International Research Journal of Multidisciplinary Scope (IRJMS), 2026; 7(1): 1666-1678

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2026.v07i01.08680

Metaheuristic Optimization of Random Forest for Lung Cancer Prediction

Balaji T[1]*, Babu P[2], Lokeshwaran K[3]

[1]Department of Computer Science and Engineering (Data Science), Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India, [2]Department of Information Technology, PSNA College of Engineering and Technology, Dindigul, Tamil Nadu, India, [3]Department of Computer Science and Engineering, School of Computing, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Tamil Nadu, India. *Corresponding Author's Email: baalaji24@gmail.com

Abstract

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, where early risk identification is critical for improving survival outcomes. While existing machine learning approaches for lung cancer prediction frequently rely on medical imaging, such methods are costly and often impractical in low-resource clinical settings. This study proposes an efficient and interpretable lung cancer risk prediction framework using demographic, lifestyle, and symptom-based data. A Genetic Algorithm (GA) is employed as a metaheuristic optimization strategy to jointly perform feature selection and hyperparameter tuning of a Random Forest (RF) classifier. To address the inherent class imbalance in the dataset, the Synthetic Minority Oversampling Technique (SMOTE) is applied exclusively to the training data to prevent information leakage and enhance minority-class learning. The proposed GA-optimized RF model is evaluated against several baseline classifiers, including Logistic Regression, Support Vector Machine, k-Nearest Neighbors, Decision Tree, standard Random Forest, and XGBoost, using accuracy, precision, recall, F1-score, and ROC-AUC as evaluation metrics. Experimental results demonstrate that the optimized RF model achieves superior performance, with an accuracy of 90.6%, an F1-score of 0.885, and a ROC-AUC of 0.917, outperforming all baseline models. Feature importance analysis identifies smoking habit, breathing difficulties, and throat discomfort as the most influential predictors, aligning with established clinical knowledge. The findings highlight that a metaheuristic-driven optimization approach applied to non-imaging data can provide a cost-effective, reliable, and interpretable solution for early lung cancer risk screening, particularly in resource-constrained healthcare environments.

Keywords: Demographic Data, Feature Selection, Genetic Algorithm, Lung Cancer Prediction, Metaheuristic Optimization, Random Forest.

Introduction

Lung cancer remains one of the chief causes of cancer-related deaths worldwide. In many countries, survival rates have improved only marginally because the disease is frequently detected at advanced stages when therapeutic options are limited and outcomes poorer. Consequently, early risk identification is crucial: timely detection permits targeted screening, earlier intervention, and better allocation of medical resources. Historically, diagnosis and risk stratification have relied on imaging modalities such as low-dose computed tomography (LDCT), histopathology, and molecular tests (1). While effective, these methods are costly and dependent on specialised equipment and trained personnel. In resource-limited settings, demographic, lifestyle, and symptom information often remain the only accessible inputs (2). Hence, there is a pressing need for accurate and interpretable prediction systems that function well with such non-imaging features. Machine learning (ML) methods have shown great promise in clinical risk modelling across various diseases. ML models trained on imaging data for lung cancer achieve impressive discriminative performance; however, their high cost and dependency on imaging infrastructure limit their widespread use (3). Alternatively, models using demographic and symptom features such as age, smoking history, occupational exposures, comorbidities, and patient-reported symptoms offer cost-effective and scalable screening tools. Building robust models on these features is challenging. First, feature selection is imperative to reduce redundancy and noise, improving both predictive accuracy and interpretability. This has been

(Received 16th October 2025; Accepted 13th January 2026; Published 31st January 2026)

extensively reviewed in medical applications emphasizing feature selection as a key step for clinical machine learning (4, 5). Second, class imbalance where negative cases significantly outnumber positive lung cancer cases poses challenges during training, as naive classifiers may bias towards the majority class. Third, hyperparameter tuning critically influences model performance; improper tuning can cause overfitting or underfitting, undermining generalisability (6). Recent comparative studies underscore the importance of effective hyperparameter optimisation methods.

Metaheuristic algorithms, particularly Genetic Algorithms (GAs), offer practical solutions for simultaneous feature selection and hyperparameter tuning. GAs navigates large, complex search spaces effectively without requiring gradient information, making them highly suitable for combinatorial optimization problems encountered in clinical ML pipelines (7). Representing feature subsets as binary chromosomes alongside encoded continuous hyperparameters allows unified optimisation, potentially discovering synergistic combinations that sequential methods might miss (8, 9). Addressing class imbalance, synthetic oversampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) enhance minority class representation by generating realistic synthetic samples, facilitating fairer decision boundaries. Random Forest (RF) classifiers naturally complement this framework: they tolerate noisy features, accommodate mixed data types, resist overfitting via ensemble averaging, and provide interpretable feature importance metrics—attributes valued by clinicians for transparency (10).

This study aims to develop an accurate and interpretable lung cancer risk predictor using demographic and symptom data under real-world constraints such as limited samples, mixed data types, and class imbalance. The primary objective is to learn a mapping from patient features to binary risk labels (high vs. low), simultaneously selecting an informative feature subset and optimising classifier hyperparameters to maximise predictive performance via cross-validation and testing. The motivations behind this work are threefold: demographic and symptom data are inexpensive and widely obtainable in primary care, enabling extensive screening where imaging is impractical; metaheuristic dual optimisation automates crucial design decisions, potentially enhancing model performance beyond manual tuning; and the interpretability of RF ensembles facilitates clinician understanding and patient communication regarding risk factors.

