AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: Please note that we cannot accept new source files as corrections for your article. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

Carefully check the page proofs (and coordinate with all authors); additional changes or updates WILL NOT be accepted after the article is published online/print in its final form. Please check author names and affiliations, funding, as well as the overall article for any errors prior to sending in your author proof corrections. Your article has been peer reviewed, accepted as final, and sent in to IEEE. No text changes have been made to the main part of the article as dictated by the editorial level of service for your publication.

AQ:1 = Please confirm whether the edits made in the current affiliation of all the authors are correct.

AQ:2 = Please confirm or add details for any funding or financial support for the research of this article.

AQ:3 = If you haven't done so already, please make sure you have submitted a graphical abstract for your paper. The GA should be a current figure or image from your accepted article. The GA will be displayed on your articles abstract page on IEEE Xplore. Please choose a current figure from the paper and supply a caption at your earliest convenience for the graphical abstract. Note that captions cannot exceed 1800 characters (including spaces). If you submitted a video as your graphical abstract, please make sure there is an overlay image and caption. Overlay images are usually a screenshot of your video that best represents the video. This is for readers who may not have access to video-viewing software. Please see an example in the link below: http://ieeeaccess.ieee.org/submitting-an-article/

AQ:4 = Please provide the publisher name and publisher location for Ref. [2].

AQ:5 = Please provide the publisher location for Ref. [4].

AQ:6 = Please note that Refs. [19] and [52] were identical in your originally submitted manuscript. Hence, we have deleted Ref. [52] and renumbered the subsequent references. This will also be reflected in the citations present in the body text.

AQ:7 = Please provide the page range for Refs. [30] and [43].

IEEE Proof

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 January 2025, accepted 24 January 2025. Date of publication 00 xxxx 0000, date of current version 00 xxxx 0000.

Digital Object Identifier 10.1109/ACCESS.2025.3536034

Integrating Advanced Techniques: RFE-SVM Feature Engineering and Nelder-Mead Optimized XGBoost for Accurate Lung Cancer Prediction

HAMDI A. AL-JAMIMI[1,2], SARAH AYAD[3], AND AMMAR EL KHEIR[4]

AQ:1 [1]Information and Computer Science, King Fahd University of Petroleum and Minerals, Dhahran 31216, Saudi Arabia
[2]Research Excellence, King Fahd University of Petroleum and Minerals, Dhahran 31216, Saudi Arabia
[3]Faculty of Computer Studies, Arab Open University, Riyadh 11681, Saudi Arabia
[4]Department of Pediatrics, Liege University Hospital Center, 4000 Liège, Belgium

Corresponding author: Sarah Ayad (s.ayad@arabou.edu.sa)

AQ:2 This work was supported by Arab Open University under Grant AOUKSA-524008. The work of Hamdi A. Al-Jamimi was supported by the King Fahd University of Petroleum and Minerals (KFUPM).

AQ:3 **ABSTRACT** Early detection of lung cancer is crucial for improving patient survival and reducing mortality. However, medical datasets often face challenges like irrelevant features and class imbalance, complicating accurate predictions. This study presents a comprehensive AI-powered lung cancer classification approach that enhances predictive accuracy and treatment planning. Our methodology combines Recursive Feature Elimination with Support Vector Machines (RFE-SVM) for effective feature selection and employs the XGBoost ensemble learning algorithm for classification, optimized using the Nelder-Mead algorithm. Evaluating the model's generalizability on two distinct lung cancer datasets, results show that our approach outperforms traditional machine learning models, achieving 100% accuracy. This research highlights the importance of advanced computational techniques in healthcare, paving the way for more personalized and effective patient care.

INDEX TERMS Early prediction, feature engineering, lung cancer, XGBoost.

