

RESEARCH ARTICLE

Grey Wolf Optimizer-Based Feature Selection for Naive Bayes Lung Cancer Risk Classification

Ilhanatus Saadah¹, Mohd Izham², Safuan^{3,*}, Rima Dias Ramadhani³ and Muhamad Arifpin Mansor⁴

¹Undergraduate Program in Informatics, Faculty of Computer Science and Digital Technology, Universitas Muhammadiyah Semarang, Semarang, Indonesia

²Universiti Malaysia Pahang Al-Sultan Abdullah, Malaysia

³Department of Informatics, Faculty of Computer Science and Digital Technology, Universitas Muhammadiyah Semarang, Semarang, Indonesia

⁴Department of Information and Computer Engineering, National Institute of Technology, Sasebo College, Japan

*Correspondence safuan@unimus.ac.id

Received: 5 September 2026; Revised: 18 September 2026; Accepted: 26 September 2025; Published: 30 September 2025

ABSTRACT

Lung cancer risk classification is an important task in health informatics because early identification can support further clinical assessment and risk management. Machine learning can process patient risk factors and symptoms, but classification performance can be affected by irrelevant or less informative input features. This study addresses this problem by applying the Grey Wolf Optimizer (GWO) as a feature-selection method for Naive Bayes classification of three lung cancer risk levels: Low, Medium, and High. The dataset contains 1,000 patient records with 25 attributes related to lung cancer risk factors and symptoms. The proposed procedure includes data preprocessing, an 80:20 training-testing split, GWO-based feature selection, and Naive Bayes classification. Each candidate solution in GWO represents a feature subset, while Naive Bayes classification accuracy is used to evaluate the candidate solutions. The baseline Naive Bayes model achieved an accuracy of 89.50%. GWO reduced the number of input attributes from 25 to 10, and the resulting Naive Bayes model achieved an accuracy of 96.50%. Precision, recall, and F1-score also increased across the Low, Medium, and High risk classes after feature selection. These results show that GWO-based feature selection reduced the input dimensionality while producing higher classification metrics than the baseline Naive Bayes model under the evaluated experimental setting.

KEYWORDS

Lung Cancer Classification; Grey Wolf Optimizer; Feature Selection; Naive Bayes; Machine Learning; Health Informatics

1. Introduction

Lung cancer remains a major health problem because of its incidence and mortality rates. Machine learning has been applied to lung adenocarcinoma diagnosis and prognosis using molecular and clinical information [1]. Computed tomography images have also been used for multi-class lung cancer detection through modified

DenseNet feature extraction and machine learning classification [2]. Explainable artificial intelligence has been applied to identify biomarkers for non-small cell lung cancer subtype classification [3]. Radiomic feature selection has been evaluated for distinguishing lung adenocarcinoma from squamous cell carcinoma using chest computed tomography images [4]. Ensemble transfer learning has also been applied to lung cancer image classification using VGG16, ResNet50, InceptionV3, and Grad-CAM [5]. The performance of a classification model depends on the information contained in its input features. Feature selection reduces the input space by retaining a subset of variables for the classification process. This process is relevant when a dataset contains redundant or less informative attributes. A binary Grey Wolf Optimizer has been applied to feature selection in biomedical data classification [6]. A hybrid gradient descent and Grey Wolf Optimizer has also been evaluated for feature selection on medical datasets [7].

The Grey Wolf Optimizer (GWO) is a population-based metaheuristic inspired by the leadership hierarchy and hunting mechanism of grey wolves [8]. The algorithm represents alpha, beta, and delta wolves as leading candidate solutions during the search process. A random-walk GWO with a dispersion factor has been evaluated for feature selection in chronic-disease prediction [9]. A two-stage improved GWO has also been developed for high-dimensional feature-selection problems [10]. An enhanced GWO has been combined with machine learning techniques for biomedical data classification [11].

Several studies have introduced additional search strategies for GWO-based feature selection. A local opposition-learning golden-sine GWO has been evaluated for feature selection in classification problems [12]. A reinforcement learning-based comprehensive learning GWO has also been developed for feature selection [13]. An S-shaped GWO-based FOX algorithm has been evaluated across multiple feature-selection datasets [14]. A role-oriented binary GWO combines foraging-following behavior with Lévy flight for feature-subset search [15]. A multi-strategy reinforced hierarchical GWO has been developed for feature selection [16]. A mixed-variable GWO has been proposed for joint feature selection and classifier-parameter tuning [17]. A Golden Jackal–Grey Wolf hybrid optimization algorithm has also been evaluated for feature selection [18].

The combination of GWO and classification algorithms has also been examined in specific application domains. A hybrid GWO and Particle Swarm Optimization method has been applied to feature selection for breast cancer classification using a Distance-Biased Naive Bayes classifier [19]. GWO has also been combined with Naive Bayes for email spam classification [20]. The reviewed studies show that feature selection has been applied to medical classification, lung cancer analysis, and high-dimensional classification problems. The reviewed GWO studies also show that different search strategies can be used to identify reduced feature subsets. However, the reviewed studies do not evaluate standard GWO-based feature selection with Naive Bayes for three-class lung cancer risk classification using the dataset examined in this study.

