JCBI
Journal of Computing & Biomedical Informatics Volume 09 Issue 02
ISSN: 2710 - 1606 2025

Research Article
https://doi.org/10.56979/902/2025

An Optimized Multistage Model for Lung Cancer Prediction with Machine Learning

Shazia Javed[1], Layla Naz[1], Sumbul Azeem[1*], and Uzma Bashir[1]

[1]Department of Mathematics, Lahore College for Women University (LCWU), Lahore, Punjab, Pakistan.
[*]Corresponding Author: Author's Name. Email: sumbul.azeem@lcwu.edu.pk

Received: July 06, 2025 **Accepted:** August 21, 2025

Abstract: Lung cancer accounts for a significant share of cancer-related deaths globally. It is one of the most prevalent and fatal illnesses. This paper presents an enhanced machine learning framework for diagnosing lung cancer. To to eliminate redundant and irrelevant features, a feature selection method inspired by natural phenomena called firefly algorithm is employed. Lung cancer patients are subsequently categorized using the K-Nearest Neighbors algorithm based on specific characteristics. The goal of the proposed model is to maximize computational efficiency while maintaining high diagnostic accuracy. Calculation metrics, such as accuracy, the number of features chosen, and computing efficiency, will be used to evaluate the model. This study supports improved treatment planning, better patient outcomes, and advances in early detection. Additionally, combining feature selection with classification reduces overfitting and improves prediction accuracy. Incorporating medical knowledge with computational intelligence provides a sustainable approach to developing scalable and clinically useful diagnostic systems. Experimental results show that the proposed model obtained 96.73% best accuracy. If applied in medical diagnosis systems, the suggested model can significantly increase the accuracy of lung cancer detection leading to operational treatment and an increase in patient survival rates.

Keywords: Feature Selection; Firefly Algorithm; K-Nearest Neighbors ; Lung Cancer; Machine Learning

1. Introduction

Lung cancer develops when lung cells multiply and grow out of control, creating tumors that may potentially spread to other regions of the body. Smoking is a significant risk factor, but anyone can develop lung cancer. Lung cancer is often not caught at an early stage. Instead, it is usually diagnosed when the tumor itself causes local problems, such as coughing and chest pain, or when it has spread to other parts of the body. The most often reported cancers in men after prostate cancer are bronchus and lung and cancer. In women, lung and bronchus cancers are the second most predominant cancer type after breast cancer. Since early-stage lung cancer usually has no noticeable indications, most lung cancer cases are only diagnosed when it has advanced rigorously. Miserably, the prediction for advanced-stage lung cancer is largely grim, with about 1 out of 5 patients living five years. Thus, early identification of lung cancer is the most crucial step in treatment. Early recognition can escalate a patient's well-being and lifespan.

The most crucial stage in creating a dataset for machine learning (ML) models is the feature selection (FS). It is based on the selection of critical and relevant characteristics of a dataset to improve the model's performance, accuracy, and explicability. By choosing only the most useful features or characteristics of a dataset, one can significantly reduce or handle the complex nature of the data, remove noise, and prevent overfitting. There are three types of feature selection methods. The Filter Selection Method involves evaluating the correlation of each feature with the target variables, considering only relevant features, and

discarding irrelevant ones. Wrapper–based methods rely on a search strategy to compute optimal subsets of features. In this technique, the search algorithms are wrapped around a classifier. Embedded methods include the feature selection step in the model construction process.

Irrespective of modernizations in medicinal imaging and investigative measures, well-timed recognition of lung cancer remains a key challenge, leading to extraordinary death rates. Metaheuristic and machine learning-based feature selection approaches have shown aptitude in improving classification accuracy. Nevertheless, prevalent practices face plentiful limitations. These comprise high computational costs, suboptimal feature selection, and a lack of generalizability across different datasets. This research aims to address these challenges by proposing a Firefly algorithm-based feature selection method that enhances predictive accuracy. Based on fireflies' flashing characteristic, the Firefly Algorithm (FA) [1] is an optimization technique inspired by nature. Created in 2010, fireflies utilize bioluminescence to communicate and attract mates. The algorithm uses this to solve various optimization problems optimally.