I. INTRODUCTION

Lung cancer is among the most common cancers worldwide and one of the major causes of cancer mortality. It affects individuals of both genders, with approximately 2.1 million diagnoses made in 2018, resulting in around 1.8 million deaths [1]. Its prevalence surpasses that of colon, breast, and prostate cancer combined, with over 40% of cases diagnosed at advanced stages, resulting in dismal long-term survival rates [2]. Typically diagnosed in its later stages, lung cancer poses significant challenges to effective treatment and management, characterized by abnormal and uncontrollable cell growth within the lungs [3]. The diagnostic journey involves steps, from symptom onset to medical consultation, imaging techniques, histopathology, and molecular diagnosis [4]. Lung cancer risk may stem from a range of factors, including lifestyle, environmental exposure, genetic predisposition, and existing medical conditions. Risk prediction assesses these factors to estimate an individual's likelihood of developing lung cancer within a specific timeframe [5].

In the quest to advance medical diagnostics and enhance patient outcomes, harnessing the power of artificial intelligence (AI) holds significant promise [6]. AI technologies, including machine learning (ML) algorithms and deep learning frameworks, offer capabilities in analyzing vast medical datasets, identifying complex patterns, and building predictive models [7]. AI-driven approaches enhance prediction accuracy for infectious and chronic diseases [8], [9], supporting early detection, personalized treatment, and better patient outcomes. The AI algorithms offer innovative solutions for feature selection, hyperparameter tuning, and model optimization [10]. Given the multifaceted nature of lung cancer prediction, involving the analysis of diverse datasets and the identification of intricate patterns, the integration of AI algorithms can play a pivotal role in enhancing the accuracy and robustness of predictive

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

IEEE Proof

models [11]. In recent years, significant strides have been made in lung cancer risk prediction, driven by advancements in computational methodologies and healthcare data analytics [12]. A multitude of studies have explored various approaches, ranging from traditional statistical models to cutting-edge ML algorithms [13], [14]. This paper proposes a novel approach to lung cancer prediction, integrating advanced techniques to achieve superior performance. The main contributions are:

- *Integrated Feature Engineering:* We combine Recursive Feature Elimination (RFE [15]) with Support Vector Machine (SVM [16], [17]) to identify the most relevant features by iteratively eliminating less impactful variables based on SVM weights [18].
- *Efficient Ensemble Learning with XGBoost:* We use the XGBoost classifier, known for its efficiency and robustness, to enhance predictive accuracy and mitigate overfitting.
- *Optimized Hyperparameter Tuning:* Our methodology optimizes the XGBoost hyperparameters using the Nelder-Mead algorithm [19] to maximize model effectiveness in predicting lung cancer risk.

Using two distinct lung cancer datasets, we aim to demonstrate the model's effectiveness across diverse scenarios. Despite minimal missing or erroneous data, the datasets are small, and one is notably imbalanced. The proposed approach addresses both feature selection and data imbalance, ensuring that the features used in the lung cancer prediction model are relevant and contribute significantly to accurate classification.

The subsequent sections of this paper are structured as follows: Section II summarizes recent related work. While Section III gives an overview of the investigated datasets, Section IV introduces the proposed methodology. Section V discusses the obtained results, and Section VI concludes the research paper.

II. RELATED WORK

This section presents a sample of recent studies in two different categories: imaging and clinical data, showcasing various approaches and methodologies employed in lung cancer prediction research.

For medical imaging analysis, approaches like computer vision, deep learning fusion, and blockchain-based federated learning show promising results in overcoming diagnostic challenges from CT scans. Researchers have employed deep learning for lung nodule detection [20], [21], [22], early-stage lung cancer identification from CT scans [23], [24], and categorization of malignant lung nodules [25], [26], [27], [28]. In addition, ANFIS-based has been used for lung and colon cancer using histopathological images, combining noise reduction, advanced feature extraction, and dimensionality reduction [29]. CNN, LSTM, and Bi-LSTM models have been integrated to detect lung cancer from CT scans [30]. A hybrid recurrent and feed-forward neural network (Hyb-RNN-FFBPNN) optimized with the glowworm swarm algorithm (GWSA) has been proposed to enhance early lung cancer detection in computer-aided diagnosis [31].These studies highlight significant advancements in using computational techniques for early detection and accurate diagnosis of lung cancer.