This study applies GWO-based feature selection to Naive Bayes for lung cancer risk classification. The dataset contains 1,000 patient records with 25 attributes related to lung cancer risk factors and symptoms. The classification task divides the observations into Low, Medium, and High risk classes. GWO searches for a reduced feature subset before the Naive Bayes classification process. The model performance is evaluated using accuracy, precision, recall, and F1-score. This study provides three contributions. First, the study applies GWO-based feature selection to Naive Bayes for three-class lung cancer risk classification. Second, the study identifies a reduced feature subset from the available lung cancer attributes. Third, the study compares Naive Bayes performance before and after GWO-based feature selection using accuracy, precision, recall, and F1-score.

2. Related Work

Machine learning and feature-selection methods have been applied to medical and high-dimensional classification problems. A modified DenseNet model combined with feature-selection methods and machine learning classifiers was evaluated for multi-class lung cancer detection from computed tomography images [2]. An Enhanced Grey Wolf Optimizer was combined with ensemble machine learning classifiers for biomedical data classification [11]. A Master–Slave Binary Grey Wolf Optimizer was also evaluated for feature selection on high-dimensional biomedical datasets [6].

Several studies have modified the Grey Wolf Optimizer to improve feature-subset search. A random-walk Grey Wolf Optimizer with a dispersion factor was evaluated for feature selection in chronic-disease prediction [9]. A two-stage improved Grey Wolf Optimizer was developed for high-dimensional feature selection [10]. An Adaptively Balanced Grey Wolf Optimizer introduced a random wolf and an adaptive coefficient to regulate exploration and exploitation in high-dimensional classification [21]. These studies apply different mechanisms to modify the search behavior of the standard Grey Wolf Optimizer.

Recent studies have introduced additional learning and population-management strategies into Grey Wolf Optimizer-based feature selection. A reinforcement learning-based comprehensive learning Grey Wolf Optimizer uses reinforcement learning to coordinate search strategies during feature selection [22]. A role-oriented binary Grey Wolf Optimizer combines foraging-following behavior and Lévy flight for feature-subset search [15]. A multi-strategy reinforced hierarchical Grey Wolf Optimizer uses memory-based diversity, Lévy flight, and hierarchical population segmentation for feature selection [23]. These approaches modify different components of the Grey Wolf Optimizer while retaining feature selection as the optimization objective.

Feature selection has also been formulated together with other optimization objectives. A mixed-variable Grey Wolf Optimizer jointly optimizes feature subsets and classifier parameters [24]. An S-shaped Grey Wolf Optimizer-based FOX algorithm combines Grey Wolf Optimizer mechanisms with the FOX algorithm and a binary transfer function for feature selection [14]. A Golden Jackal–Grey Wolf hybrid algorithm combines two population-based optimization mechanisms to remove irrelevant and redundant features [18].

Other studies have extended Grey Wolf Optimizer-based feature selection to different high-dimensional application settings. A hybrid rice optimization and Grey Wolf Optimizer framework combines filter and wrapper mechanisms for high-dimensional biomedical feature selection [25]. A multi-strategy Grey Wolf Optimizer has been applied to feature selection for performance evaluation of open-ended funds [26]. Modified Grey Wolf Optimizer variants have also been evaluated with an optimized Decision Tree for feature selection in peer-to-peer lending credit-risk assessment [27].

The reviewed studies demonstrate that Grey Wolf Optimizer-based feature selection has been evaluated in biomedical data, chronic-disease prediction, high-dimensional classification, financial data, and general classification datasets. The studies also show that GWO variants use different mechanisms to regulate population diversity, exploration, exploitation, binary representation, and classifier parameters. However, the reviewed studies do not evaluate standard GWO-based feature selection with Naive Bayes for three-class lung cancer risk classification using the dataset examined in this study. This study therefore compares Naive Bayes before and after GWO-based feature selection using accuracy, precision, recall, and F1-score.

3. Methods

3.1. Proposed Research Framework

Figure 1 presents the experimental workflow used in this study. The framework compares two classification settings: Naive Bayes without feature selection and Naive Bayes with GWO-based feature selection. Both settings use the same training and testing partitions to maintain a consistent evaluation procedure.

3.2. Dataset

This study uses a secondary lung cancer dataset obtained from Kaggle. The dataset contains 1,000 patient records with 25 attributes related to lung cancer risk factors and symptoms. The target variable consists of three lung cancer risk classes: Low, Medium, and High. The dataset is used to evaluate two classification conditions. The first condition applies Naive Bayes without feature selection. The second condition applies Naive Bayes using the features selected by the Grey Wolf Optimizer. The comparison between these two conditions is used to evaluate the effect of GWO-based feature selection on lung cancer risk classification.

