17}$$

The fitness function $f(X)$ aims to minimize the classification error rate during cross-validation. It is defined in (18).

$$f(X) = 1 - \text{Accuracy}_{CV} \tag{18}$$

where Accuracy_{CV} is the mean accuracy obtained from 5-fold Stratified Cross-Validation. This ensures that the selected k generalizes well to unseen data segments.

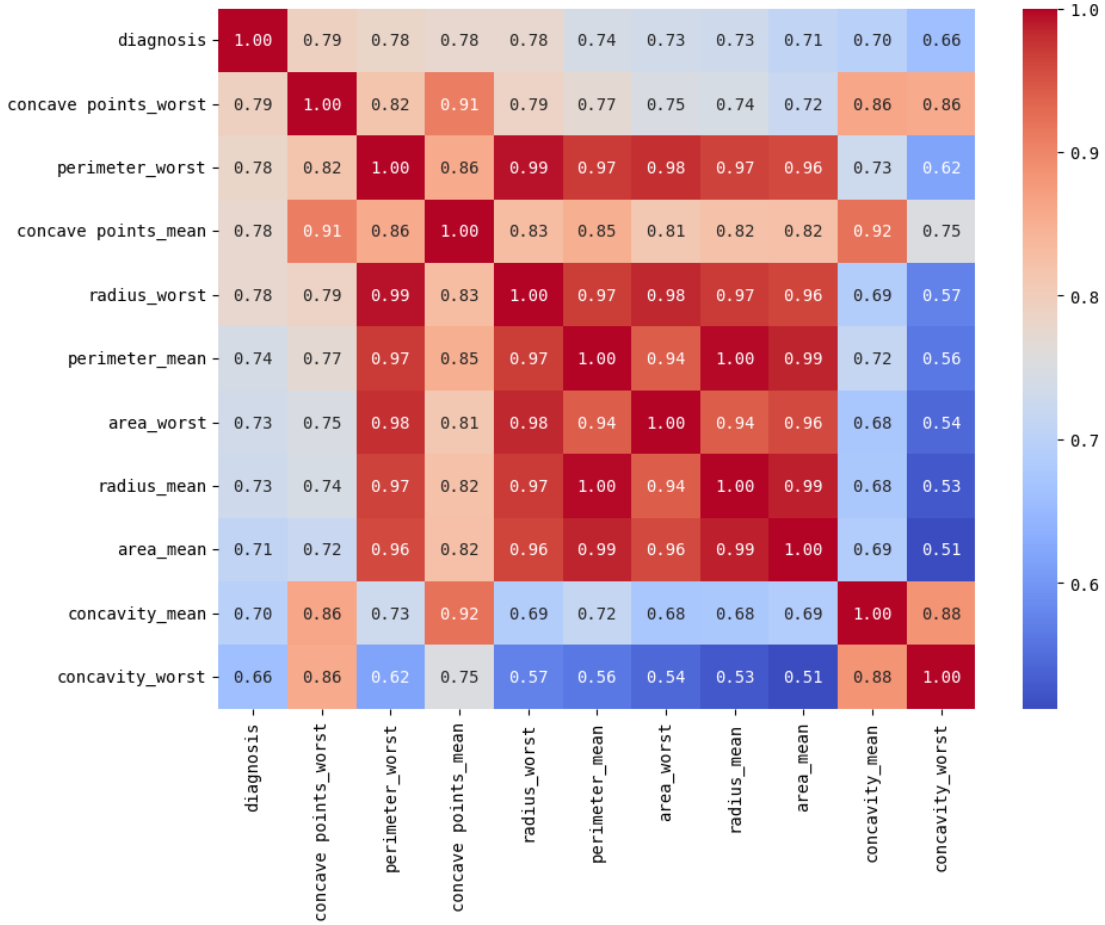


Fig. 4. Top ten correlated features based on dataset

3.2.2. Determination Parameter *K* values of KNN

According to Chen [66], selecting the right K value that reflect the number of neighbors is essential to make a prediction, balance bias and variance, steer clear of typical, challenges, and maintain strong performance. To explore the impact of the K value on the accuracy of the model, an evaluation was conducted using various k values 3, 5, 7, 9, 11, and 13 [67]. As the value of K increases [68], the decision boundary becomes smoother because more neighbors are considered, which can help reduce the effect of noise on the data. However, the relationship between K and "distance" is more about how many data points influence the prediction, rather than the distance from its current position. The fittest K value controls the optimal number of neighbors that would improve the predictive power of KNN enhancement model as shown in Table 1.

Table 1. Accuracy experiment results k values

KNN Enhancement	*K* values					
	3	**5**	**7**	**9**	**11**	**13**
KNN and GWO	**0.970760**	**0.970760**	**0.970760**	0.964912	0.964912	0.959064
KNN and PCA	**0.959064**	**0.959064**	0.953216	0,953216	0.947368	0.947368
KNN, GWO and PCA	**0.959064**	**0.959064**	0.953216	0,953216	0.947368	0.947368

Table 1 presents the accuracy results of the K-Nearest Neighbors (KNN) classification enhanced with different optimization or dimensionality-reduction techniques, measured across a range of odd K values (3, 5, 7, 9, 11, 13). The three enhancement scenarios are: KNN and GWO (Grey Wolf Optimizer), KNN and PCA (Principal Component Analysis), and a combined KNN, GWO, and PCA approach. Examining the numerical patterns in this matrix reveals several clear trends and subtle differences that can guide interpretation. Across all methods, accuracy is highest at the smallest K values and gradually decreases as K increases.