Specific goals include: (a) designing a GA encoding to jointly represent feature selection and RF hyperparameters, evolving near-optimal solutions under cross-validation; (b) evaluating SMOTE's impact on minority detection and predictive metrics with properly applied oversampling; (c) benchmarking the GA-optimised RF against baseline classifiers (Logistic Regression, SVM, k-NN, Decision Tree, standard RF, XGBoost) using Accuracy, Precision, Recall, F1-score, and ROC-AUC; (d) analysing the final features to provide clinical insights into influential demographic and symptom predictors. Several prior studies have explored lung cancer risk prediction using demographic, lifestyle, or questionnaire-based data, typically employing conventional classifiers such as logistic regression, support vector machines, or tree-based ensembles with predefined feature sets. While these models demonstrate reasonable predictive performance, most rely on manual or filter-based feature selection and standard hyperparameter tuning strategies. Recent ensemble-based approaches using Random Forest or gradient boosting have reported improved discrimination; however, optimization is often performed sequentially or limited to either feature selection or parameter tuning. In contrast, the present study introduces a unified genetic algorithm-driven framework that simultaneously optimizes feature subsets and classifier hyperparameters under balanced training conditions. This joint optimization enables the discovery of synergistic features–parameter configurations, leading to improved screening-level performance while retaining interpretability based on clinically meaningful demographic and symptom variables.

The literature relevant to this study spans three principal areas: (a) metaheuristic algorithms for feature selection and hyperparameter tuning, (b) methods to handle class imbalance (with emphasis on SMOTE and its variants), and (c) machine learning approaches for lung cancer prediction using demographic/tabular features and complementary imaging-based studies that provide context and performance baselines.

Metaheuristic Algorithms for Feature Selection and hyper Parameter Optimisation

Genetic Algorithms (GAs) and other evolutionary methods have a long history in computational optimization and have been applied extensively to feature selection. Guyon and Elisseeff provided a foundational perspective on variable and feature selection, outlining filter, wrapper and embedded strategies, and pointing out the utility of search-based wrappers in domains where subset interactions matter (11). In healthcare analytics, wrapper methods using GAs have been shown to effectively reduce dimensionality while preserving classification performance, for example in microarray and clinical datasets where combinatorial interactions are important (12). Recent comparative studies show that GAs often outperforms simple greedy or filter approaches when the feature space is moderately large and features interact in nonlinear ways (13). Hybrid GA frameworks that combine filters for initial pruning and a GA wrapper for fine search have been proposed to make the search tractable on higher-dimensional data; such hybrids strike a balance between computational cost and selection quality (14). In parallel, more sophisticated evolutionary schemes (multi-objective GA, island models, and adaptive mutation) have been adapted to maintain population diversity and to avoid premature convergence in complex search landscapes (15). The idea of applying metaheuristics for hyper parameter tuning of ensemble models is also well explored. Evolutionary algorithms can tune both discrete and continuous hyper parameters jointly and are particularly useful when the search space is multi-modal or non-differentiable. Studies comparing GA, particle swarm optimisation (PSO), Bayesian optimisation and random/grid search show that no single method dominates universally; however, GA provides strong exploration capability and flexible encoding of mixed-type parameters, which makes it attractive for joint feature-hyper parameter optimisation in wrapper settings (16, 17). Application-oriented works demonstrate that GA-tuned Random Forests or boosted trees often achieve better generalisation than models tuned by naive grid search, especially when combined with cross-validated fitness metrics that penalise complexity (18).

Several medical domain studies are notable: GA wrappers for ECG and EEG feature selection, and GA-guided hyper parameter search for ensemble classifiers in cardiovascular and pulmonary disease prediction show consistent performance gains, particularly in sensitivity or recall for minority classes (19, 20). These examples illustrate the potential of a GA to simultaneously trim irrelevant features and find stable model settings that generalise well.

Class Imbalance Handling: SMOTE and Enhancements

Class imbalance is a persistent issue in medical datasets. The SMOTE algorithm remains a standard technique: it generates synthetic minority class samples along the line segments joining minority class neighbours and has been shown to enhance classifier sensitivity in numerous studies (21). Since its introduction, several SMOTE variants have been proposed to reduce generation of noisy examples and to pay attention to borderline or hard-to-classify minority examples: Borderline-SMOTE focuses synthesising samples near classification boundaries; ADASYN adapts the amount of oversampling to local difficulty; safe-level and density-based SMOTE variants attempt to avoid generating samples in regions dominated by majority class points (22, 23).

Empirical analyses emphasise careful experimental practice: oversampling must be applied within training folds during cross-validation to avoid information leakage, and hybrid strategies (SMOTE combined with controlled under-sampling or with noise filters) often improve robustness (24). In healthcare prediction tasks, improving recall (sensitivity) for the positive class is usually a priority; SMOTE and its variants have been instrumental in achieving balanced performance measures (precision/recall trade-offs) rather than merely improving accuracy (25). Recent reviews also caution that synthetic oversampling can sometimes produce over fitting if the synthetic samples are not representative of real minority distributions; hence, pairing SMOTE with robust model selection and validation procedures is advisable (26).