3.3. Data Preprocessing

The preprocessing stage prepares the dataset before the classification process. The procedure consists of data encoding, removal of irrelevant attributes, data normalization, and dataset splitting. Categorical attributes are converted into numerical representations through an encoding process. The *Patient Id* and *Index* attributes are removed because they are not used as predictive variables in the classification process. The numerical attributes are normalized to place the input values on a comparable scale. The dataset is then divided into training and testing subsets using an 80:20 ratio. The training subset is used to build the classification model, while the testing subset is used to evaluate model performance.

3.4. Grey Wolf Optimizer

GWO is a population-based metaheuristic algorithm inspired by the social hierarchy and hunting mechanism of grey wolves [8]. In this study, GWO is used as a feature-selection method to identify a reduced subset of

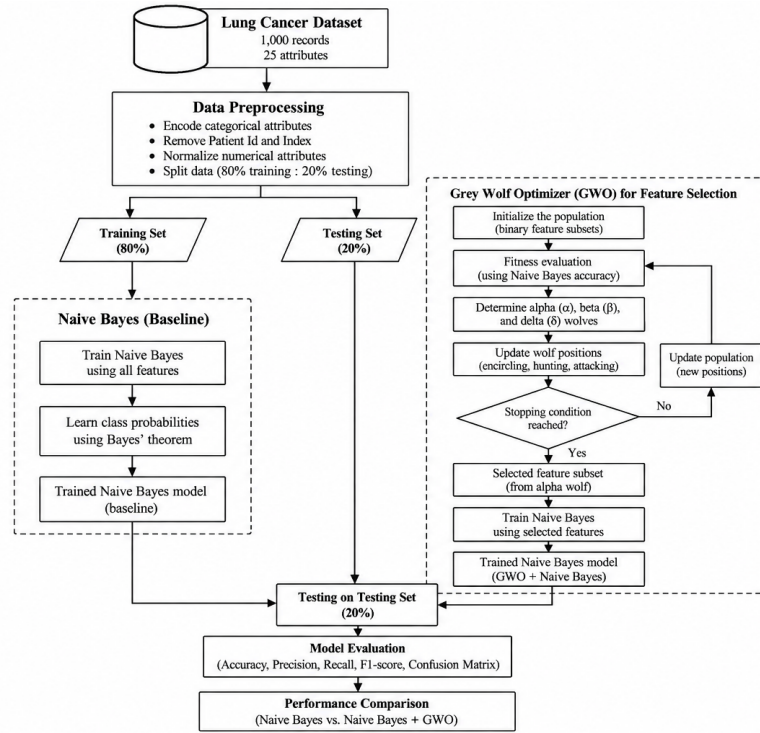


Figure 1. Proposed framework for GWO-based feature selection and Naive Bayes classification of lung cancer risk.

attributes for lung cancer risk classification. The GWO population is organized into four hierarchical levels. The best candidate solution is represented by the alpha wolf (α), while the second-best and third-best solutions are represented by the beta wolf (β) and delta wolf (δ), respectively. The remaining candidate solutions are represented by omega wolves (ω). The alpha, beta, and delta wolves guide the search process because they represent the three best solutions obtained during optimization [8].

The hunting mechanism of GWO consists of three main processes: encircling the prey, hunting the prey, and attacking the prey. The encircling process models the movement of wolves toward a candidate solution. The mathematical formulation is defined as:

$$\mathbf{D} = |\mathbf{C} \cdot \mathbf{X}_p(t) - \mathbf{X}(t)|, \quad (1)$$

$$\mathbf{X}(t+1) = \mathbf{X}_p(t) - \mathbf{A} \cdot \mathbf{D}, \quad (2)$$

where $\mathbf{X}(t)$ represents the current position of a wolf, $\mathbf{X}_p(t)$ represents the estimated position of the prey, and t represents the current iteration.

The coefficient vectors $\mathbf{A}$ and $\mathbf{C}$ are calculated as follows:

$$\mathbf{A} = 2\mathbf{a} \cdot \mathbf{r}_1 - \mathbf{a}, \quad (3)$$

$$\mathbf{C} = 2\mathbf{r}_2, \quad (4)$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are random vectors in the interval $[0, 1]$. The control parameter $\mathbf{a}$ decreases linearly from 2 to 0 during the optimization process:

$$a = 2 - 2 \left(\frac{t}{T} \right), \quad (5)$$

where T denotes the maximum number of iterations. The hunting process uses the positions of the alpha, beta, and delta wolves to estimate a new candidate position. The distances between a wolf and the three leaders are calculated as:

$$\mathbf{D}_\alpha = |\mathbf{C}_1 \mathbf{X}_\alpha - \mathbf{X}|, \quad (6)$$

$$\mathbf{D}_\beta = |\mathbf{C}_2\mathbf{X}_\beta - \mathbf{X}|, \quad (7)$$

$$\mathbf{D}_\delta = |\mathbf{C}_3\mathbf{X}_\delta - \mathbf{X}|. \quad (8)$$

The candidate positions relative to the three leaders are then calculated as:

$$\mathbf{X}_1 = \mathbf{X}_\alpha - \mathbf{A}_1\mathbf{D}_\alpha, \quad (9)$$

$$\mathbf{X}_2 = \mathbf{X}_\beta - \mathbf{A}_2\mathbf{D}_\beta, \quad (10)$$

$$\mathbf{X}_3 = \mathbf{X}_\delta - \mathbf{A}_3\mathbf{D}_\delta. \quad (11)$$

The final position of each wolf is updated by averaging the three candidate positions:

$$\mathbf{X}(t+1) = \frac{\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3}{3}. \quad (12)$$