For KNN and GWO**,** the accuracy remains at a strong 0.970760 for $K = 3, 5, 7$, then lower slightly to 0.964912 at $K = 9$ and $K = 11$ and decreases again to 0.959064 at $K = 13$. This shows a plateau of excellent performance up to $K = 7$, followed by a gentle downward slope. The small stepwise decline indicates that adding more neighbors introduces a bit more variability or noise, which is consistent with the general behavior of KNN: a higher K can reduce variance but may increase bias. The KNN and PCA configuration starts slightly lower than the GWO-enhanced version, achieving 0.959064 at $K = 3$ and 5, dipping modestly to 0.953216 at $K = 7$ and 9, and then falling further to 0.947368 for $K = 11$ and 13. This is a steady monotonic decline, suggesting that PCA dimensionality reduction alone cannot maintain top performance as neighborhood size grows. The combined KNN, GWO and PCA method mirrors the PCA-only pattern almost exactly: 0.959064 at $K = 3$ and 5, 0.953216 at $K = 7$ and 9, and 0.947368 at $K = 11$ and 13. This indicates that adding GWO optimization after PCA dimensionality reduction does not yield additional accuracy gains. This outcome contradicts the initial hypothesis of hybridization but offers a critical insight: PCA effectively aggregates the dataset's variance into the principal components, thereby removing the redundant noise that GWO would typically filter out. Consequently, applying GWO on top of PCA-transformed data introduces computational overhead without contributing to feature discriminability. This finding underscores that GWO is most effective when applied to raw features (preservation of physical meaning) rather than the latent, pre-compressed feature space created by PCA.

3.2.3. Measurement Model Using Confusion Matrix

A confusion matrix [69] is an essential statistical method for assessing the effectiveness of a classification model by comparing its predicted labels with the actual labels in a dataset. This method is still preferable for measurement performance [70], [71]. It is usually displayed as a square matrix in which the rows correspond to the real classes and the columns represent the predicted classes, providing a clear picture of how accurately the model separates distinct categories, as shown in Fig. 5. In a binary classification setting, the matrix is made up of four main elements: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). True Positives represent cases where the model successfully identifies positive instances, while True Negatives refer to correctly recognized negative instances. False Positives often called Type I errors that occur when the model mistakenly predicts a positive outcome, whereas False Negatives, known as Type II errors, arise when the model fails to detect a positive case.

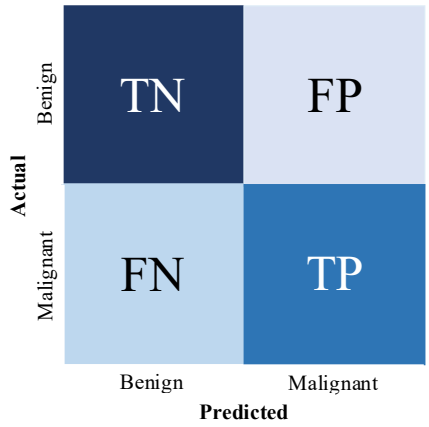
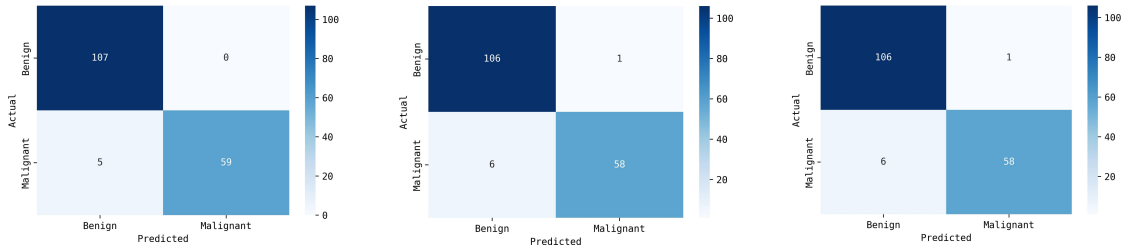


Fig. 5. Statistical method confusion matrix

Detail description Confusion Matrix shown in Fig. 6 is exhibit in Table 2. The KNN enhancement of the model with GWO which produces optimum *K* values at 3, 5, and 7 shows the predicted malignant number for 59 samples and the number of samples truly identified as malignant tumors according to the predicted malignant and actual malignant (TP). Moreover, based on predicted benign and actual benign (TN) present 107 samples, which means that the model correctly identified 107 samples as Benign when they were Benign. Meanwhile, FP represent cases where the model incorrectly predicted a result as positive (malignant) when it was negative (benign). KNN and GWO result in zero (0) false positives, which means the model did not incorrectly classify any benign cases as malignant. FN of KNN and GWO shows cases where the model incorrectly predicted a result as negative (benign) when it was positive (malignant) with 5 false negatives, meaning the model failed to identify 5 malignant cases and classified them as benign. However, optimum *K* values for the model of KNN and PCA and the model of KNN, GWO and PCA produce the same values at 3 and 5. Additionally, the numbers of values TP, TN, FP and FN for those two models generate 58, 106, 1 and 6, respectively. Thus, KNN model enhanced with GWO indicates accurate malignant detection with only missing five malignant cases.