By efficiently selecting the most relevant attributes and balancing the dataset, this study aims to contribute to more accurate and consistent lung cancer classification, thereby facilitating timely diagnosis and treatment. The study uses a lung cancer dataset comprising 309 instances. The dataset initially includes 15 features such as GENDER, WHEEZING COUGH, YELLOW FINGERS, and others. The firefly algorithm is applied to identify the most relevant features. After feature selection, the data is then partitioned into 75% for training and 25% for testing, and K-NN techniques are employed to train and evaluate the model. The performance of the proposed method is evaluated using metrics, including accuracy, the number of features selected, and computational time.

The main objectives of this research are:

· To develop a framework for lung cancer identification.

· To apply the firefly algorithms for feature selection and optimize the feature selection process.

· To analyze the performance of the suggested framework using the K-NN classifier.

This article is divided as follows: Section II is based on a literature review, Section III shows research methodology, Section IV presents the results and discussion, and Section V concludes the paper.

2. Literature Survey

Numerous studies have explored ML and optimization techniques for lung cancer diagnosis and image analysis. Senthil et al. [2] applied segmentation using five optimization algorithms and found that GCPSO produced the highest accuracy (95.89%), demonstrating the potential of advanced swarm-based methods. Similarly, Patra et al. [3] showed that ML classifiers could perform reasonably well on UCI lung cancer data, with RBF achieving 81.25% accuracy, though performance remained limited compared to more recent approaches.

Lavanya et al. [4] investigated lung nodule segmentation and classification through an integrated approach combining preprocessing, FA-FCM clustering, and SVM classification. While effective for accurate nodule detection, the study placed greater emphasis on methodological detail than on comparative performance with other state-of-the-art techniques. Dutta et al. [5] advanced the field by integrating IoT-enabled CNN feature extraction with Random Forest classification, achieving 93.25% accuracy. However, the reliance on IoT infrastructure limits its adaptability in purely clinical settings.

Other works have applied diverse ML classifiers. Ojha et al. [6] compared several algorithms and reported Logistic Regression as the most effective (94.7% accuracy), highlighting that traditional statistical models can still be competitive under certain conditions. Gopinath et al. [7] demonstrated that Firefly-optimized transfer learning achieved superior diagnostic accuracy (98.5%) compared to well-established deep learning models such as DenseNet and ResNet, suggesting that optimization strategies can significantly enhance deep learning architectures. Similarly, Slama et al. [8] combined pre-trained VGG19 with SVM for chest X-ray classification, attaining 93.5% accuracy and outperforming standard CNN models.

Research has also focused on feature engineering and hybrid approaches. Mony et al. [9] applied feature selection via LASSO and PCA on the PLCO dataset, where ensemble learning delivered strong results (97.66% accuracy), reinforcing the importance of balancing feature reduction with model complexity. Azeem et al. [10] proposed a hybrid PSO-HHO framework for feature selection in lung cancer datasets, demonstrating the advantages of hybrid metaheuristic approaches over conventional methods.

Likewise, Aftab et al. [11] combined advanced CNN structures with attention mechanisms and optimization-based feature selection, reaching 95.0% accuracy on CT images, further validating the role of hybrid deep learning in early cancer detection.

Generally, current studies establish considerable progress in lung cancer detection using optimization, ML, and deep learning approaches. Still, many works either focus heavily on methodological details without critical comparisons, or they emphasize accuracy without addressing computational efficiency, or feature reduction across datasets. These gaps highlight the need for hybrid optimization-based feature selection approaches that balance interpretability, efficiency, and accuracy an area where the proposed technique targets to contribute.

