ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

OPTIMIZED FEATURE SELECTION FOR LUNG CANCER CLASSIFICATION USING NATURE-INSPIRED ALGORITHMS

[1]Sumbul Azeem
[2]Shazia Javed
[1,2] Department of Mathematics, Lahore College for Women University, Lahore, Pakistan
[1]Email: Sumbul.azeem@lcwu.edu.pk
[2]Email: Shazia.javed@lcwu.edu.pk

Abstract
Lung cancer is still one of the most fatal forms of cancer around the world. Early detection is indeed a crucial aspect that can help improve survival rates in cancer patients. Modern imaging and genetic technologies have made it feasible to have access to a wide variety of datasets associated with the detection of lung cancer. But unwanted or redundant features can act as a deteriorating factor that might reduce the performance level of classification methods. Therefore, appropriate feature selection methods are required to curate dimensionality, efficiency, and accuracy in cancer detection methods as well. Meta-heuristic search methods, inspired by nature, have proven to be a powerful aid in overcoming difficult optimization tasks, including feature selection problems in particular. Their stochastic nature makes them more apt to deal with health-related databases efficiently. In this context, this research aims to use two meta-heuristic optimization methods, Particle Swarm Optimization (PSO) and Genetic Algorithm (GA), to find out optimal feature sets from a lung cancer database to judge performance associated with feature sets identified by selecting suitable features that are gauged with K-Nearest Neighbours classification technique individually. Notably, PSO attained a classification accuracy of 97.82%, outperforming GA with 96.74%. These findings underscore the effectiveness of metaheuristic optimization-based feature selection in elevating the diagnostic performance of lung cancer classification systems.

Keywords
Feature Selection, Genetic Algorithm, Lung Cancer, Machine Learning, Particle Swarm Optimization, Firefly Algorithm.

1. Introduction

Cancer is a chronic disease that poses a significant threat to human life. It is usually discovered much later in the development of the disease and has a very high recurrence and fatality rate. Inhaling of smoke, family history, exposure to hazardous chemicals and atmospheric contamination are some causes of lung cancer. Lung and bronchus cancers are the second most prevalent cancer type in females after breast cancer, and the most frequently reported tumours in males after prostate cancer are lung and bronchus cancers. Early-stage lung cancer has no visible symptoms, so most lung cancer cases are only identified when it has severely advanced. Consequently, early lung cancer detection is the only way to increase a patient's health and life expectancy. One of the significant challenges for researchers is to extract meaningful information from data sets efficiently. This procedure involves separating relevant subsets of the original data while ignoring non-essential parts. Highly relevant features considerably impact the accuracy of classification models. Conversely, irrelevant features provide no beneficial information and fail to contribute to classification models. Thus, removing all irrelevant and extraneous features during the attribute selection mechanism is essential. The feature selection process includes four main steps, which are: subset generation, evaluation of created subsets based on a fitness function, formulation of a stopping criterion, and finally, result verification. Feature selection methods are classified into three categories. These categories include wrapper approaches, filter approaches, and embedded approaches. Wrapper approaches use a result obtained from a classifier to effectively search for optimal feature

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

subsets. They are referred to as wrappers because a search algorithm is used along with a classifier, which is treated as a black box. Usually, this methodology results in the identification of feature subsets that improve the value
of classification [1]. Well-known techniques in this category include Particle Swarm Optimization (PSO) [2], Genetic Algorithm (GA) [3]. Filter techniques are based on the integral characteristics of the data. They do not rely on a machine learning model. This approach is usually speedy because it selects features without running a learning algorithm. Examples include the Fisher Score, Relief and ReliefF, and the Gini Index. Embedded methods aim to balance the strengths of filter and wrapper methods by integrating machine learning algorithms with the exploration of fundamental data properties within a single method. Examples include multinomial logistic regression, random forest and CART.
In this paper, we implement the PSO and GA for the feature selection (FS) procedure. Each algorithm draws motivation from natural phenomena. The PSO algorithm reflects the flocking attitude of fish or birds searching for food; GA is motivated by natural selection and the principles of genetics from evolutionary biology. Evolutionary algorithms are heuristic search methods motivated by biological evolution [4].