(a) *K*=3,5,7 (b) *K*=3,5 (c) *K*=3,5

Fig. 6. Confusion matrix: (a) KNN and GWO; (b) KNN and PCA; (c) KNN, GWO and PCA

Table 2. Detail description confusion matrix enhancement of KNN

KNN enhancement	*K* Values	True Positives (TP)	True Negatives (TN)	False Positives (FP)	False Negatives (FN)
KNN and GWO	3, 5, 7	59	107	0	5
KNN and PCA	3, 5	58	106	1	6
KNN, GWO, and PCA	3, 5	58	106	1	6

3.2.4. Model Evaluation Matrix: Precision Recall, F1-Score

The evaluation of machine learning models using precision, recall, and the F-measure have been widely applied [72], especially when dealing with imbalanced datasets where overall accuracy may give a distorted view of performance. Precision represents the ratio of correctly identified positive cases to the total number of positive predictions, while recall reflects the percentage of actual positive instances accurately recognized by the classifier. The F-measure serves as the harmonic means of precision and recall, providing a single metric that balances the trade-off between these two complementary indicators. Table 3 presents the performance metrics for the KNN and GWO model. The performance of the model is measured across on different metrics for two classes: class 0 (benign) and class 1 (malignant). These metrics include precision, recall, F1-score, and support, with an overall accuracy of 97.77%. Based on precision values, the model demonstrates impressive performance across both classes, with class 0 achieving a precision of 0.9554 and class 1 attaining a perfect precision of 1. This suggests that the model is particularly accurate in predicting class 1, with no false positives. For recall, the model performs slightly better for class 1 with a value of 0.9219, indicating it correctly identifies about 92% of all malignant instances. Class 0 achieves perfect recall at 1, meaning it correctly classifies all benign instances. In terms of the F1-score, which is the harmonic mean of precision and recall, the results are highly favorable. Class 1 achieves an F1-score of 0.9594, while class 0 reaches 0.9772, reinforcing the strength of the model in accurately distinguishing between the two classes. The macro average, which gives equal weight to both classes, is 0.9708,

highlighting the balanced performance of the model across both benign and malignant cases. Similarly, the weighted average F1-score is 0.9772, which is close to the F1-score for class 0, again showing that the model is well-optimized for the dataset. These results demonstrate a high performance of KNN and GWO in high accuracy, precision, recall, and F1-score values. Therefore, KNN and GWO is well-suited for detecting malignant cases while minimizing errors, especially false positives determined clearly in Table 3.

Table 3. Performance model KNN and GWO

	Precision	Recall	F1-Score	Support
Class 0 (Benign)	0.955357	1.000000	0.977169	107
Class 1 (Malignant)	1.000000	0.921875	0.959350	64
Accuracy	0.977679	0.960938	0.968259	171
Macro avg	0.972066	0.970760	0.970500	171
Weighted avg	0.955357	1.000000	0.977169	107

Meanwhile, Table 4 exhibits the performance model of KNN and PCA and model of KNN, GWO and PCA. Based on the experiment results, these two models produce equal values of precision, recall, F1-score, and accuracy. For class 0, the two models in Table 4 achieve a precision of 0.9464, which is slightly lower than the KNN and GWO model's precision of 0.9554. This indicates that although the two are highly effective in classifying benign cases, they have a slightly higher false positive rate compared to the KNN and GWO model. The recall class 0 of KNN-GWO produces it perfect recall, catching every true positive without missing any. It is desirable to see situations where missing positive cases are unacceptable. For Class 1 in Table 4 exhibits a precision of 0.9831, which is a bit lower than the KNN and GWO model's perfect precision of 1. However, their recall for Class 1 0.9063, which is lower than KNN-GWO model (0.9219), indicating that KNN and GWO is slightly better at identifying malignant cases, even though it is not perfect in avoiding false negatives. The F1-score for Class 1 in Table 3 and Table 4 show quite close with difference values 0.016261, indicating that the KNN and GWO model still performs marginally better in balancing precision and recall for malignant cases. Moreover, the accuracy of the model in Table 4 is 0.9647, which indicates a small decrease compared to the KNN and GWO model with values 0.9777. Overall, all three models perform excellently. However, KNN-GWO model slightly outperforms compared to KNN and PCA model and KNN, GWO, and PCA in terms of precision for both classes and overall accuracy. Additionally, the performance of the three models is summarized and visualized in Fig. 7.

Table 4. Performance model of KNN and PCA and model of KNN, GWO and PCA

	Precision	Recall	F1-Score	Support
Class 0 (Benign)	0.946429	0.990654	0.968037	107
Class 1 (Malignant)	0.983051	0.906250	0.943089	64
Accuracy	0.964740	0.948452	0.955563	171
Macro avg	0.960135	0.959064	0.958700	171
Weighted avg	0.946429	0.990654	0.968037	107

3.2.5. Comparison with State-of-the-Art

To validate the proposed KNN-GWO model, we compared our optimal result (97.07%) against recent state-of-the-art methods applied to the same Wisconsin Diagnostic Breast Cancer dataset. As detailed in Table 2, our KNN-GWO approach outperforms the standard KNN optimization found in Assegie (94.35%) [24] and the Multi-objective KNN by Nataliani et al. (93.7%) [25]. Furthermore, it significantly exceeds the PSO-optimized KNN reported by Rizki et al. (73.61%) [26]. While formal t-tests were not feasible due to the single-dataset protocol, the robustness of the KNN-GWO approach is evident in its consistency; it maintained superior accuracy across all tested k values (3 to 13) compared to the PCA and baseline models. It is acknowledged that some recent studies utilizing Deep Learning (DL) or ensemble Neural Networks have reported accuracies exceeding 98% on this dataset. However, such "black box" models often sacrifice feature traceability for marginal accuracy gains. In medical diagnostics, the ability to trace a decision back to specific physical attributes (e.g.,

concavity_worst) is often prioritized over raw accuracy. Our result of 97.07% represents a competitive optimal point that maintains clinical interpretability—a feature lost in PCA-based or Deep Learning approaches—while significantly outperforming standard geometric classifiers.