3. Materials and Methods

Firefly algorithm is nature-inspired optimization method which clones the flashing patterns of fireflies. FA interprets this behavior in an optimization framework, with each firefly symbolizing a possible solution, and its luminosity reflecting the quality, or fitness of that solution as per a specified objective function. A mathematical formula that postulates the attraction of a single firefly to another based on relative brightness powers the Firefly Algorithm's movement mechanism. The procedure integrates a number of factors, such as the solution's attractiveness, the firefly's closeness to one another, and a certain amount of impulsiveness to save the ability to explore. The fundamental equation principal the movement of firefly toward a brighter firefly is given as:

$$x_i = x_i + \beta e^{-\gamma r_{ij}^2}\left(x_j - x_i\right) + \alpha\epsilon$$

Where

- x_i and $x_j$is the current position and brighter position of firefly i and j respectively.
- β is the base attractiveness at $r = 0$
- γ is the Light absorption coefficient, controlling the decrease in attractiveness with distance.
- $r_{ij}$is the Euclidean distance between fireflies i and j.
- α is the random parameter.
- The random vector, typically denoted by ϵ taken from a Gaussian or uniform distribution.

The separation between fireflies can be defined as:

$$r_{ij} = \left|\left|x_i - x_j\right|\right|$$

This distance influence the degree in which one firefly is drawn to another. The attractiveness term $\beta e^{-\gamma r_{ij}^2}$ decreases exponentially with the square of the distance, ensuring that fireflies closer to each other have stronger interactions, while those farther apart exert weaker influence. This mimics natural behavior and helps preserve local search focus in dense solution regions.

The process of lung cancer classification begins with lung cancer data set [12]. The lung cancer data set is preprocessed at initial stage. Outliers are removed. Missing values are filled. After this. Fire fly algorithm is applied to choose optimal feature subset. Then for classification K-NN are employed. A K-NN classifier is used to find the accuracy of the solution obtained using the feature selection model. In the experiments, k = 5 is chosen for optimal performance, following the procedure adopted from [13] and [14].

This whole process is repeated twenty times and accuracy, number of features selected and time (in seconds) taken by the algorithm has been evaluated. This process is presented in figure below.

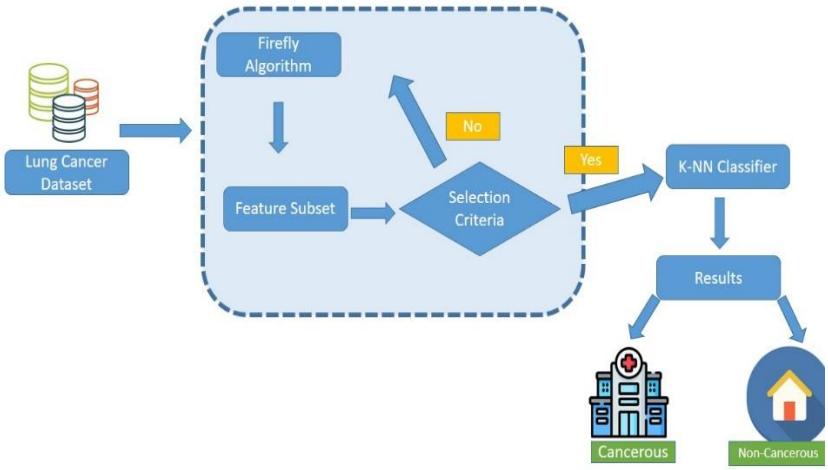

Figure 1. Lung Cancer Prediction Model

4. Results

In this methodology, the simulation work is executed using MATLAB R2023a. The suggested model used K-NN for classification. In Table 1 shows the number of runs where the run no. represents the index of each independent execution of the proposed method, the accuracy obtained in corresponding run, the time taken in that run and the number of features selected in that run.

Table 1. Accuracy, Time And Number Of Features Chosen During Lung Cancer Classification.