This paper seeks to create a multi-stage lung cancer prediction model. The proposed model receives lung cancer patient data as input and first processes it through the processing stage. In this stage, outliers are removed, missing values are filled out if there are any, and the data is normalized. The following step will be the selection of features. This will include using metaheuristic algorithms to perform the selection. The PSO, GA, and FFA algorithms will be applied independently to the lung cancer dataset. This process will enable the reduction of features, hence minimizing dimensionality. At this point, valuable pieces of information will be preserved. The objective will be to find features that are most relevant. The last step will be to proceed and determine accuracy, classifying the dataset into malignant and non-malignant cases using K-Nearest Neighbours. This study makes the following primary contributions:

- Developing a lung cancer detection model.
- Generating the reduced feature subsets by applying PSO and GA during the feature selection process.
- Evaluating the performance of the obtained feature subsets by implementing the K-NN classifier.
- Evaluating the outcomes and demonstrating that the recommended method reduces feature dimensionality while increasing classification accuracy.

This paper is structured as follows: **Section 2** contains Literature Survey. In **Section 3**, the proposed method is discussed. Experimental results and their analysis are presented in **Section 4**. Finally, the conclusion and future work are presented in **Section 5**.

2. Literature Survey

Agarwal et al. [5] suggested a Fuzzy rule-based bPSO for FS by integrating fuzzy logic to enhance decision-making. This approach was explicitly useful for health-care data sets due to the importance of variation in features related to disease progression and population size. Khammassi et al. [6] suggested GA based wrapper FS for intrusion identification. The methodology achieved favorable results for large datasets. Wang et al. [7] gave a summary of PSO, stating challenges and improvements for FS of high-dimensional datasets. Kadam et al. [8] proposed improved GA with soft margin SVM based feature selection for enhancing classification accuracy for medical datasets. A novel optimization technique based on PSO was proposed by Huda et al. [9]. This approach improved exploration and avoided premature convergence, which is a common limitation for PSO algorithm. Shaheen et al. [10] projected the MinMaxScaler Binary PSO for FS. It showed promising results in handling imbalanced datasets. Ghosh et al. [11] presented a hybridized GA and PSO. This approach combined the

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

strengths of both algorithms in order to increase accuracy. A system review on the PSO algorithm was performed by Gad et al. [12] on the feature selection methods using PSO for high-dimensional datasets. A literature review on the applications and feature selection methods of radiomics, with emphasis on the applications of PSO and Genetic Algorithm for improving the lung cancer detection models is given by Zhang et al. [13]. For high-dimensional feature subsets and feature classification without losses in computation and convergence speed, the adaptive pyramid PSO (AP-PSO) technique for feature selection is proposed by Jin et al. [14]. This system had a multi-level structure that increased the overall convergence and diversity of solutions for all processes and subsets involved. A system named Guided Particle Adaptation PSO (GPA-PSO) for high-dimensional classification problems is completed by Huang et al. [15]. The approach included an adaptive learning procedure that enabled the fine-tuning of the search procedure of the swarm system. This provided additional optimal subsets of features with less idle time.

3. Materials and Methods

The key components of the current research include data acquisition and preprocessing, feature selection by nature-inspired algorithms and techniques, and further evaluating the number of features selected and the improvement in accuracy for classifying lung cancer.

3.1 Dataset

The proposed model used a dataset from Kaggle [16] to predict lung cancer. This dataset is readily available to the public. This dataset has information from 309 patients, including both individuals with cancer and without cancer. The dataset is based on 270 lung cancer patients and 39 healthy individuals.

The dataset comprises 15 independent variables and one dependent variable, classified into two classes: benign (non-cancerous) and malignant (cancerous). In order to prepare the data for prediction and training, the recommended model employed a range of methods to handle missing values, address outliers, and normalize the data.

3.2 Feature Selection

Dimensionality reduction as a part of pre-processing stage to machine learning is beneficial in eliminating unnecessary and redundant data, improving correct classification rate, and improving results to increase interpretability [17]. On the other hand, the recent escalation of dimensionality of data puts an extreme challenge to numerous current feature selection techniques with respect to efficacy and effectiveness. In the area of machine learning dimensionality reduction is significant domain, where many approaches have been proposed. In this proposed approach, three nature-inspired algorithms are used for feature selection. After applying feature selection, reduced feature subsets are obtained.

These obtained subsets of features are then evaluated using the K-NN classifier with the purpose of how effectively, these feature selection techniques can be used to achieve high performance in lung cancer diagnosis.

3.3 Particle Swarm Optimization (PSO)

The PSO algorithm is a stochastic optimization approach motivated by the collective behaviour of swarms. It was presented by Eberhart and Kennedy in 1995 [18].