Machine Learning for Lung Cancer Prediction: Demographic and Imaging Studies

Machine learning in lung cancer spans imaging-heavy radiomics/deep learning work and demographic/clinical tabular studies. On the imaging side, radiomics approaches extract quantitative features from CT images and have been successful in predicting nodule malignancy and outcomes; radiomic features contain prognostic information across cancer types, including lung cancer, and can be used to predict survival and phenotypes (27). Similarly, deep learning models trained on LDCT volumes deliver state-of-the-art performance in screening tasks, as shown in large-scale studies where convolutional networks matched or outperformed radiologists on certain tasks (28). While these studies set high performance benchmarks, their dependence on large annotated imaging datasets and high computational costs limits applicability in low-resource clinical environments. Conversely, models trained on demographic, lifestyle and symptom variables are attractive for primary screening. Several recent studies demonstrate that ensemble methods — Random Forests and gradient boosted trees perform well on such tabular datasets, giving stable performance and interpretable importance rankings (29, 30). For instance, multi-model ensemble studies on elderly cohorts found that combining demographic, environmental and clinical features yield ROC-AUC values in the 0.9 range for cohort-level incidence prediction when external validation is available (31). Other works that explicitly compare multiple classifiers report that RF and XGBoost frequently outperform linear models and single decision trees on mixed-type demographic data, largely due to their ability to capture nonlinear interactions and to tolerate correlated predictors (32).

Finally, there is a smaller body of work in which metaheuristics and oversampling is combined with ensembles for lung cancer and related respiratory disease prediction. Studies applying GA-based feature selection together with SMOTE and RF report improvements in recall and overall balanced metrics on non-imaging datasets, suggesting that the combined pipeline is promising for screening tasks where sensitivity is essential (33). Nevertheless, many published studies do not perform joint GA optimisation of both features and hyper parameters; instead, they tune hyper parameters separately or use simpler feature selection techniques. This gap motivates a unified GA dual-optimisation approach in the present work. Table 1 provides a comparative summary of key works referenced in this article.

The literature indicates strong evidence for three facts: (a) SMOTE and its variants are useful for improving minority-class detection in medical datasets; (b) ensemble methods like Random Forest and XGBoost are reliable choices for tabular clinical data; and (c) Genetic Algorithms and other metaheuristics are well suited for feature selection and hyperparameter tuning, but their joint application in a single pipeline for demographic-based lung cancer prediction remains under-represented. Thus, we proceed to implement and evaluate a GA-driven dual optimisation pipeline (feature selection + RF hyperparameter tuning), combined with SMOTE balancing, and benchmark against standard baselines.

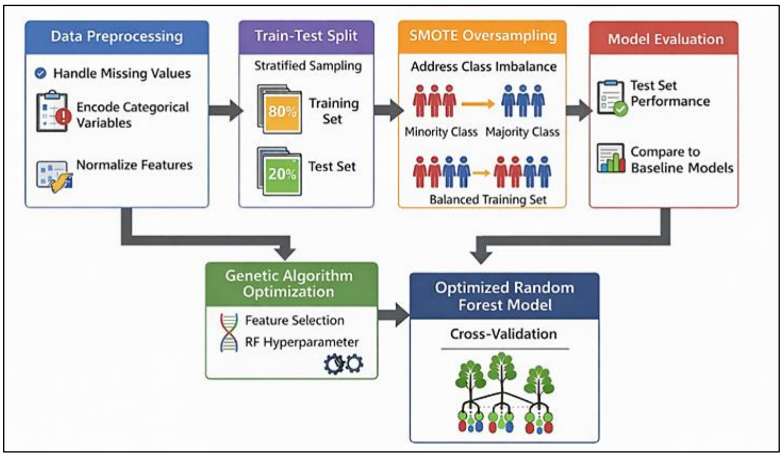

Figure 1: Overall Workflow of the Proposed Lung Cancer Prediction Framework

Methodology

Research Framework

The present study was designed as a well-organized pipeline that links various stages of machine learning aimed at lung cancer prediction. Initially, the demographic and symptom-related dataset is preprocessed to handle missing values and prepare the features appropriately. Following this, the issue of class imbalance is addressed using the Synthetic Minority Oversampling Technique (SMOTE). To enhance model accuracy and efficiency, a Genetic Algorithm is incorporated for both feature selection and hyperparameter tuning of the Random Forest classifier. Finally, the performance of the optimized model is compared against multiple baseline classifiers to demonstrate its robustness and effectiveness. The entire experiment was carried out on a system equipped with an Intel i7 processor and 16 GB of RAM. A graphical illustration of this comprehensive machine learning workflow is presented in Figure 1.

Dataset Description

The dataset used in this study contained demographic details such as age, gender, and lifestyle attributes, along with symptom variables like persistent cough, chest pain, fatigue, throat discomfort, and breathing difficulties. Each patient record was marked either as positive for lung cancer risk or negative. The class distribution was skewed with more negative than positive cases, which reflects the real-world scenario where the prevalence of lung cancer is lower compared to the healthy population. Such imbalance needed to be handled carefully before model training. This dataset was downloaded from Kaggle and consists of a total of 5,000 patient records. The class distribution is imbalanced, with 2,963 instances labeled as negative and 2,037 instances labeled as positive for lung cancer risk. This imbalance posed a challenge for model training and motivated the use of balancing techniques such as SMOTE to generate a fairer training dataset. A descriptive overview of dataset statistics is presented in Table 1.