Run No.	Accuracy	Time (in sec.)	No. of features selected
1	0.9130	12.51	5
2	0.9239	26.91	5
3	0.9239	17.07	5
4	0.9347	25.05	5
5	0.9347	17.17	5
6	0.9456	17.56	5
7	0.9456	14.99	6
8	0.9565	16.03	6
9	0.9565	41.54	6
10	0.9565	20.79	7
11	0.9565	20.75	7
12	0.9565	20.49	7
13	0.9565	14.43	7
14	0.9565	21.22	7
15	0.9673	18.08	7
16	0.9673	24.93	7
17	0.9673	17.14	7
18	0.9673	11.77	8
19	0.9673	13.87	8
20	0.9673	17.69	9
Average	0.9510	19.50	6.45
SD	0.02	6.62	1.19
Maximum	0.9673	41.54	9
Minimum	0.9130	11.77	5

From table 1 it is observed that the best accuracy of fire fly algorithm is 96.73 % and the best time taken by algorithm is 11.77 seconds. The minimum number of features selected by algorithm is 5. Table 1 also shows maximum accuracy, minimum accuracy, average accuracy, maximum time, minimum time, and average time, maximum numbers of features selected, minimum numbers of features selected and average

number of features selected. From table 1 it is noted that minimum accuracy achieved by proposed methodology is 91.30 % and maximum accuracy achieved by algorithm is 96.74 %. Average accuracy obtained by the suggested method is 95.11 %. Also the minimum and maximum time Taken for the proposed methodology is 11.77 seconds and 41.54 seconds the average time taken by the whole procedure is 19.50 seconds. The minimum number of features selected by the suggested methodology is five and the maximum number of features selected by the Suggested methodology is 9 and average number of features selected by the suggested process is 6.45.

The accuracy exhibited a very low SD of ±0.0167, representing that the classification results were very steady across multiple runs. The computational time showed a higher SD of 6.62 seconds, which can be credited to the stochastic nature of metaheuristic optimization, where different runs may converge at varying speeds. The number of features selected showed an SD of 1.19, reflecting minor fluctuations in the subsets chosen by the optimizer. These variations are expected in evolutionary search processes and, importantly, do not compromise accuracy, which remained consistently high across all runs.

Table 2 shows comparative analysis of the proposed method and the method already existed in literature. The suggested algorithm achieves better accuracy than the algorithms mentioned in the table below.

Table 2. Comparative Analysis

Algorithm	Accuracy
Proposed Algorithm	96.73%
RBF classifier[3]	81.25%
CNN and Random Forest [5]	93.25%
Logistic Regression [6]	94.70%

Fig. 2 shows convergence curve of firefly algorithm across twenty runs.

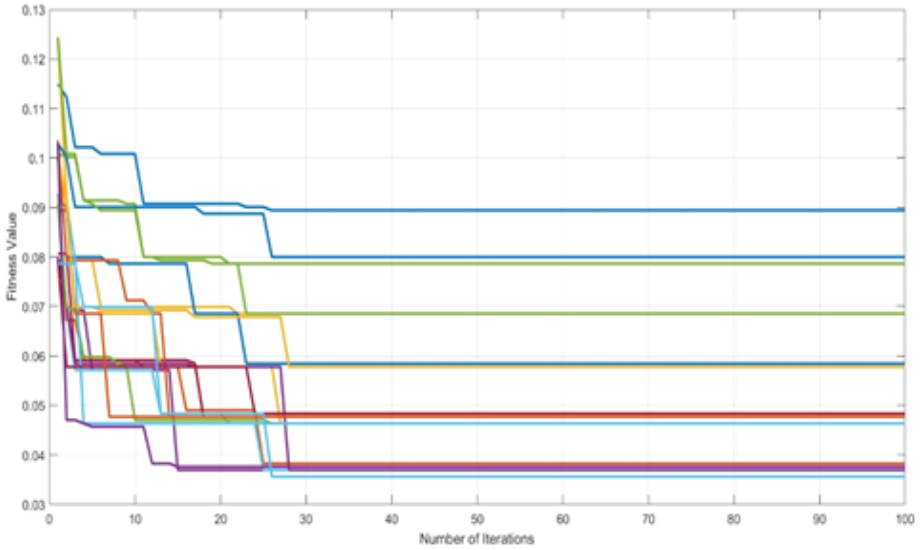

Figure 2. Convergence curve across 20 runs

Fig 3 shows graph of run number vs accuracy, time and number of featuers chosen during the algorithm.