In the feature-selection process, each wolf represents a candidate feature subset. The candidate solution is encoded as a binary vector:

$$\mathbf{X}_i = [x_{i1}, x_{i2}, \dots, x_{id}], \quad x_{ij} \in \{0, 1\}, \quad (13)$$

where d denotes the number of candidate features. A value of $x_{ij} = 1$ indicates that the j -th feature is selected, while $x_{ij} = 0$ indicates that the feature is excluded. The continuous position generated by GWO must be converted into a binary value before the candidate subset can be evaluated. A binary transformation can be expressed through a transfer function:

$$S(x) = \frac{1}{1 + \exp(-x)}, \quad (14)$$

followed by a thresholding rule:

$$x_{ij} = \begin{cases} 1, & \text{if } S(x_{ij}) \geq 0.5, \\ 0, & \text{if } S(x_{ij}) < 0.5. \end{cases} \quad (15)$$

Each binary vector is evaluated using the Naive Bayes classifier. In this study, the fitness evaluation is based on the classification accuracy produced by Naive Bayes. A candidate subset that produces a higher classification accuracy is considered a better solution under the fitness formulation used in the experiment. The optimization process begins by generating an initial population of candidate feature subsets. Each candidate subset is evaluated using Naive Bayes. The three best candidate solutions are assigned as α , β , and δ . The remaining wolves update their positions according to Equations 6–12. The updated positions are converted into binary feature-selection vectors before the next fitness evaluation. The optimization process continues until the stopping condition is reached. The best solution obtained at the end of the process is used as the selected feature subset for the Naive Bayes classification model.

Algorithm 1 summarizes the GWO-based feature-selection process used in this study. Each wolf represents a candidate feature subset. The fitness value of each candidate is calculated using the classification accuracy produced by the Naive Bayes classifier. The three best solutions guide the position updates during each iteration. The optimization process continues until the maximum number of iterations is reached, and the alpha wolf represents the selected feature subset.

3.5. Naive Bayes

Naive Bayes is used as the classification algorithm in this study. The algorithm estimates the probability of each lung cancer risk class based on the observed features. The classification task consists of three classes: Low, Medium, and High. Naive Bayes applies Bayes' theorem to calculate the posterior probability of a class given an observed feature vector. The posterior probability is expressed as:

$$P(C_k | X) = \frac{P(C_k)P(X | C_k)}{P(X)}, \quad (16)$$

Algorithm 1: GWO-Based Feature Selection with Naive Bayes

Input: Training dataset D , number of wolves n , maximum iterations T
Output: Best feature subset $\mathbf{X}_\alpha$

- 1 Initialize the wolf population $\mathbf{X}_i, i = 1, 2, \dots, n$;
- 2 Convert each candidate solution into a binary feature vector;
- 3 Evaluate each candidate subset using Naive Bayes accuracy;
- 4 Determine the three best solutions: $\mathbf{X}_\alpha, \mathbf{X}_\beta$, and $\mathbf{X}_\delta$;
- 5 Set $t \leftarrow 0$;
- 6 **while** $t < T$ **do**
 - 7 Update the control parameter a using Equation 5;
 - 8 **for** $i \leftarrow 1$ **to** n **do**
 - 9 Calculate $\mathbf{D}_\alpha$ using Equation 6;
 - 10 Calculate $\mathbf{D}_\beta$ using Equation 7;
 - 11 Calculate $\mathbf{D}_\delta$ using Equation 8;
 - 12 Calculate $\mathbf{X}_1$ using Equation 9;
 - 13 Calculate $\mathbf{X}_2$ using Equation 10;
 - 14 Calculate $\mathbf{X}_3$ using Equation 11;
 - 15 Update $\mathbf{X}_i$ using Equation 12;
 - 16 Convert the updated position into a binary feature vector;
 - 17 Evaluate the selected feature subset using Naive Bayes accuracy;
 - 18 Update $\mathbf{X}_\alpha, \mathbf{X}_\beta$, and $\mathbf{X}_\delta$;
 - 19 Set $t \leftarrow t + 1$;
- 20 **return** $\mathbf{X}_\alpha$;

where C_k denotes the k -th class and X denotes the observed feature vector. Under the conditional-independence assumption of Naive Bayes, the likelihood of the feature vector is expressed as:

$$P(X | C_k) = \prod_{i=1}^n P(X_i | C_k), \quad (17)$$

where X_i denotes the i -th feature and n denotes the number of input features. The posterior probability for each class is therefore calculated as:

$$P(C_k | X) = \frac{P(C_k) \prod_{i=1}^n P(X_i | C_k)}{P(X)}. \quad (18)$$

The denominator $P(X)$ is identical for all candidate classes during classification. The predicted class can therefore be determined by maximizing the numerator:

$$\hat{C} = \arg \max_{C_k} \left[P(C_k) \prod_{i=1}^n P(X_i | C_k) \right]. \quad (19)$$

In this study, the Naive Bayes model is evaluated under two experimental conditions. The first condition uses the available input features without GWO-based feature selection. The second condition uses the feature subset selected by GWO. The same classification procedure is applied to both conditions. The trained model assigns each test instance to the Low, Medium, or High lung cancer risk class. The classification results are then evaluated using accuracy, precision, recall, and F1-score.