Table 1: Dataset Characteristics

Attribute	Description	Type
AGE	Age of patient	Continuous
GENDER	Male (1) / Female (0)	Categorical
SMOKING	Smoking habit	Binary
FINGER_DISCOLORATION	Discoloration due to smoking	Binary
MENTAL_STRESS	Presence of mental stress	Binary
EXPOSURE_TO_POLLUTION	Long-term exposure to pollutants	Binary
LONG_TERM_ILLNESS	Existing chronic illness	Binary
ENERGY_LEVEL	Self-reported energy level	Continuous
IMMUNE_WEAKNESS	Weakness in immune system	Binary
BREATHING_ISSUE	Difficulty in breathing	Binary
ALCOHOL_CONSUMPTION	Alcohol usage	Binary
THROAT_DISCOMFORT	Presence of throat discomfort	Binary
OXYGEN_SATURATION	Blood oxygen saturation (%)	Continuous
CHEST_TIGHTNESS	Feeling of chest tightness	Binary
FAMILY_HISTORY	Family history of lung disease	Binary
SMOKING_FAMILY_HISTORY	Family history of smoking habit	Binary
STRESS_IMMUNE	Stress–immune interaction indicator	Binary
PULMONARY_DISEASE (Target)	Presence of lung cancer risk (Yes/No)	Binary

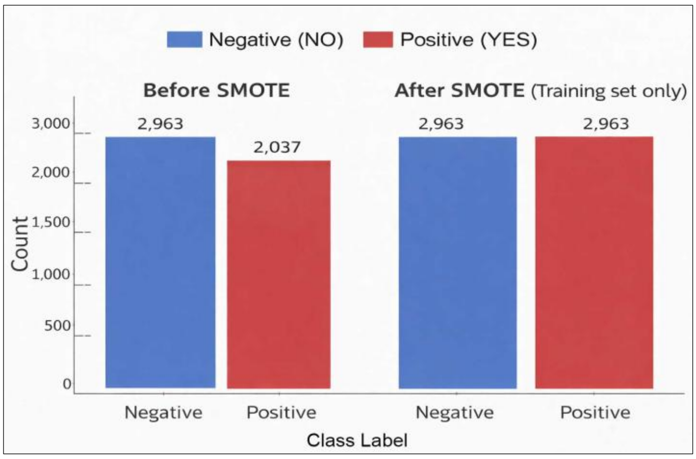

Figure 2: Distribution of Lung Cancer Risk Classes Before and After Applying the Synthetic Minority Oversampling Technique (SMOTE) to the Training Data

Data Pre-processing

Data preprocessing was carried out in a stepwise manner to make the dataset suitable for machine learning tasks. The missing values present in the dataset were replaced through imputation. For continuous variables, mean imputation was used, while for categorical variables the most frequent value was substituted. Categorical variables such as gender were encoded using one-hot encoding so that they could be used in the models. Certain numerical features were normalized to ensure that they remained within similar ranges. After preprocessing, the dataset was divided into training and testing sets in an 80:20 ratio. Stratified splitting was performed so that both sets preserved the same proportion of positive and negative classes. To address the prominent class imbalance, Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data. Figure 2 illustrates the effect of SMOTE on class distribution: prior to balancing, the dataset contained 2,963 negative (majority) and 2,037 positive (minority) instances. After applying SMOTE, both classes had 2,963 samples in the training set, resulting in a balanced distribution and supporting fairer model training.

Class Imbalance Handling with SMOTE

One of the critical challenges of the dataset was the imbalance between positive and negative cases. This was overcome using Synthetic Minority Oversampling Technique (SMOTE). SMOTE generates synthetic records of the minority class by interpolating between nearby positive instances. In the present study, SMOTE was applied only within training folds during cross-validation to prevent any information leakage into the test set. By creating synthetic but realistic minority samples, SMOTE allowed the classifiers to learn fairer decision boundaries and improved their ability to detect true positive cases. Figure 3 provides a schematic illustration of the balancing process.

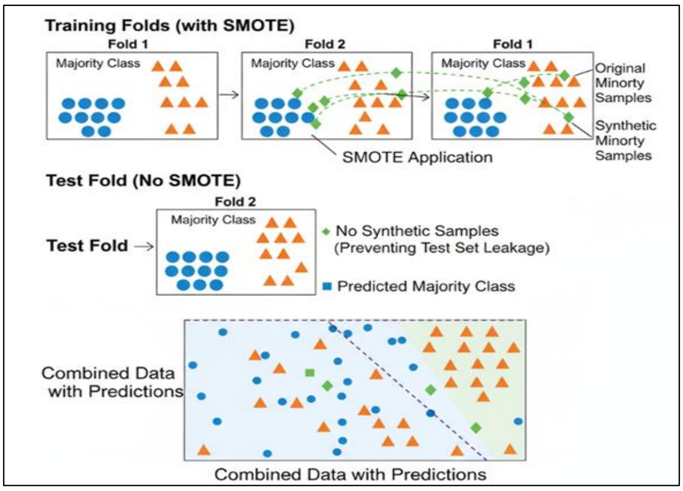

Figure 3: Schematic Illustration of the Synthetic Minority Oversampling Technique

Although Synthetic Minority Oversampling Technique (SMOTE) may generate artificial samples that do not perfectly correspond to real clinical profiles, its controlled use remains effective for mitigating class imbalance in medical screening tasks. In this study, SMOTE was applied exclusively to the training data within each cross-validation fold, ensuring that no synthetic information leaked into the test set. The primary objective was to improve sensitivity toward high-risk lung cancer cases, as missing true positives is more detrimental than producing false alarms in screening contexts. Furthermore, the Random Forest classifier's ensemble structure and robustness to noise help reduce sensitivity to potentially imperfect synthetic samples. This careful application allows SMOTE to enhance minority class learning while preserving clinical plausibility and model generalizability.