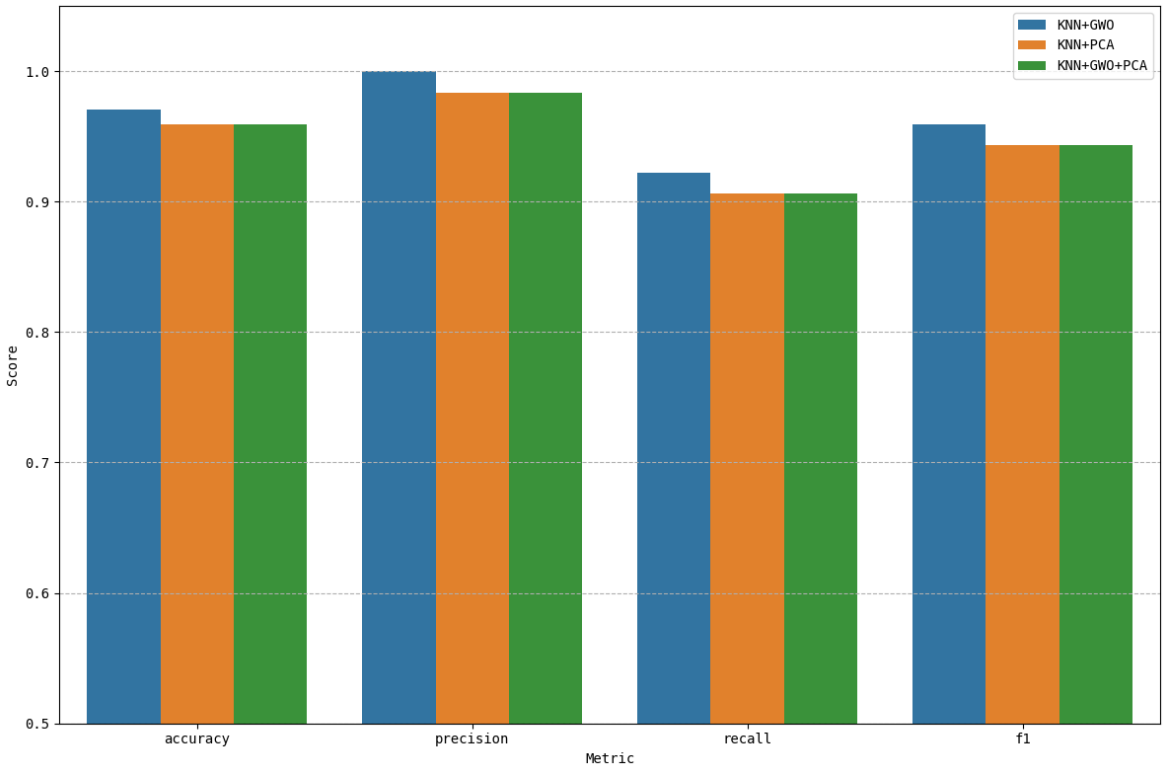


Fig. 7. Results of the metric experiment KNN and GWO, KNN and PCA, and KNN, GWO, and PCA

3.2.6. Feature Interpretability Analysis

A key advantage of GWO over PCA is clinical interpretability. Although PCA components are linear combinations of all features, GWO identified a specific subset of features as the most diagnostic. The algorithm converged on critical morphological features, specifically concave points_worst, concave points_mean, perimeter_worst, texture_worst, and radius_worst. The selection of these specific attributes suggests that they are the primary biological indicators of malignancy, providing valuable, traceable feedback for pathologists that PCA-transformed data cannot offer.

4. Conclusion

This research systematically evaluated the performance of the KNN classification under three optimization frameworks: (a) KNN enhanced with GWO, (b) KNN with PCA, and (c) a hybrid KNN-GWO-PCA model. Experimental results in varying K values (3, 5, 7, 9, 11, 13) demonstrate that the KNN-GWO wrapper approach produces superior stability and accuracy, achieving a peak performance of 0.9707 at K=3, 5, and 7. Unlike PCA-based models, which exhibited a steeper performance decline as neighborhood size increased, the GWO-optimized model effectively isolated diagnostically relevant features, maintaining robustness. Crucially, confusion matrix analysis confirmed that the KNN-GWO model offers the best safety profile for medical diagnosis, achieving zero false positives for optimal K values and outperforming comparative models in precision and recall for malignant cases (Class 1). However, it is acknowledged that this wrapper-based improvement comes with a higher computational cost during the training phase compared to the filter-based PCA approach, as the GWO fitness evaluation requires iterative model training. Future research will address this computational trade-off by exploring hybrid filter-wrapper techniques and benchmarking GWO against other emerging swarm intelligence algorithms, such as the Whale

Optimization Algorithm or Harris Hawks Optimization, to further validate these findings on larger, multi-center datasets.

Supplementary Materials: The following supporting information can be downloaded at: https://breastcancerml-u5p7fiedlzhqyyxjuhnjnt.streamlit.app/.

Author Contribution: All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding: This research was funded by DPA LPPM Universitas Negeri Semarang.