The algorithm operates such that particles exhibiting higher velocities and positions closer to the food source are identified as the best candidates, while others with lower velocities are adjusted to enhance their performance. Each particle's velocity $v_i(t)$and position $x_i(t)$ are updated according to the optimal fitness values as expressed in Eq. (1) and (2):

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

$$v_i(t+1) = w\, v_i(t) + r_1 c_1 (P_{best,i}(t) - x_i(t)) + r_1 c_1 (G_{best}(t) - x_i(t)) \quad \ldots(1)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad \ldots(2)$$

In Eq. (2), r_1 and r_2 are random values distributed uniformly in [0,1], ω is the inertia weight, and c_1, c_2 are cognitive and social acceleration coefficients, respectively.

3.4 Genetic Algorithm (GA)

The GA starts by randomly generating the initial population X of m chromosomes[19]. Each chromosome's fitness is evaluated, and the chromosomes, C_1 and C_2, are chosen due to their fitness. A one-point breeding operator with breeding probability C_P is implented to generate an offspring$O$. The crossover rate R is defined as:

$$R = \frac{\gamma + 2\sqrt{\delta}}{3\gamma} \quad \ldots(3)$$

3.5 Classification

After applying feature selection using PSO and GA, the obtained feature subsets are further evaluated using the K-NN classifier. K-NN is a machine learning approach that works by finding the, K, nearest data points to a new data point and making predictions based on their values or labels.

3.6 Experimental Setup

In this study, MATLAB ® 2023, is used for the computational simulation. Experiments are performed using a lung dataset on a computer functioning on Windows 10 Pro, an x64-based processor, and a 64-bit operating system. The computer featured 8.00 GB RAM and a processor speed of 2.40 GHz. The optimal solutions were found using the K-nearest neighbour (K-NN) algorithm with K = 5 and Euclidean distance. The k parameter is determined by trial and error. Across all data sets, K = 5, steadily produced better results.

For examining the performance of the suggested methodology, metrics like the minimum number of features selected, the maximum number of features selected, the average value of a number of chosen features, minimum accuracy, maximum accuracy, and average accuracy are used.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

The methodology of proposed model is outlined in fig. 1.

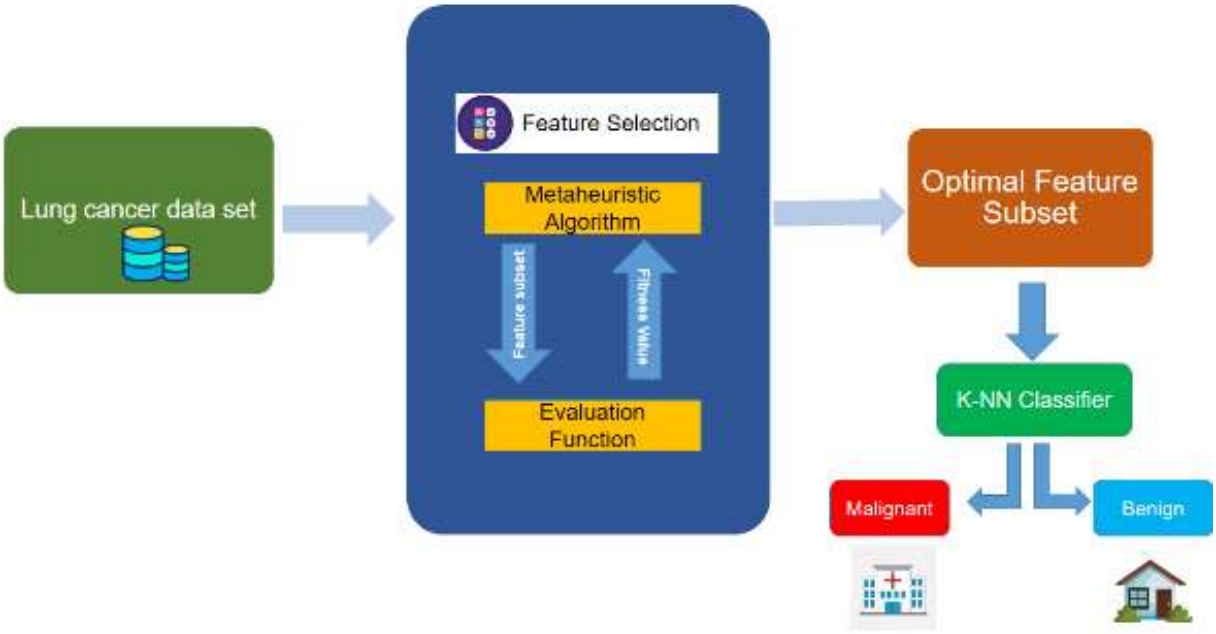

Fig. 1 Lung Cancer Classification Model

The table 1. Outlines the parameter values of PSO, GA and other common settings.