Genetic Algorithm for Optimization

A Genetic Algorithm was employed to serve two purposes: feature selection and hyperparameter tuning of the Random Forest classifier. GA is an evolutionary metaheuristic which simulates natural selection through operations like selection, crossover and mutation. In this study, the chromosomes were designed to contain both binary bits representing inclusion or exclusion of features and numerical values representing hyperparameters of the Random Forest model, such as number of trees, maximum depth and minimum samples required for split. The fitness of a chromosome was calculated based on cross-validated accuracy and F1 score. Tournament selection was used to identify strong parent chromosomes, and one-point crossover exchanged information between parents. Mutation was applied to randomly alter feature bits or change parameter values. The GA was allowed to evolve for a fixed number of generations or until convergence was reached. This design ensured that the algorithm simultaneously discovered the best subset of features and the most appropriate hyperparameters. The overall flow of this process is depicted in Algorithm 1 and Algorithm 2.

Algorithm 1: Genetic Algorithm for Feature Selection

Input and Output:

The inputs to the proposed algorithm consist of the datasetX, y, where X represents the feature matrix and y denotes the corresponding class labels. In addition, the Genetic Algorithm parameters include the population size N, which determines the number of candidate solutions evaluated in each generation, the maximum number of generations Gcontrolling the evolutionary search process, and the early stopping patience parameter P, which terminates the optimization when no improvement in fitness is observed over a predefined number of consecutive generations. The output of the algorithm is the optimal selected feature subset S^*, representing the most informative set of features identified through the evolutionary optimization process.

Procedure:

a) Initialize a population of N chromosomes: each chromosome is a binary vector indicating selected features.

b) For each generation g=1 to G:

   a. Evaluate fitness of each chromosome (model performance on selected features via cross-validation).

   b. Keep track of the best chromosome and its fitness score.

   c. If no improvement in best fitness over last P generations, break early.

   d. Select parents using tournament selection.

   e. Generate offspring: Offspring are generated through crossover between selected parent chromosomes, followed by mutation at a predefined rate to preserve population diversity and avoid premature convergence.

   f. Form a new population from offspring and a portion of elite chromosomes.

c) Return the best feature subset chromosome S^*.

Algorithm 2: Genetic Algorithm for Hyper parameter Optimization

Input and Output:

The inputs to the hyper parameter optimization process include the predefined hyper parameter boundsΘ, which specify the allowable ranges for the Random Forest parameters to be optimized. The Genetic Algorithm is further configured with a population size N, determining the number of candidates hyperparameter configurations evaluated in each generation, a maximum number of generations Gthat governs the extent of the evolutionary search, and an early stopping patience parameter P, which halts the optimization process if no improvement in fitness is observed over a specified number of consecutive generations. The output of this procedure is the optimal hyperparameter set θ^*, representing the most effective configuration identified through the evolutionary optimization.

Procedure:

a) Initialize a population of N chromosomes: each chromosome encodes hyperparameter values within bounds Θ.
b) For each generation g=1to G:
   a. Evaluate fitness of each chromosome (model performance with hyperparameters via cross-validation).
   b. Keep track of the best chromosome and its fitness score.
   c. If no improvement in best fitness over last P generations, break early.
   d. Select parents using tournament selection.
   e. Generate offspring: Offspring are generated through crossover between selected parent chromosomes, followed by mutation at a predefined rate to preserve population diversity and avoid premature convergence.
   f. Form a new population from offspring and a portion of elite chromosomes.
c) Return the best hyper parameter chromosome θ^*.

Random Forest Classifier

Random Forest was selected as the main predictive classifier because of its ability to handle mixed-type data and its resistance to overfitting. It is an ensemble approach where multiple decision trees are constructed on bootstrapped subsets of the training data. Predictions are then obtained through majority voting or averaging across trees. Another strength of Random Forest is its interpretability, as it provides feature importance rankings, which are valuable for clinical understanding. In this research, Random Forest was used in its GA-optimized form as well as in its standard form to compare the advantage of optimization.

Baseline Models for Comparison

To demonstrate the strength of the proposed GA-RF framework, several baseline models were implemented. These included Logistic Regression, Support Vector Machine, k-Nearest Neighbors, Decision Tree, and standard Random Forest without optimization and XGBoost. The baseline models were trained and tuned through conventional procedures and evaluated under the same protocol to ensure fairness in comparison. Their results provide a benchmark against which the performance of the GA-optimized RF can be measured.

Evaluation Metrics and Validation Strategy

The performance of all models was assessed using accuracy, precision, recall, F1 score and ROC-AUC. Accuracy reflects the proportion of correctly classified cases, precision measures the correctness of predicted positives, recall denotes the proportion of actual positives that were correctly detected, while F1 score provides a balance between precision and recall. ROC-AUC was also considered as it indicates the overall discriminative power of the model across different thresholds. The models were evaluated using 5-fold cross-validation during training to avoid overfitting, and the final metrics were reported on the independent test set.

Results

Comparative Model Performance

The comparison of different models is shown in Table 2. Among all classifiers, the GA-optimized Random Forest achieved the highest accuracy of 90.6 percent, an F1 score of 0.885, and an ROC-AUC of 0.917. Logistic Regression and SVM showed moderate performance, while k-Nearest Neighbors and Decision Tree performed relatively weaker. Standard Random Forest without optimization gave satisfactory results, but once combined with GA optimization it achieved a clear performance gain. XGBoost also performed well, yet it was slightly inferior to GA-RF. These results indicate that the combined effect of GA-based feature selection and parameter tuning significantly enhanced the predictive ability of Random Forest.