Recent advancements in clinical data mining have significantly improved the detection and diagnosis of lung cancer. For instance, "DeepXplainer," an interpretable hybrid deep learning model that combines convolutional neural networks with XGBoost, has been employed for lung cancer prediction while also providing explanations for its predictions using the SHapley Additive exPlanation (SHAP) method [32]. This approach enhances the effectiveness of lung cancer detection and treatment for patients. Additionally, the integration of SHAP within Explainable ML techniques has further improved the interpretability of lung cancer detection systems [33]. An automated model for early-stage lung cancer detection, employing nine ML algorithms, highlighting the potential for early and cost-effective detection [34]. Another approach leveraged deep learning and natural language processing to extract and predict lung cancer instances from electronic health records, demonstrating improved prediction accuracy [35].

An Artificial Neural Network (ANN)-based approach used symptoms and relevant information to detect lung cancer presence with 96.67% accuracy [36]. Furthermore, the Rotation Forest algorithm was employed to identify high-risk individuals, achieving an impressive AUC of 99.3% and high performance across various metrics [37]. Ensemble techniques, including XGBoost and LightGBM, applied to a survey dataset using the synthetic minority oversampling (SMOTE) method, showed that XGBoost outperformed other techniques, achieving 94.42% accuracy, demonstrating its effectiveness in predicting lung cancer [38]. These studies collectively highlight the significant potential of advanced computational techniques in improving lung cancer detection and diagnosis.

That combines the optimized NM-XGBoost classifier with SVM-based RFE for feature selection. This methodology distinguishes our work from previous studies by introducing a novel integration of these two techniques, significantly enhancing the accuracy of lung cancer prediction within the clinical data mining field.

III. MATERIALS

In our research, we utilized two distinct lung cancer datasets to train and validate our model, each presenting unique challenges and characteristics. The objective is to evaluate the model's performance across diverse scenarios and assess its capacity to generalize effectively across varying dataset conditions.

A. LUNG CANCER RISK DATASET

The dataset encompasses a wide range of features related to individuals, intended to explore potential factors associated

IEEE Proof

TABLE 1. Patient characteristics of lung cancer risk dataset.

Feature	Summary Statistics
Gender	Male: 47 (47%), Female: 53 (53%)
Age	Mean: 62.1, SD: 8.6, Range: 21-81
Smoking	Yes: 57 (57%), No: 43 (43%)
Yellow finger	Yes: 49 (49%), No: 51 (51%)
Anxiety	Yes: 48 (48%), No: 52 (52%)
Peer pressure	Yes: 36 (36%), No: 64 (64%)
Chronic diseases	Yes: 39 (39%), No: 61 (61%)
Fatigue status	Yes: 51 (51%), No: 49 (49%)
Allergy	Yes: 38 (38%), No: 62 (62%)
Wheezing	Yes: 46 (46%), No: 54 (54%)
Alcohol consuming	Yes: 43 (43%), No: 57 (57%)
Coughing	Yes: 65 (65%), No: 35 (35%)
Shortness of breath	Yes: 57 (57%), No: 43 (43%)
Swallowing difficulty	Yes: 34 (34%), No: 66 (66%)
Chest pain	Yes: 55 (55%), No: 45 (45%)
Lung cancer (Target)	Yes: 69 (69%), No: 31 (31%)

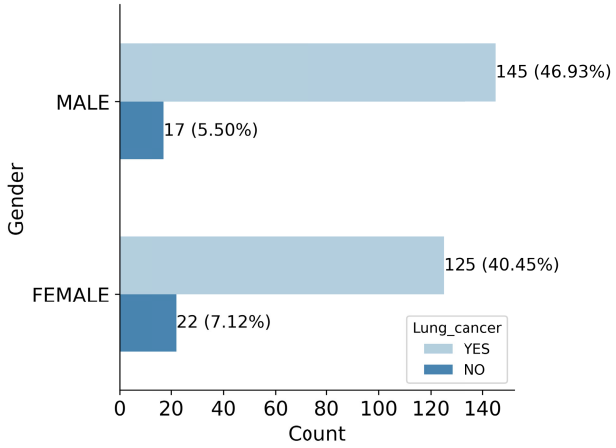


FIGURE 1. Distribution of gender with cancer status. The figure illustrates the proportion of male and female individuals categorized by their cancer status.