Acknowledgments: The authors would like to thank Advance Smart Computing College of Computing, Khon Kaen University, for supporting this research collaboration work.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] S. Nardin, E. Mora, F. M. Varughese, F. D'Avanzo, A. R. Vachanaram, V. Rossi, C. Saggia, S. Rubinelli, and A. Gennari, "Breast cancer survivorship, quality of life, and late toxicities," *Frontiers in Oncology*, vol. 10, p. 864, 2020, https://doi.org/10.3389/fonc.2020.00864.

[2] K. M. Cuthrell and N. Tzenios, "Breast cancer: Updated and deep insights," *International Research Journal of Oncology*, vol. 6, no. 1, pp. 104–118, 2023, https://journalirjo.com/index.php/IRJO/article/view/129.

[3] M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, J. Vignat, J. R. Gralow, F. Cardoso, S. Siesling, and I. Soerjomataram, "Current and future burden of breast cancer: Global statistics for 2020 and 2040," *The Breast*, vol. 66, pp. 15–23, 2022, https://doi.org/10.1016/j.breast.2022.08.010.

[4] J. Kim, A. Harper, V. McCormack, H. Sung, N. Houssami, E. Morgan, M. Mutebi, G. Garvey, I. Soerjomataram, and M. M. Fidler-Benaoudia, "Global patterns and trends in breast cancer incidence and mortality across 185 countries," *Nature Medicine*, vol. 31, no. 4, pp. 1154–1162, 2025, https://doi.org/10.1038/s41591-025-03502-3.

[5] P. H. Tan, I. Ellis, K. Allison, E. Brogi, S. B. Fox, S. Lakhani, A. J. Lazar, E. A. Morris, A. Sahin, R. Salgado, A. Sapino, H. Sasano, S. Schnitt, C. Sotiriou, P. van Diest, V. A. White, D. Lokuhetty, and I. A. Cree, "The 2019 World Health Organization classification of tumours of the breast," *Histopathology*, vol. 77, no. 2, pp. 181–185, 2020, https://doi.org/10.1111/his.14091.

[6] J. Shelton, E. Zotow, L. Smith, S. A. Johnson, C. S. Thomson, A. Ahmad, L. Murdock, D. Nagarwalla, and D. Forman, "25 year trends in cancer incidence and mortality among adults aged 35–69 years in the UK, 1993–2018: Retrospective secondary analysis," *BMJ*, vol. 384, p. e076962, 2024, https://doi.org/10.1136/bmj-2023-076962.

[7] E. I. Obeagu and G. U. Obeagu, "Breast cancer: A review of risk factors and diagnosis," *Medicine (United States)*, vol. 103, no. 3, p. E36905, 2024, https://doi.org/10.1097/MD.0000000000036905.

[8] C. Hu, S. N. Hart, R. Gnanaolivu, H. Huang, K. Y. Lee, J. Na, C. Gao, J. Lilyquist, S. Yadav, N. J. Boddicker, R. Samara, J. Klebba, C. B. Ambrosone, H. Anton-Culver, P. Auer, E. V. Bandera, et al., "A population-based study of genes previously implicated in breast cancer," *New England Journal of Medicine*, vol. 384, no. 5, pp. 440–451, 2021, https://doi.org/10.1056/NEJMoa2005936.

[9] C. Hu, E. C. Polley, S. Yadav, J. Lilyquist, H. Shimelis, J. Na, S. N. Hart, D. E. Goldgar, S. Shah, T. Pesaran, J. S. Dolinsky, H. LaDuca, and F. J. Couch, "The contribution of germline predisposition gene mutations to clinical subtypes of invasive breast cancer from a clinical genetic testing cohort," *JNCI: Journal of the National Cancer Institute*, vol. 112, no. 12, pp. 1231–1241, 2020, https://doi.org/10.1093/jnci/djaa023.

[10] L. J. Medeiros, D. P. O'Malley, N. P. Caraway, F. Vega, K. S. J. Elenitoba-Johnson, and M. S. Lim, *Tumors of the Lymph Nodes and Spleen*. American Registry of Pathology, 2017, https://doi.org/10.55418/9781933477381.

[11] A. Patel, "Benign vs malignant tumors," *JAMA Oncology*, vol. 6, no. 9, p. 1488, 2020, https://doi.org/10.1001/jamaoncol.2020.2592.

[12] G. B. Pierce and C. Wallace, "Differentiation of malignant to benign cells," *Cancer Research*, vol. 31, no. 2, pp. 127–134, 1971, https://aacrjournals.org/cancerres/article/31/2/127/478292/Differentiation-of-Malignant-to-Benign-Cells1.

[13] A. Hussein, S. Alomari, M. Almomani, R. Zitar, H. Migdady, A. Smerat, V. Snasel, and L. Abualigah, "A hybrid PSO-GCRA framework for optimizing control systems performance," *International Journal of Robotics and Control Systems*, vol. 5, no. 1, pp. 459–478, 2025, https://doi.org/10.31763/ijrcs.v5i1.1738.

[14] D. Lestari, S. Sendari, I. A. E. Zaeni, S. Arifin, and R. D. I. Sari, "Improving efficiency and effectiveness of wheeled mobile robot pathfinding in grid space using a genetic algorithm with dynamic crossover and mutation rates," *International Journal of Robotics and Control Systems*, vol. 5, no. 1, pp. 407–425, 2025, https://doi.org/10.31763/ijrcs.v5i1.1573.

[15] Y. Bhardwaj, O. A. Shah, and R. Kumar, "Multi-objective particle swarm optimization for enhancing chiller plant efficiency and energy savings," *International Journal of Robotics and Control Systems*, vol. 4, no. 3, pp. 1319–1336, 2024, https://doi.org/10.31763/ijrcs.v4i3.1501.