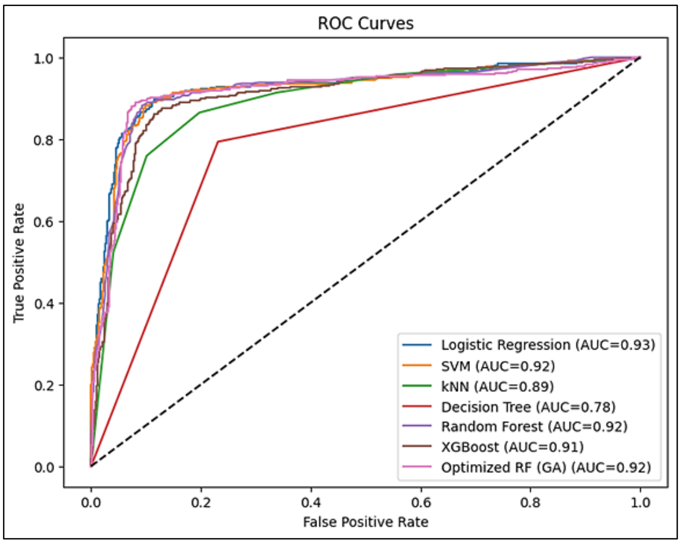

Figure 4: ROC Curves of Various Classifiers for Lung Cancer Prediction

The ROC curves of the evaluated classifiers are presented in Figure 4. It is observed that the standard Random Forest model achieves a slightly higher ROC-AUC value (0.921) compared to the GA-optimized Random Forest (0.917). This suggests that while the Genetic Algorithm optimization improved other metrics such as accuracy and F1 score, the discriminative ability of the optimized model as measured by ROC-AUC is marginally reduced. Such a trade-off is common in model optimization, where gains in prediction accuracy might come with subtle changes in classification thresholds and sensitivity-specificity balance. Overall, both Random Forest models demonstrate strong discriminatory power, with the GA-optimized version offering enhanced overall predictive performance.

Table 2: Performance Comparison of Models

Model	Accuracy	F1 Score	ROC-AUC
Logistic Regression	0.872	0.852	0.925
Support Vector Machine	0.885	0.863	0.923
k-Nearest Neighbors	0.828	0.804	0.892
Decision Tree	0.779	0.745	0.781
Random Forest	0.894	0.872	0.921
XGBoost	0.873	0.847	0.908
Optimized RF (GA)	0.906	0.885	0.917

Classification Report

The classification report of GA-RF given in Table 3 highlights that the model was able to maintain high precision and recall simultaneously. This shows that it is not only predicting correctly but also reducing the chances of missing true positive cases. The confusion matrix in Figure 5 further supports this by showing that a majority of cancer positive cases were correctly classified. Even though a few false positives were present, such an outcome is acceptable in a medical screening scenario where the consequences of false negatives are much more severe.

Table 3: Classification Report of Optimized RF (GA)

Class	Precision	Recall	F1-Score	Support
Negative (NO)	0.91	0.93	0.92	2963
Positive (YES)	0.86	0.84	0.85	2963
Accuracy	–	–	**0.906**	5926
Macro Avg	0.883	0.887	0.885	–
Weighted Avg	0.904	0.906	0.905	–

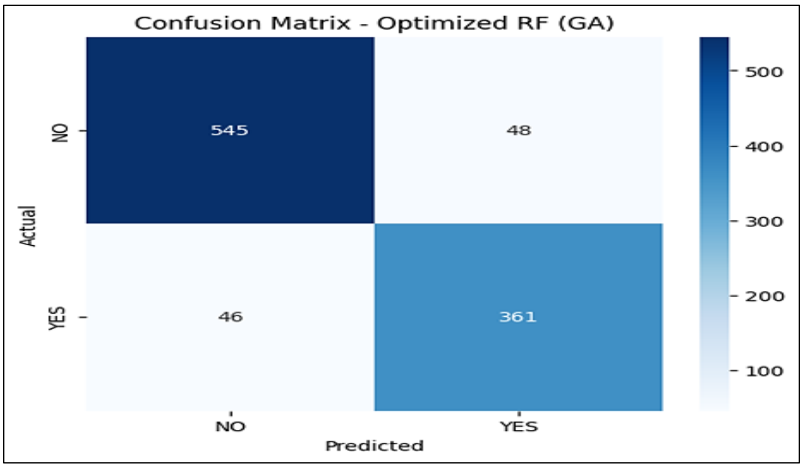

Figure 5: Confusion Matrix for Optimized RF (GA)

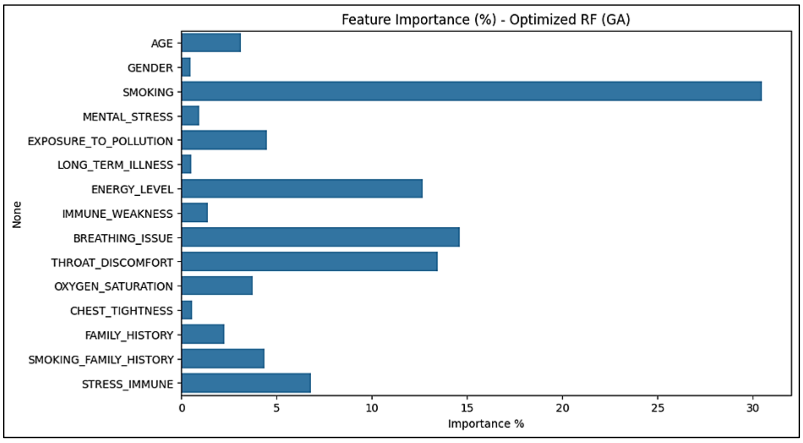

Figure 6: Feature Importance of Optimized RF (GA)

Feature Importance Analysis

The analysis of feature importance revealed smoking habit as the top predictor of lung cancer risk. This was followed by breathing difficulty and throat discomfort, which also carried high importance. Other features such as chest pain, fatigue and gender played supportive roles in prediction. This order of importance is in line with existing medical evidence, where smoking and respiratory symptoms are identified as leading indicators of lung cancer. The ranking of features is shown in Figure 6. Such interpretability adds practical value, as clinicians can relate the predictions to patient history and symptoms.