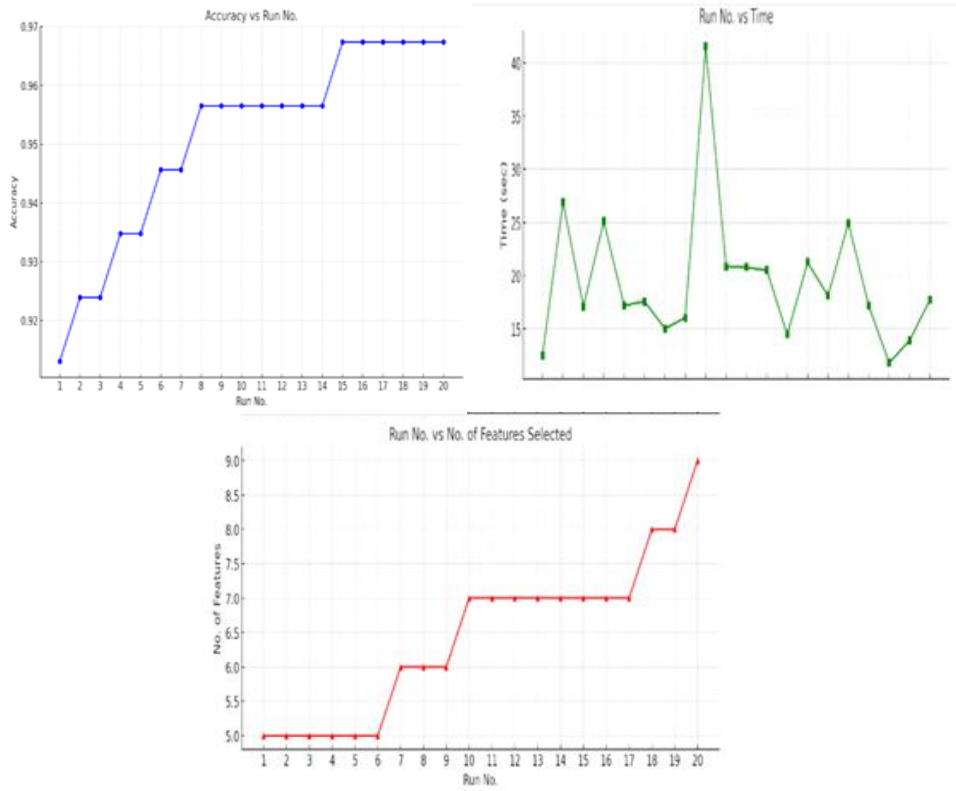

Figure 3. Graph of Run Numbers vs Accuracy, Time and Numbers of Features

5. Conclusions

Lung cancer continues to pose a major risk to worldwide public health, frequently identified in late stages when treatment alternatives are scarce and outcomes are bleak. The intricate and varied characteristics of lung cancer, coupled with elevated mortality rates, highlight the critical demand for smart, effective, and precise diagnostic systems capable of identifying the disease during its initial stages. Combining machine learning with optimization methods inspired by nature offers a promising solution to tackle these issues. This study focused on creating and validating an increase in machine learning structure for lung cancer diagnosis, using nature-inspired algorithms, especially the Firefly Algorithm (FA), to enhance feature selection and classification precision. Several critical insights emerged from this study. The use of Firefly algorithm for feature selection significantly increased classification accuracy across all machine learning models, reducing misclassification rates. Firefly algorithm effectively reduced the original feature space, enhancing model interpretability and reducing computational time. The framework performed consistently across different classifiers, indicating strong potential for generalization across algorithms and datasets. When benchmarked against other nature-inspired algorithms and feature selection techniques, the firefly-based system demonstrated competitive or superior results in key performance metrics. We aim to further develop this research in the future to forecast lung cancer utilizing additional patient data and various biomarkers using transfer learning and fuzzification. Also, we plan to extend the evaluation to multiple publicly available and real-world datasets, perform cross-validation, and collaborate with clinical experts to assess the model's practical applicability.