[16] A. El Romeh, V. Snášel, and S. Mirjalili, "Dholes-inspired optimization (DIO): a nature-inspired algorithm for engineering optimization problems," *Cluster Computing*, vol. 28, no. 13, p. 853, 2025, https://doi.org/10.1007/s10586-025-05543-2.

[17] N. A. Mansour, M. S. Saraya, and A. I. Saleh, "Groupers and moray eels (GME) optimization: A nature-inspired metaheuristic algorithm for solving complex engineering problems," *Neural Computing and Applications*, vol. 37, no. 1, pp. 63–90, 2025, https://doi.org/10.1007/s00521-024-10384-y.

[18] M. Ghorbian and S. Ghorbian, "Usefulness of machine learning and deep learning approaches in screening and early detection of breast cancer," *Heliyon*, vol. 9, no. 12, p. e22427, 2023, https://doi.org/10.1016/j.heliyon.2023.e22427.

[19] M. Tahmooresi, A. Afshar, B. B. Rad, K. B. Nowshath, and M. A. Bamiah, "Early Detection of Breast Cancer Using Machine Learning Techniques," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 10, no. 3-2, pp. 21–27, Jul. 2023, https://doi.org/10.47839/ijc.22.2.3093.

[20] X. Jia, X. Sun, and X. Zhang, "Breast Cancer Identification Using Machine Learning," *Mathematical Problems in Engineering*, vol. 2022, no. 1, p. 8122895, 2022, https://doi.org/10.1155/2022/8122895.

[21] R. R. Kadhim and M. Y. Kamil, "Comparison of machine learning models for breast cancer diagnosis," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 1, pp. 415–421, 2023, https://doi.org/10.11591/ijai.v12.i1.pp415-421.

[22] D. Houfani, S. Slatnia, O. Kazar, N. Zerhouni, H. Saouli, and I. Remadna, "Breast cancer classification using machine learning techniques: A comparative study," *Medical Technologies Journal*, vol. 4, no. 2, pp. 535–544, 2020, https://doi.org/10.26415/2572-004X-vol4iss2p535-544.

[23] A. Bokhare and P. Jha, "Machine learning models applied in analyzing breast cancer classification accuracy," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 3, pp. 1370–1377, 2023, https://doi.org/10.11591/ijai.v12.i3.pp1370-1377.

[24] T. A. Assegie, "An optimized k-nearest neighbor based breast cancer detection," *Journal of Robotics and Control*, vol. 2, no. 3, pp. 115–118, 2021, https://doi.org/10.18196/jrc.2363.

[25] Y. Nataliani, C. Arthur, T. Wellem, K. D. Hartomo, and N. H. A. Wahab, "Multi-objective k-nearest neighbor for breast cancer detection," *International Journal of Informatics and Visualization*, vol. 9, no. 1, pp. 241–247, 2025, https://doi.org/10.62527/joiv.9.1.2669.

[26] M. Rizki, A. Hermawan, and D. Avianto, "Optimization of hyperparameter K in K-nearest neighbor using particle swarm optimization," *JUITA Journal of Informatics*, vol. 12, no. 1, p. 71, 2024, https://doi.org/10.30595/juita.v12i1.20688.

[27] A. Khamparia, A. Khanna, N. G. Nguyen, and B. L. Nguyen, *Nature-inspired optimization algorithms: Recent advances in natural computing and biomedical applications*. Berlin, Boston: De Gruyter, 2021, https://doi.org/10.1515/9783110676112.

[28] S. Anvari, M. A. Azgomi, M. R. E. Dishabi, and M. Maheri, "Weighted K-nearest neighbors classification based on Whale optimization algorithm," *Iranian Journal of Fuzzy Systems*, vol. 20, no. 3, pp. 61–74, 2023, https://doi.org/10.22111/ijfs.2023.7639.

[29] H. Ye, P. Wu, T. Zhu, Z. Xiao, X. Zhang, L. Zheng, R. Zheng, Y. Sun, W. Zhou, Q. Fu, X. Ye, A. Chen, S. Zheng, A. A. Heidari, M. Wang, J. Zhu, H. Chen, and J. Li, "Diagnosing coronavirus disease 2019 (COVID-19): Efficient Harris hawks-inspired fuzzy K-nearest neighbor prediction methods," *IEEE Access*, vol. 9, pp. 17787–17802, 2021, https://doi.org/10.1109/ACCESS.2021.3052835.

[30] M. Behrouzian Nejad and M. E. Shiri, "A new enhanced learning approach to automatic image classification based on salp swarm algorithm," *Computer Systems Science and Engineering*, vol. 34, no. 2, pp. 91–100, 2019, https://doi.org/10.32604/csse.2019.34.091.

[31] Q. Al-Tashi, H. Md Rais, S. J. Abdulkadir, S. Mirjalili, and H. Alhussian, "A review of Grey Wolf Optimizer-based feature selection methods for classification," in *Evolutionary Machine Learning Techniques: Algorithms and Applications*, S. Mirjalili, H. Faris, and I. Aljarah, Eds. Singapore: Springer Singapore, 2020, pp. 273–286, https://doi.org/10.1007/978-981-32-9990-0_13.

[32] N. M. Sallam, A. I. Saleh, H. A. Arafat Ali, and M. M. Abdelsalam, "An efficient strategy for blood diseases detection based on Grey Wolf Optimization as feature selection and machine learning techniques," *Applied Sciences*, vol. 12, no. 21, p. 10760, 2022, https://doi.org/10.3390/app122110760.