3.6. Model Evaluation

The classification performance is evaluated using a confusion matrix, accuracy, precision, recall, and F1-score. The confusion matrix summarizes the relationship between the actual classes and the predicted classes. Accuracy measures the proportion of correctly classified instances among all evaluated instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (20)$$

Algorithm 2: Naive Bayes Classification for Lung Cancer Risk

Input: Training dataset D_{train} , testing dataset D_{test} , feature set F
Output: Predicted lung cancer risk classes

- 1 Extract the selected features F from D_{train} and D_{test} ;
- 2 Define the class set $\mathcal{C} = \{\text{Low, Medium, High}\}$;
- 3 **foreach** $C_k \in \mathcal{C}$ **do**
- 4 Calculate the prior probability $P(C_k)$;
- 5 Estimate the conditional probability $P(X_i | C_k)$ for each feature X_i ;
- 6 **foreach** test instance $X \in D_{\text{test}}$ **do**
- 7 Calculate the posterior probability for each class using Equation 18;
- 8 Determine the predicted class using Equation 19;
- 9 Return the predicted classes;

where TP denotes true positive, TN denotes true negative, FP denotes false positive, and FN denotes false negative. Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive:

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (21)$$

Recall measures the proportion of correctly identified positive instances among all actual positive instances:

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (22)$$

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

$$\text{F1-score} = 2 \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right). \quad (23)$$

The evaluation is performed for two experimental conditions. The first condition evaluates Naive Bayes without GWO-based feature selection. The second condition evaluates Naive Bayes using the feature subset selected by GWO. The resulting metric values are then compared to examine the effect of feature selection on classification performance.

4. Results and Discussion

This section presents the classification results obtained from Naive Bayes before and after GWO-based feature selection. The comparison uses precision, recall, F1-score, and accuracy to evaluate the two experimental conditions. The performance of Naive Bayes and Naive Bayes with GWO-based feature selection is presented in Table 1. The evaluation covers the Low, Medium, and High lung cancer risk classes.

Table 1. Classification performance of Naive Bayes before and after GWO-based feature selection.

Class	Naive Bayes			Naive Bayes + GWO		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Low	1.0000	0.8545	0.9216	1.0000	1.0000	1.0000
Medium	0.8548	0.8413	0.8480	0.9841	0.9841	0.9841
High	0.8681	0.9634	0.9133	0.9878	0.9878	0.9878

The Naive Bayes model achieved an accuracy of 89.50%. After GWO-based feature selection, the accuracy increased to 96.50%. The precision, recall, and F1-score values also increased across the three evaluated classes. For the Low-risk class, the F1-score increased from 0.9216 to 1.0000. For the Medium-risk class, the F1-score increased from 0.8480 to 0.9841. For the High-risk class, the F1-score increased from 0.9133 to 0.9878. These results show that the model using GWO-selected features produced higher metric values than the model using the original feature set under the reported experimental setting.

Figure 2 presents the confusion matrices of Naive Bayes before and after GWO-based feature selection. The matrices show the distribution of correct and incorrect predictions for the Low, Medium, and High lung cancer risk classes. Before feature selection, the model correctly classified 47 Low-risk instances, 53 Medium-risk instances, and 79 High-risk instances. After GWO-based feature selection, the model correctly classified 51 Low-risk instances, 60 Medium-risk instances, and 82 High-risk instances. The matrix after feature selection shows fewer misclassified instances than the baseline model. These results are consistent with the higher classification metrics obtained after GWO-based feature selection.