Table 1
Parameter settings

Algorithm	**Parameter**	**Value**
PSO	Cognitive Factor	2
	Social Factor	2
	Inertia Weight	0.9
GA	Crossover Rate	0.8
	Mutation Rate	0.01
Common Parameters	Alpha in fitness value	0.99
	Number of K in K-NN	5
	Ratio of validation data	0.3
	Lower Bound	0
	Upper Bound	1
	Number of Runs	20
	Problem Dimensions	No. of Features
	Maximum Iterations	100

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

4. Results and Discussion

The minimum, maximum and mean accuracy and number of features selected obtained by PSO, and GA on lung cancer data set is shown in table 2.

Table 2

Results

Algorithm	Accuracy	Value
PSO	Min.	91.30%
	Max.	97.82%
	Avg.	94.18%
GA	Min.	91.30%
	Max.	96.74%
	Avg.	94.18%
Algorithm	Feature Size	Value
PSO	Min.	5
	Max.	9
	Avg.	6.80
GA	Min.	4
	Max.	11
	Avg.	6.90

GA exhibits the broadest range in terms of selecting features. The number of features chosen ranged from 4 to 11 features, showing the flexibility of the GA in inspecting different feature subsets. The average value of a number of features chosen by the GA is 6.90. The range shows the flexibility of the Genetic Algorithm in inspecting different feature subsets. The PSO selects features ranging from 5 to 9, implying consistency in selecting smaller feature subsets. The mean number of features chosen by PSO is 6.80, respectively. Overall GA shows diverse selection behaviour, whereas the PSO proposed moderate variability.

The comparative analysis of proposed method with other methodologies already presented in literature is shown in table 3.

Table 3

Comparative Analysis

Category	Methodology	Accuracy
Proposed Algorithm	PSO	97.82%
	GA	96.74%
Kabir et al. [20]	PSO with Gradient Boost	93.00%
	PSO with Random Forest	92.00%
Sachdeva et al. [21]	Naïve Bayes	95.16%
	K-NN	91.93

5. Conclusion

This work focused on improving lung cancer classification by using two nature-inspired optimization methods, PSO and GA to determine an optimal subset of features from a lung cancer dataset. These obtained features were then used as an input for the K-Nearest Neighbours (K-NN) classifier, aiming to improve accuracy while reducing data complexity. This approach provides significant evaluation by comparing well-known

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

feature selection methods and examining their direct impact on classification performance. This approach will benefit the scientific community by allowing them to compare optimization methods, which are mostly reviewed individually, side by side. This approach will help researchers and medical personnel make well-informed decisions while designing diagnostic systems for medical datasets. In the set of investigated algorithms, PSO produced the best result, achieving an accuracy of 97.82%, which clearly proved its efficiency in feature identification for classification problems related to lung cancer. For the genetic algorithm, it achieved an accuracy of 96.74%. This study has some limitations, as it has been conducted on one dataset only, not on general applicability. For future studies, it is hoped that this approach will be conducted on more than one dataset, as well as investigating adaptive optimization methods for improving feature selection.

References

[1]. D. Bajer, M. Dudjak, and B. Zorić, "Wrapper-based feature selection: how important is the wrapped classifier," *Proc. Int. Conf. Smart Syst. Technol. (SST)*, pp. 97–105, 2020, doi: 10.1109/SST49455.2020.9264072.

[2]. P. Moradi and M. Gholampour, "A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy," *Appl. Soft Comput.*, vol. 43, pp. 117–130, 2016, doi: 10.1016/j.asoc.2016.02.002.

[3]. J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intell. Syst. Appl.*, vol. 13, no. 2, pp. 44–49, 1998, doi: 10.1109/5254.671091.

[4]. Y. Zhang, C. Zhu, S. Tang, Y. Ran, D. Z. Du, and Z. Zhang, "Evolutionary algorithm on general cover with theoretically guaranteed approximation ratio," *INFORMS J. Comput.*, vol. 36, no. 2, pp. 510–525, 2024, doi: 10.1287/ijoc.2022.0327.

[5]. S. Agarwal, R. Rajesh, and P. Ranjan, "FRBPSO: A fuzzy rule-based binary PSO for feature selection," *Proc. Natl. Acad. Sci. India Sect. A Phys. Sci.*, vol. 87, pp. 221–233, 2017, doi: 10.1007/s40010-017-0347-8.