Funding: This research received no external funding.
Data Availability Statement: Data is available online [12].
Conflicts of Interest: The authors declare no conflict of interest

References

1. Yang, X.-S. Firefly algorithm, stochastic test functions and design optimisation. Int. J. Bio-Inspired Comput. 2010, 2(2), 78–84.
2. Senthil Kumar, K.; Venkatalakshmi, K.; Karthikeyan, K. Lung cancer detection using image segmentation by means of various evolutionary algorithms. *Computational and Mathematical Methods in Medicine* 2019, 2019(1), 4909846. doi: 10.1155/2019/4909846.
3. Patra, R. Prediction of lung cancer using machine learning classifier. In Computing Science, Communication and Security. COMS2 2020. Communications in Computer and Information Science; Chaubey, N., Parikh, S., Amin, K., Eds.; Springer: Singapore, 2020; Vol. 1235, pp. 123–134. doi: 10.1007/978-981-15-6648-6_11.
4. Lavanya, M.; Kannan, P.M.; Arivalagan, M. Lung cancer diagnosis and staging using firefly algorithm fuzzy C-means segmentation and support vector machine classification of lung nodules. *Int. J. Biomed. Eng. Technol.* 2021, 37(2), 185–200. doi: 10.1504/IJBET.2021.119504.
5. A. K. Dutta, "Detecting lung cancer using machine learning techniques," *Intell. Autom. Soft Comput.*, vol. 31, no. 2, 2022.
6. Ojha, T.R. Machine learning based classification and detection of lung cancer. *Journal of Artificial Intelligence and Capsule Networks* 2023, 5(2), 110–128.
7. Gopinath, A.; Gowthaman, P. Transfer learning and optimised firefly neural network for lung cancer. *J. Intell. Syst. Internet Things* 2024, 13(2). doi: 10.54216/JISIoT.130213.
8. Slama, A.B.; Amri, Y.; Barbaria, S.; Rahmouni, H.B.; Trabelsi, H. Lung diseases classification using pre-trained based deep learning model and support vector machine. *Polish Journal of Medical Physics and Engineering* 2025, 31(3), 178–194. doi: 10.2478/pjmpe-2025-0021.
9. Mony, M.J.I.; Chowdhury, M.J.U. Improving lung cancer prediction with SMOTE, LASSO, and PCA: A data-driven machine learning framework. 2025.
10. Azeem, S.; Javed, S.; Azeem, S. Hybrid PSO-HHO Based Machine Learning Optimization for Lung Cancer Diagnosis. *Proceedings of the 2025 International Conference on Emerging Technologies in Electronics, Computing, and Communication (ICETECC)*, pp. 1–6, IEEE, 2025.
11. Aftab, J.; Khan, M.A.; Arshad, S.; Rehman, S.U.; AlHammadi, D.A.; Nam, Y. Artificial intelligence-based classification and prediction of medical imaging using a novel framework of inverted and self-attention deep neural network architecture. *Scientific Reports* 2025, 15, 8724.
12. Lung Cancer Prediction Dataset. Available https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer (accessed-on-15 July 2025).
13. Azeem, S.; Javed, S.; Naseer, I.; Ali, O.; Ghazal, T.M. A new hybrid PSO-HHO wrapper based optimization for feature selection. *IEEE Access* 2025, 13, 87090–87099, doi: 10.1109/ACCESS.2025.3570901.
14. Q. Al-Tashi, S. J. Abdul Kadir, H. M. Rais, S. Mirjalili, and H. Alhussian, ''Binary optimization using hybrid grey wolf optimization for fea ture selection,'' *IEEE Access*, vol. 7, pp. 39496–39508, 2019, doi: 10.1109/ACCESS.2019.2906757.