[33] Q. Al-Tashi, S. J. Abdulkadir, H. M. Rais, S. Mirjalili, H. Alhussian, M. G. Ragab, and A. Alqushaibi, "Binary multi-objective grey wolf optimizer for feature selection in classification," *IEEE Access*, vol. 8, pp. 106247–106263, 2020, https://doi.org/10.1109/ACCESS.2020.3000040.

[34] C. Shen and K. Zhang, "Two-stage improved grey wolf optimization algorithm for feature selection on high-dimensional classification," *Complex & Intelligent Systems*, vol. 8, no. 4, pp. 2769–2789, 2022, https://doi.org/10.1007/s40747-021-00452-4.

[35] K. Prokop, "Grey wolf optimizer combined with k-nn algorithm for clustering problem," in *Proceedings of the 27th International Conference on Information Society and University Studies (IVUS 2022)*, Kaunas, Lithuania, 2022, pp. 14–19. CEUR Workshop Proceedings, vol. 3611, 2022, https://ceur-ws.org/Vol-3611/paper3.pdf.

[36] R. Aswani, S. P. Ghrera, and S. Chandra, "A novel approach to outlier detection using modified grey wolf optimization and k-nearest neighbors algorithm," *Indian Journal of Science and Technology*, vol. 9, no. 44, pp. 1–8, 2016, https://doi.org/10.17485/ijst/2016/v9i44/105161.

[37] R. R. Rajammal, S. Mirjalili, G. Ekambaram, and N. Palanisamy, "Binary grey wolf optimizer with mutation and adaptive k-nearest neighbour for feature selection in Parkinson's disease diagnosis," *Knowledge-Based Systems*, vol. 246, pp. 108701, 2022, https://doi.org/10.1016/j.knosys.2022.108701.

[38] H. Tahraoui, S. Toumi, A. H. Hassein-Bey, A. Bousselma, A. N. E. H. Sid, A.-E. Belhadj, Z. Triki, M. Kebir, A. Amrane, J. Zhang, A. Sheta, A. Kouadri, A. Ghenai, and M. N. Sahmoune, "Advancing water quality research: K-nearest neighbor coupled with the improved grey wolf optimizer algorithm model unveils new possibilities for dry residue prediction," *Water*, vol. 15, no. 14, p. 2631, 2023, https://doi.org/10.3390/w15142631.

[39] H. Kraiem, F. Aymen, L. Yahya, A. Triviño, M. Alharthi, and S. S. M. Ghoneim, "A comparison between particle swarm and grey wolf optimization algorithms for improving the battery autonomy in a photovoltaic system," *Applied Sciences*, vol. 11, no. 16, p. 7732, 2021, https://doi.org/10.3390/app11167732.

[40] A. Pati, A. Panigrahi, M. Parhi, J. Giri, H. Qin, S. Mallik, S. R. Pattanayak, and U. K. Agrawal, "Performance assessment of hybrid machine learning approaches for breast cancer and recurrence prediction," *PLOS ONE*, vol. 19, no. 8, pp. 1–29, 2024, https://doi.org/10.1371/journal.pone.0304768.

[41] R. Oktafiani and E. I. Sela, "Breast cancer classification with principal component analysis and SMOTE using random forest method and support vector machine," *International Journal of Computer Applications*, vol. 186, no. 16, pp. 1-8, 2024, https://doi.org/10.5120/ijca2024923537.

[42] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: Review of a decade of research," *Artificial Intelligence Review*, vol. 57, no. 10, 2024, https://doi.org/10.1007/s10462-024-10884-2.

[43] P. Boileau, N. S. Hejazi, and S. Dudoit, "Exploring high-dimensional biological data with sparse contrastive principal component analysis," *Bioinformatics*, vol. 36, no. 11, pp. 3422–3430, 2020, https://doi.org/10.1093/bioinformatics/btaa176.

[44] S. Kumar and M. Singh, "Breast cancer detection based on feature selection using enhanced grey wolf optimizer and support vector machine algorithms," *Vietnam Journal of Computer Science*, vol. 8, no. 2, pp. 177–197, 2021, https://doi.org/10.1142/S219688882150007X.

[45] O. A. Alomari, S. N. Makhadmeh, M. A. Al-Betar, Z. A. A. Alyasseri, I. A. Doush, A. K. Abasi, M. A. Awadallah, and R. A. Zitar, "Gene selection for microarray data classification based on gray wolf optimizer enhanced with TRIZ-inspired operators," *Knowledge-Based Systems*, vol. 223, pp. 107034, 2021, https://doi.org/10.1016/j.knosys.2021.107034.

[46] R. Bro and A. K. Smilde, "Principal component analysis," *Analytical Methods*, vol. 6, no. 9, pp. 2812–2831, 2014, https://doi.org/10.1039/C3AY41907J.

[47] M. F. Kabir, T. Chen, and S. A. Ludwig, "A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction," *Healthcare Analytics*, vol. 3, p. 100125, 2023, https://doi.org/10.1016/j.health.2022.100125.

[48] M. H. Alshayeji, H. Ellethy, S. Abed, and R. Gupta, "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach," *Biomedical Signal Processing and Control*, vol. 71, pp. 103141, 2022, https://doi.org/10.1016/j.bspc.2021.103141.

[49] Y. Cakmak and I. Pacal, "Enhancing breast cancer diagnosis: A comparative evaluation of machine learning algorithms using the Wisconsin dataset," *Journal of Operational Intelligence*, vol. 3, no. 1, pp. 175–196, 2025, https://doi.org/10.31181/jopi31202539.