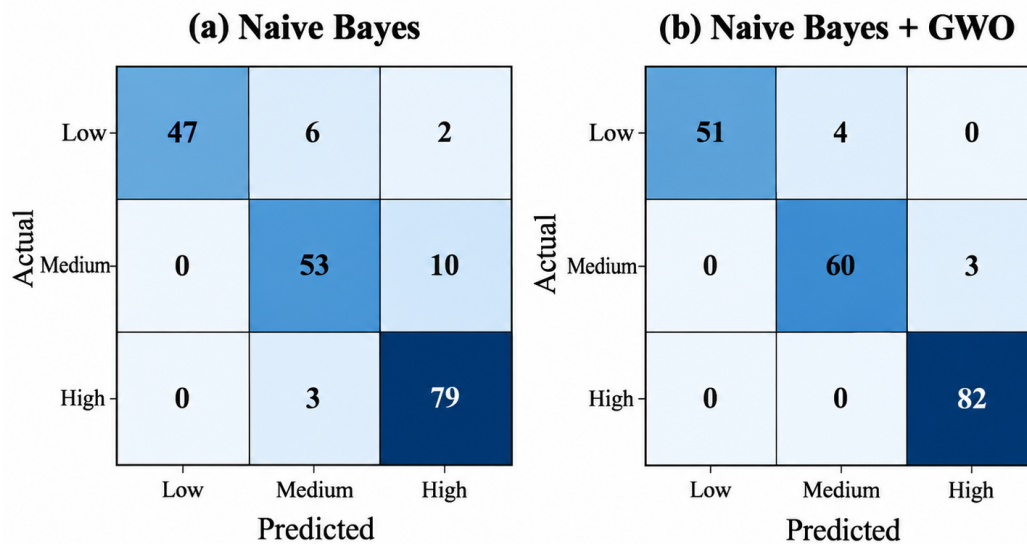


Figure 2. Confusion matrices of Naive Bayes before and after GWO-based feature selection.

GWO-based feature selection reduced the number of input attributes from 25 to 10. The selected subset was subsequently used as the input for the Naive Bayes classifier. The Naive Bayes model achieved an accuracy of 89.50% before feature selection. The model using the GWO-selected features achieved an accuracy of 96.50%. The classification metrics reported for the Low, Medium, and High classes also increased after feature selection. The reduction in the number of input features was accompanied by higher classification performance under the experimental setting used in this study. This result indicates that the reduced feature subset contained sufficient information for the Naive Bayes classifier to distinguish the three lung cancer risk classes. The experiment also shows that using all available attributes did not produce the highest classification performance. The model obtained higher metric values after the input space was reduced through the GWO search process. However, the reported results describe the performance of the evaluated dataset and experimental configuration.

Figure 3 summarizes the effect of GWO-based feature selection on model complexity and classification accuracy. The number of input features decreased from 25 to 10, while the classification accuracy increased from 89.50% to 96.50%. The figure therefore illustrates the relationship between feature reduction and classification performance under the evaluated experimental setting.

The experimental results show that GWO-based feature selection changed both the dimensionality of the input data and the classification performance of Naive Bayes. The number of input features decreased from 25 to 10, while the reported classification metrics increased after feature selection. The confusion matrix shows that the number of misclassified instances decreased after GWO-based feature selection. The reduction in classification errors occurred across the Low, Medium, and High risk classes. This pattern is consistent with the higher precision, recall, F1-score, and accuracy values obtained after feature selection.

The results indicate that the reduced feature subset retained information that was sufficient for the Naive Bayes classifier under the evaluated experimental setting. The use of fewer features also reduced the number of input variables processed by the classifier. The findings are consistent with previous studies that applied GWO-based feature selection to biomedical and high-dimensional classification problems. Binary and modified GWO variants have been used to identify reduced feature subsets before classification [6, 9]. Enhanced GWO mechanisms have also been evaluated with machine learning models for biomedical classification [11]. These studies support the use of population-based optimization for searching feature subsets in classification tasks.

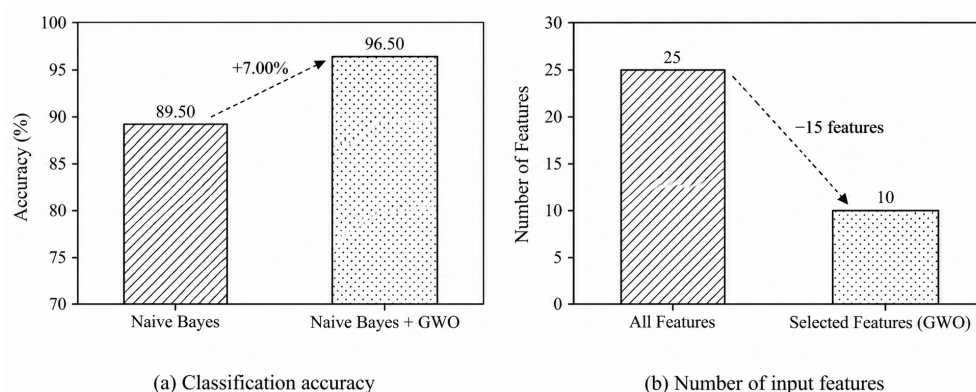


Figure 3. Effect of GWO-based feature selection on (a) classification accuracy and (b) number of input features.

The results of this study should be interpreted within the evaluated dataset and experimental configuration. The dataset contains 1,000 patient records and uses a single 80:20 training-testing split. Therefore, the reported performance does not directly represent performance on other lung cancer datasets or clinical populations. The comparison between the baseline and GWO-based models shows that feature selection contributed to a different classification outcome under the same dataset partition. Further evaluation using repeated cross-validation, external datasets, and alternative classifiers is required to examine the consistency of the observed performance.