[6]. C. Khammassi and S. Krichen, "A GA-LR wrapper approach for feature selection in network intrusion detection," *Comput. Secur.*, vol. 70, pp. 255–277, 2017, doi: 10.1016/j.cose.2017.06.005.

[7]. D. Wang, D. Tan, and L. Liu, "Particle swarm optimization algorithm: an overview," *Soft Comput.*, vol. 22, pp. 387–408, 2017, doi: 10.1007/s00500-016-2474-6.

[8]. V. J. Kadam, S. S. Yadav, and S. M. Jadhav, "Soft-margin SVM incorporating feature selection using improved elitist GA for arrhythmia classification," in *Proc. Int. Conf. Intell. Syst. Des. Appl. (ISDA)*, Vellore, India, Dec. 6–8, 2018, vol. 2, pp. 965–976, Springer, 2020, doi: 10.1007/978-3-030-16660-1_94.

[9]. R. K. Huda and H. Banka, "New efficient initialization and updating mechanisms in PSO for feature selection and classification," *Neural Comput. Appl.*, vol. 32, no. 8, pp. 3283–3294, 2020, doi: 10.1007/s00521-019-04395-3.

[10]. H. Shaheen, S. Agarwal, and P. Ranjan, "Ensemble maximum likelihood estimation based logistic MinMaxScaler binary PSO for feature selection," in T. K. Sharma, C. W. Ahn, O. P. Verma, and B. K. Panigrahi, Eds., *Soft Computing: Theories and Applications*, Advances in Intelligent Systems and Computing, vol. 1380, Springer, Singapore, 2022, doi: 10.1007/978-981-16-1740-9_58.

[11]. M. Ghosh, R. Guha, I. Alam, P. Lohariwal, D. Jalan, and R. Sarkar, "Binary genetic swarm optimization: a combination of GA and PSO for feature selection," *J. Intell. Syst.*, vol. 29, no. 1, pp. 1598–1610, 2020, doi: 10.1515/jisys-2019-0062.

[12]. A. G. Gad, "Particle swarm optimization algorithm and its applications: a systematic review," *Arch. Comput. Methods Eng.*, vol. 29, pp. 2531–2561, 2022, doi: 10.1007/s11831-021-09694-4.

[13]. Y. Zhang, S. Wang, and G. Ji, "A comprehensive survey on particle swarm optimization algorithm and its applications," *Math. Probl. Eng.*, vol. 2015, Article ID 931256, 38 pages, 2015, doi: 10.1155/2015/931256.

[14]. G. Ge and J. Zhang, "Feature selection methods and predictive models in CT lung cancer radiomics," *J. Appl. Clin. Med. Phys.*, vol. 24, e13869, 2023, doi: 10.1002/acm2.13869.

[15]. X. Jin, B. Du, Y. Zhang, S. Li, and J. Wu, "An adaptive pyramid PSO for high-dimensional feature selection," *Expert Syst. Appl.*, vol. 257, 125084, 2024, doi: 10.1016/j.eswa.2024.125084.

[16]. S. G. Nelson, "Lung cancer prediction," Kaggle, 2025. [Online]. Available: https://www.kaggle.com/code/sandragracenelson/lung-cancer-prediction [Accessed: Apr. 2, 2025].

[17]. S. Azeem, S. Javed, I. Naseer, O. Ali, and T. M. Ghazal, "A new hybrid PSO-HHO wrapper based optimization for feature selection," *IEEE Access*, vol. 13, pp. 87090–87099, 2025, doi: 10.1109/ACCESS.2025.3570901.

[18]. R. Eberhart and J. Kennedy, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Netw.*, Perth, Australia, 1995, pp. 1942–1948.

[19]. J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Ann Arbor, MI: Univ. Michigan Press, 1975.

[20]. S. R. R. Kabir, H. Mondal, A. Nag, S. M. H. Jamil, and P. Das, "Identification of lung cancer using particle swarm optimization and machine learning technique," in *Proc. Int. Conf. Trends Electron. Health Informatics*, pp. 421–430, Springer, 2023, doi: 10.1007/978-981-97-3937-0_29.

[21]. R. K. Sachdeva, P. Bathla, P. Rani, et al., "A novel K-nearest neighbor classifier for lung cancer disease diagnosis," *Neural Comput. Appl.*, vol. 36, pp. 22403–22416, 2024, doi: 10.1007/s00521-024-10235-w.