[50] A. Al Mamun, T. Bhuiyan, M. M. Hassan, and S. I. Anik, "Exploring the best machine learning models for breast cancer prediction in Wisconsin," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 1, pp. 1362–1368, 2025, https://doi.org/10.14569/IJACSA.2025.01601129.

[51] V. R. Hulipalled and G. Naganandini, "Breast cancer diagnosis using supervised machine learning for benign and malignant classification," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 25634–25640, 2025, https://doi.org/10.48084/etasr.11563.

[52] C. Syms, "Principal components analysis," *Encyclopedia of Ecology*, vol. 3, pp. 566–573, 2019, https://doi.org/10.1016/B978-0-12-409548-9.11152-2.

[53] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, "An overview of principal component analysis," *Journal of Signal and Information Processing*, vol. 4, no. 3, pp. 173–175, 2013, https://doi.org/10.4236/jsip.2013.43B031.

[54] J. F. Kenney and E. S. Keeping, *Mathematics of Statistics, Part Two*, 2nd ed. New Delhi, India: W. D. ten Broeck for Affiliated East-West Press Private Ltd., https://books.google.co.id/books?id=Jxx6nQEACAAJ.

[55] M. G. Kendall, and A. Stuart, "The advanced theory of statistics," *Population (French Edition)*, vol. 18, no. 2, p. 396, 1963, https://doi.org/10.2307/1527192.

[56] J. R. Magnus, "On differentiating eigenvalues and eigenvectors," *Econometric Theory*, vol. 1, no. 2, pp. 179–191, 1985, https://doi.org/10.1017/S0266466600011129.

[57] Z. Zhang, "Introduction to machine learning: K-nearest neighbors," *Annals of Translational Medicine*, vol. 4, no. 11, pp. 1–7, 2016, https://doi.org/10.21037/atm.2016.03.37.

[58] P. Cunningham and S. J. Delany, "K-nearest neighbour classifiers: A tutorial," *ACM Computing Surveys*, vol. 54, no. 6, pp.1–25, 2022, https://doi.org/10.1145/3459665.

[59] A. R. Lubis, M. Lubis, and Al-Khowarizmi, "Optimization of distance formula in k-nearest neighbor method," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 1, pp. 326–338, 2020, https://doi.org/10.11591/eei.v9i1.1464.

[60] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014, https://doi.org/10.1016/j.advengsoft.2013.12.007.

[61] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing K-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications," *Journal of Big Data*, vol. 11, no. 1, p. 113, 2024, https://doi.org/10.1186/s40537-024-00973-y.

[62] F. K. Ahmad and N. Yusoff, "Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier," in *Proceedings of the International Conference on Intelligent Systems Design and Applications*, 2014, pp. 121–125, https://doi.org/10.1109/ISDA.2013.6920720.

[63] A. I. Jony and A. K. B. Arnob, "Deep learning paradigms for breast cancer diagnosis: A comparative study on Wisconsin diagnostic dataset," *Malaysian Journal of Science and Advanced Technology*, pp. 109–117, 2024, https://doi.org/10.56532/mjsat.v4i2.245.

[64] N. Gupta, H. K. Gupta, R. Srivastava, C. Saxena, and Surjeet, "Design and implementation of artificial neural network classifiers based on hypertuning parameters for breast cancer diagnosis," *Procedia Computer Science*, vol. 233, pp. 929–938, 2024, https://doi.org/10.1016/j.procs.2024.03.282.

[65] A. M. Mohsen and A. S. K. A. Alhurdi, "Using an adaptive linear support vector machine algorithm for predicting breast cancer," *Arab Journal of Management, Banking and Financial Studies*, vol. 1, no. 1, pp. 90–103, 2025, https://doi.org/10.59559/ajmbfs.1.1.6.

[66] G. H. Chen and D. Shah, "Explaining the success of nearest neighbor methods in prediction," *Foundations and Trends in Machine Learning*, vol. 10, no. 5–6, pp. 337–588, May 2018, https://doi.org/10.1561/2200000064.

[67] F. A. Siregar, A. Candra, and Sutarman, "Determining the parameter k of k-nearest neighbors (KNN) using random grid search," in *Proceedings of ICSINTESA 2024 — 4th International Conference on Science, Information Technology and Smart Administration*, 2024, pp. 662–665, https://doi.org/10.1109/ICSINTESA62455.2024.10748027.

[68] A. A. Amer, S. D. Ravana, and R. A. A. Habeeb, "Effective k-nearest neighbor models for data classification enhancement," *Journal of Big Data*, vol. 12, no. 1, p. 86, 2025, https://doi.org/10.1186/s40537-025-01137-2.

[69] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, https://doi.org/10.1016/j.patrec.2005.10.010.

[70] I. David and R. Gelbard, "Using machine learning for systematic literature review case in point: Agile software development," *WIREs Data Mining and Knowledge Discovery*, vol. 15, no. 1, p. e1569, 2025, https://doi.org/10.1002/widm.1569.

[71] M. A. Jawaid, S. A. Naqvi, M. Safdar, and M. Haroon, "Enhancing credit card transaction security using support vector machines and feature engineering techniques," *International Journal of Innovative Research in Computer Science & Technology*, vol. 13, no. 3, pp. 82–88, 2025, https://doi.org/10.55524/ijircst.2025.13.3.14.

[72] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint,* arXiv:2010.16061, 2020, https://arxiv.org/abs/2010.16061.