Discussion

The results of this study demonstrate that the proposed Genetic Algorithm–optimized Random Forest (GA-RF) model achieves robust performance for lung cancer risk prediction using demographic and symptom-based data. The obtained accuracy, F1-score, and ROC-AUC are consistent with recent studies reporting that ensemble learning models outperform linear and single-tree classifiers when applied to heterogeneous clinical datasets. In particular, ensemble-based approaches have been shown to effectively capture nonlinear relationships and feature interactions in demographic and lifestyle data, leading to improved predictive reliability in lung cancer risk assessment tasks (34, 35).

The performance gains observed through Genetic Algorithm-based optimization align with earlier research emphasizing the effectiveness of evolutionary metaheuristics for feature selection and hyperparameter tuning in medical prediction problems (36, 37). Prior studies indicate that GA-based optimization is particularly advantageous in complex search spaces involving mixed-type variables, where conventional grid or random search strategies may fail to identify optimal configurations. The results of this study further support these findings, showing that evolutionary optimization can enhance model generalization and balanced performance metrics in disease screening scenarios.

Class imbalance handling using SMOTE played a critical role in improving the detection of high-risk lung cancer cases. The improvement in recall and F1-score observed in this work is consistent with comparative studies demonstrating that SMOTE-based techniques significantly enhance minority-class learning in imbalanced clinical datasets when applied carefully within the training process (38). By restricting oversampling to training folds only, the present study mitigates the risk of information leakage and overfitting, thereby maintaining the validity of the evaluation.

Although genetic algorithm–based feature selection and Random Forest classifiers have been explored independently in prior studies, most existing works apply feature selection and hyperparameter tuning as separate or sequential processes. In contrast, the present study formulates lung cancer risk prediction as a joint optimization problem, in which a single Genetic Algorithm simultaneously evolves both the feature subset and the Random Forest hyperparameters under class-imbalanced training conditions. This integrated design facilitates the identification of synergistic features–parameter combinations while preserving model interpretability and ensuring suitability for screening-level applications based on demographic and symptom data.

Feature importance analysis revealed smoking habit, breathing difficulties, and throat discomfort as dominant predictors, which is consistent with findings reported in earlier demographic-based lung cancer prediction studies (34, 35). Minor differences in feature ranking compared to previous work may be attributed to variations in dataset composition, population characteristics, and feature encoding strategies. While advanced interpretability methods such as SHAP and partial dependence plots have been shown to provide instance-level explanations in recent ensemble-based healthcare studies (39), the present study deliberately employs Random Forest feature importance to maintain simplicity and transparency. This global interpretability approach aligns well with the screening-oriented objective of the proposed framework. Nevertheless, incorporating SHAP or partial dependence analysis to provide individualized explanations is identified as a promising direction for future research.

Conclusion

The present study attempted to design an efficient and interpretable framework for predicting lung cancer risk using simple demographic and symptom-based information. By employing Genetic Algorithm for simultaneous feature selection and hyperparameter optimization of Random Forest, and combining it with SMOTE to overcome class imbalance, the proposed model achieved superior performance compared to conventional classifiers. The results clearly indicated that smoking habit, breathing issues and throat discomfort are dominant predictors, which is also consistent with medical knowledge. The GA-optimized Random Forest not only improved predictive accuracy but also provided interpretability, thereby making it a practical tool for early detection in low-resource settings where costly imaging-based approaches may not be feasible. From a broader perspective, such models can support community-level screening programmes and contribute to better allocation of healthcare resources.

Abbreviations

ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced, CT: Computed tomography, ECG: Electrocardiogram, EEG: Electroencephalogram, GA-RF: Random Forests (RF) and Genetic Algorithms (GA), KNN: K Nearest Neighbour, PSO: Particle Swarm Optimisation, ROC-AUC: Area under the Receiver Operating Characteristic Curve, SVM: Support Vector Machine.

Acknowledgement

None.

Author Contributions

All authors equally contributed to the conception, design of the study, data collection, pre-processing, model development, experimentation, analysis, interpretation of results, manuscript preparation, revision.

Conflict of Interest

The authors declare no conflict of interest.

Declaration of Artificial Intelligence (AI) Assistance

AI tools were used for language editing only, and all content was verified by the authors.

Ethics Approval

Not applicable.

Funding

This research received no external funding.