5. Conclusion and Future Work

This study evaluated GWO-based feature selection with Naive Bayes for three-class lung cancer risk classification. The dataset contained 1,000 patient records with 25 attributes and three target classes: Low, Medium, and High. The GWO-based feature-selection process reduced the number of input attributes from 25 to 10. The Naive Bayes model achieved an accuracy of 89.50% before feature selection. After GWO-based feature selection, the accuracy increased to 96.50%. The precision, recall, and F1-score values also increased across the three evaluated classes. The results show that the reduced feature subset was associated with higher classification performance under the experimental setting used in this study. The confusion matrix also showed fewer classification errors after GWO-based feature selection.

Future work can evaluate the proposed approach using repeated cross-validation and external lung cancer datasets. Additional classifiers can also be compared using the same selected feature subset. Alternative feature-selection methods can be evaluated under the same experimental conditions. Future studies can also examine the selected attributes in greater detail to determine their contribution to lung cancer risk classification.

Author Contributions

Conceptualization: I.S., S., and R.D.R.; Methodology: I.S., S., and R.D.R.; Software: I.S.; Validation: M.I., S., R.D.R., and M.A.M.; Formal analysis: I.S., M.I., and M.A.M.; Investigation: I.S.; Resources: M.I. and M.A.M.; Data curation: I.S.; Writing—original draft preparation: I.S.; Writing—review and editing: M.I., S., R.D.R., and M.A.M.; Visualization: I.S.; Supervision: S. and R.D.R.; Project administration: I.S., S., and R.D.R.; Funding acquisition: Not applicable. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The lung cancer dataset used in this study was obtained from Kaggle. The dataset contains 1,000 patient records with 25 attributes related to lung cancer risk factors and symptoms.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

ChatGPT, developed by OpenAI, was used to assist with language editing and manuscript formatting. The authors reviewed and verified all generated content and remain fully responsible for the accuracy, interpretation, and integrity of the manuscript.

References

- [1] Lin L, Bao Y. Development and Validation of Machine Learning Models for Diagnosis and Prognosis of Lung Adenocarcinoma, and Immune Infiltration Analysis. *Scientific Reports*. 2024;14:22081. Available from: <https://doi.org/10.1038/s41598-024-73498-2>.
- [2] Lanjewar MG, Panchbhai KG, Charanarur P. Lung Cancer Detection from CT Scans Using Modified DenseNet with Feature Selection Methods and ML Classifiers. *Expert Systems with Applications*. 2023;224:119961. Available from: <https://doi.org/10.1016/j.eswa.2023.119961>.
- [3] Dwivedi K, Rajpal A, Rajpal S, Agarwal M, Kumar V, Kumar N. An Explainable AI-Driven Biomarker Discovery Framework for Non-Small Cell Lung Cancer Classification. *Computers in Biology and Medicine*. 2023;153:106544. Available from: <https://doi.org/10.1016/j.compbiomed.2023.106544>.
- [4] Selvam M, Sadanandan A, Chandrasekharan A, Ramesh S, Murali A, Krishnamurthi G. Radiomics for Differentiating Adenocarcinoma and Squamous Cell Carcinoma in Non-Small Cell Lung Cancer Beyond Nodule Morphology in Chest CT. *Scientific Reports*. 2024;14:32088. Available from: <https://doi.org/10.1038/s41598-024-83786-6>.
- [5] Kumaran S Y, Jeya JJ, Mahesh TR, Khan SB, Alzahrani S, Alojail M. Explainable Lung Cancer Classification with Ensemble Transfer Learning of VGG16, ResNet50 and InceptionV3 Using Grad-CAM. *BMC Medical Imaging*. 2024;24(1):176. Available from: <https://doi.org/10.1186/s12880-024-01345-x>.
- [6] Momanyi E, Segera D. A Master-Slave Binary Grey Wolf Optimizer for Optimal Feature Selection in Biomedical Data Classification. *BioMed Research International*. 2021;2021:5556941. Available from: <https://doi.org/10.1155/2021/5556941>.
- [7] Kitonyi PM, Segera DR. Hybrid Gradient Descent Grey Wolf Optimizer for Optimal Feature Selection. *BioMed Research International*. 2021;2021:2555622. Available from: <https://doi.org/10.1155/2021/2555622>.
- [8] Mirjalili S, Mirjalili SM, Lewis A. Grey Wolf Optimizer. *Advances in Engineering Software*. 2014;69:46-61. Available from: <https://doi.org/10.1016/j.advengsoft.2013.12.007>.
- [9] Preeti, Deep K. A Random Walk Grey Wolf Optimizer Based on Dispersion Factor for Feature Selection on Chronic Disease Prediction. *Expert Systems with Applications*. 2022;206:117864. Available from: <https://doi.org/10.1016/j.eswa.2022.117864>.
- [10] Shen C, Zhang K. Two-Stage Improved Grey Wolf Optimization Algorithm for Feature Selection on High-Dimensional Classification. *Complex & Intelligent Systems*. 2022;8:2769-89. Available from: <https://doi.org/10.1007/s40747-021-00452-4>.
- [11] Chakraborty C, Kishor A, Rodrigues JJPC. Novel Enhanced-Grey Wolf Optimization Hybrid Machine Learning Technique for Biomedical Data Computation. *Computers & Electrical Engineering*. 2022;99:107778. Available from: <https://doi.org/10.1016/j.compeleceng.2022.107778>.

- [12] A Local Opposition-Learning Golden-Sine Grey Wolf Optimization Algorithm for Feature Selection in Data Classification. *Applied Soft Computing*. 2023;142:110319. Available from: <https://doi.org/10.1016/j.asoc.2023.110319>.
- [13] Hu Z, Yu X. Reinforcement Learning-Based Comprehensive Learning Grey Wolf Optimizer for Feature Selection. *Applied Soft Computing*. 2023;149:110959. Available from: <https://doi.org/10.1016/j.asoc.2023.110959>.
- [14] Feda AK, Adegboye M, Adegboye OR, Agyekum EB, Mbasso WF, Kamel S. S-Shaped Grey Wolf Optimizer-Based FOX Algorithm for Feature Selection. *Heliyon*. 2024;10(2):e24192. Available from: <https://doi.org/10.1016/j.heliyon.2024.e24192>.
- [15] Wang Y, Ran S, Wang GG. Role-Oriented Binary Grey Wolf Optimizer Using Foraging-Following and Lévy Flight for Feature Selection. *Applied Mathematical Modelling*. 2024;126:310-26. Available from: <https://doi.org/10.1016/j.apm.2023.08.043>.
- [16] A Multi-Strategy Driven Reinforced Hierarchical Operator in the Grey Wolf Optimizer for Feature Selection. *Information Sciences*. 2024;677:120924. Available from: <https://doi.org/10.1016/j.ins.2024.120924>.
- [17] Single-Objective and Multi-Objective Mixed-Variable Grey Wolf Optimizer for Joint Feature Selection and Classifier Parameter Tuning. *Applied Soft Computing*. 2024;165:112121. Available from: <https://doi.org/10.1016/j.asoc.2024.112121>.
- [18] Liu G, Guo Z, Liu W, Jiang F, Fu E. A Feature Selection Method Based on the Golden Jackal-Grey Wolf Hybrid Optimization Algorithm. *PLOS ONE*. 2024;19(1):e0295579. Available from: <https://doi.org/10.1371/journal.pone.0295579>.
- [19] Devi RS, Saragih TH, Faisal MR. Feature Selection Using Hybrid Grey Wolf Optimization and Particle Swarm Optimization in Distance-Biased Naive Bayes for Breast Cancer Classification. *Jurnal Informatika Polinema*. 2024;10(2):307-14.
- [20] Zraqou J, Al-Helali AH, Maqableh W, Fakhouri H, Alkhadour W. Robust Email Spam Filtering Using a Hybrid of the Grey Wolf Optimizer and the Naive Bayes Classifier. *Cybernetics and Information Technologies*. 2023;23(4):79-90.
- [21] Wang J, Lin D, Zhang Y, Huang S. An Adaptively Balanced Grey Wolf Optimization Algorithm for Feature Selection on High-Dimensional Classification. *Engineering Applications of Artificial Intelligence*. 2022;114:105088. Available from: <https://doi.org/10.1016/j.engappai.2022.105088>.
- [22] Hu Z, Yu X. Reinforcement Learning-Based Comprehensive Learning Grey Wolf Optimizer for Feature Selection. *Applied Soft Computing*. 2023;149:110959. Available from: <https://doi.org/10.1016/j.asoc.2023.110959>.
- [23] Yu X, Hu Z. A Multi-Strategy Driven Reinforced Hierarchical Operator in the Grey Wolf Optimizer for Feature Selection. *Information Sciences*. 2024;677:120924. Available from: <https://doi.org/10.1016/j.ins.2024.120924>.
- [24] Li H, Kang H, Li J, Pang Y, Sun G, Liang S. Single-Objective and Multi-Objective Mixed-Variable Grey Wolf Optimizer for Joint Feature Selection and Classifier Parameter Tuning. *Applied Soft Computing*. 2024;165:112121. Available from: <https://doi.org/10.1016/j.asoc.2024.112121>.
- [25] Ye Z, Huang R, Zhou W, Wang M, Cai T, He Q, et al. Hybrid Rice Optimization Algorithm Inspired Grey Wolf Optimizer for High-Dimensional Feature Selection. *Scientific Reports*. 2024;14:30741. Available from: <https://doi.org/10.1038/s41598-024-80648-z>.
- [26] Chang D, Rao C, Xiao X, Hu F, Goh M. Multiple Strategies Based Grey Wolf Optimizer for Feature Selection in Performance Evaluation of Open-Ended Funds. *Swarm and Evolutionary Computation*. 2024;86:101518. Available from: <https://doi.org/10.1016/j.swevo.2024.101518>.

- [27] Sam'an M, Mat Deris M, Farikhin. Feature Selection in Peer-to-Peer Lending Based on Hybrid Modified Grey Wolf Optimization with Optimized Decision Tree for Credit Risk Assessment. *International Journal of Management Science and Engineering Management*. 2025;20(1):141-58. Available from: <https://doi.org/10.1080/17509653.2024.2381742>.