References

1. Ardila D, Kiraly AP, Bharadwaj S, *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019;25(6):954–61.
2. Aerts HJ, Velazquez ER, Leijenaar RT, *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
3. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016;785–94. https://medial-earlysign.github.io/MR_Wiki/attachments/5537821/5537823.pdf
4. Chawla NV, Bowyer KW, Hall LO, *et al.* SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
5. Breiman L. Random forests. Mach Learn. 2001;45(1): 5–32.
6. Suryadi MK, Herteno R, Saputro SW, *et al.* Comparative study of various hyperparameter tuning on random forest classification with SMOTE and feature selection using genetic algorithms in software defect prediction. J Electron Electromed Eng Med Inform. 2024;6(2):137–47.
7. Sampson JR. Adaptation in natural and artificial systems (John H. Holland). SIAM Review. 1976;18(3): 529–30. https://doi.org/10.1137/1018105
8. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput Biol Med. 2019;112:103375.
9. Taha ZY, Abdullah AA, Rashid TA. Optimizing feature selection with genetic algorithms: A review of methods and applications. Knowl Inf Syst. 2025;67(11):1–40.
10. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.
11. Wutzl B, Leibnitz K, Rattay F, *et al.* Genetic algorithms for feature selection when classifying severe chronic disorders of consciousness. PLoS One. 2019;14(7):e0219683.
12. Mao Y, Yang Y. A wrapper feature subset selection method based on randomized search and multilayer structure. Biomed Res Int. 2019;2019:9864213.
13. Singh N, Singh P. A hybrid ensemble-filter wrapper feature selection approach for medical data classification. Chemom Intell Lab Syst. 2021;217:104396.
14. Mandal M, Singh PK, Ijaz MF, *et al.* A tri-stage wrapper–filter feature selection framework for disease classification. Sensors. 2021;21(16):5571.
15. Xue B, Zhang M, Browne WN, *et al.* A survey on evolutionary computation approaches to feature selection. IEEE Trans Evol Comput. 2015;20(4):606–26.
16. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13: 281–305.
17. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. Adv Neural Inf Process Syst (NIPS). 2012;2951–9. https://proceedings.neurips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf
18. Cagnini HE, Dôres SC, Freitas AA, *et al.* A survey of evolutionary algorithms for supervised ensemble learning. Knowl Eng Rev. 2023;38: e1.
19. Arabasadi Z, Alizadehsani R, Roshanzamir M, *et al.* Computer-aided decision making for heart disease detection using hybrid neural network–genetic algorithm. Comput Methods Programs Biomed. 2017;141:19–26.
20. Wang Z, Zhou Y, Takagi T, *et al.* Genetic algorithm-based feature selection with manifold learning for cancer classification using microarray data. BMC Bioinformatics. 2023; 24:139.
21. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013; 14:106.
22. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–82.
23. Han H, Wang WY, Mao BH. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. Int Conf Intell Comput. 2005;878–87. https://sci2s.ugr.es/keel/keel-dataset/pdfs/2005-Han-LNCS.pdf
24. Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl. 2004;6(1):20–9.
25. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18(17):1–5.
26. Torgo L, Ribeiro RP, Pfahringer B, *et al.* SMOTE for regression. In: Proc Portuguese Conf Artif Intell. Berlin, Heidelberg: Springer. 2013;378–389. https://link.springer.com/chapter/10.1007/978-3-642-40669-0_33
27. Aerts HJ, Velazquez ER, Leijenaar RT, *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014; 5:4006.
28. Huang P, Lin CT, Li Y, *et al.* Prediction of lung cancer risk at follow-up screening with low-dose CT: A deep learning method. Lancet Digit Health. 2019;1(7): e353–62.
29. Chen S, Wu S. Ensemble machine learning models for lung cancer incidence risk prediction in the elderly: A retrospective longitudinal study. BMC Cancer. 2025;25(1):126.
30. Mohan K, Thayyil B. Machine learning techniques for lung cancer risk prediction using text dataset. Int J Data Informatics Intell Comput. 2023;2(3):47–56.
31. Quasar SR, Sharma R, Mittal A, *et al.* Ensemble methods for computed tomography scan images to improve lung cancer detection and classification. Multimed Tools Appl. 2024;83(17):52867–97.
32. Mamun M, Farjana A, Al Mamun M, *et al.* Lung cancer prediction model using ensemble learning techniques and a systematic review analysis. IEEE World AI IoT Congr (AIIoT). 2022;187–93.

https://ieeexplore.ieee.org/abstract/document/9817326/

33. Hashmi A, Ali W, Abulfaraj A, Binzagr F, Alkayal E. Enhancing cancerous gene selection and classification for high-dimensional microarray data using a novel hybrid filter and differential evolutionary feature selection. Cancers. 2024; 16(23):3913.
34. Chandran U, Reps J, Yang R, Vachani A, Maldonado F, Kalsekar I. Machine learning and real-world data to predict lung cancer risk in routine care. Cancer Epidemiol Biomarkers Prev. 2023;32(3):337–343.
35. Dubey A, Yadav P, Patel SC, Bhargava CP, Tomar A. Identifying lung cancer: A review on classification and detection. Traitement du Signal. 2024;41(4):2023-2034. https://doi.org/10.18280/ts.410431
36. Taleb AW. Investigating early lung cancer detection through feature selection and ensemble machine learning. Int J Appl Math. 2025;38(7s):505–518.
37. Murad SH, Tayfor NB, Mahmood NH, Arman L. Hybrid genetic algorithms-driven optimization of machine learning models for heart disease prediction. MethodsX. 2025;15:103510.
38. Khushi M, Shaukat K, Alam TM, *et al.* A comparative performance analysis of data resampling methods on imbalanced medical data. IEEE Access. 2021; 9:109960–109975.
39. Ganie SM, Pramanik PKD, Zhao Z. Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets. Sci Rep. 2025;15(1):13912.

How to Cite: Balaji T, Babu P, Lokeshwaran K. Metaheuristic Optimization of Random Forest for Lung Cancer Prediction. Int Res J Multidiscip Scope. 2026; 7(1): 1666-1678.
DOI: 10.47857/irjms.2026.v07i01.08680