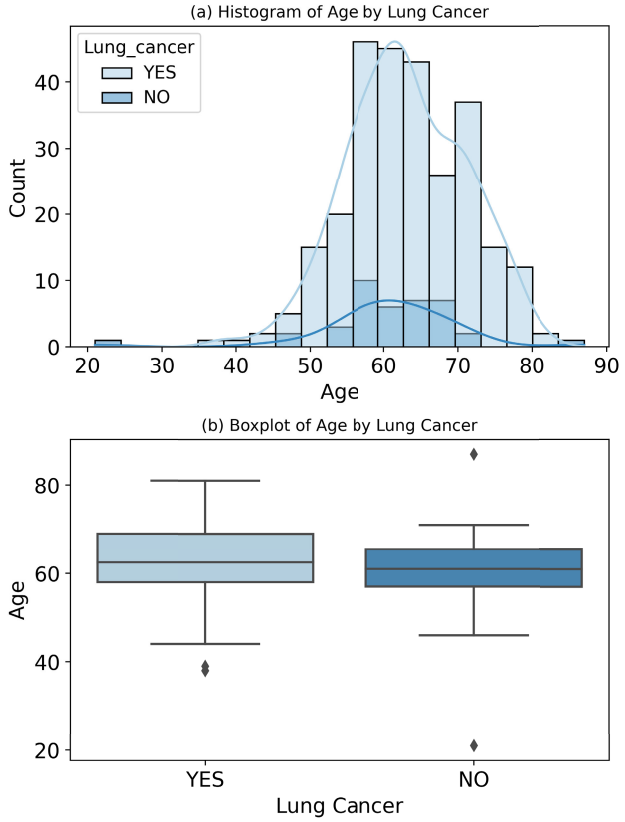


FIGURE 2. a) The distribution of lung cancer cases by age, and b) A boxplot depicting the distribution of age among cancer cases.

with the development of lung cancer. The dataset comprises 309 entries, representing individuals both affected by lung cancer and those unaffected by the disease. It includes 16 attributes, with 15 being predictive variables designed to identify potential risk factors, presented in Table 1.The dataset was collected from Kaggle repository [39].

In the dataset analysis, it was observed that among males, approximately 5.5% were identified as healthy individuals, while a significantly higher proportion, approximately 46.93%, were found to have been diagnosed with lung cancer. In contrast, among females, around 7.12% were categorized as healthy, while approximately 40.45% were reported to have lung cancer based on the dataset under scrutiny. These statistics underline a substantial disparity in lung cancer prevalence between genders within the dataset, highlighting a notably higher incidence among males compared to females, Figure 1. The dataset exhibited significant class imbalance. However, our proposed approach addresses this issue by employing the XGBoost classifier and ensuring effective handling of class imbalance.

In exploring the distribution of age among individuals with lung cancer and those classified as healthy within the dataset, a distinct pattern emerged. The peak age range for individuals diagnosed with lung cancer was notably between 55 and 75 years old, Figure 2. This range stood out as the focal point where a significant proportion of lung cancer cases were prevalent within the dataset see Figure 2a. The plots in Figure 2b indicate that lung cancer is more common in older individuals. Age distribution for lung cancer skews older, while those without are more varied in age.

B. LUNG CANCER PATIENT DATASET

The dataset provides detailed information on patients diagnosed with lung cancer, covering factors such as age, gender, environmental exposure, lifestyle choices, and health conditions. Analyzing this dataset can offer insights into the causes of lung cancer. Derived from a study published in Nature Medicine, the dataset underscores the significance of environmental factors, particularly air pollution, in lung cancer risk, highlighting its relevance in understanding the disease's multifaceted nature. The dataset consists of 1000 cases and 23 features, with one feature representing

the target variable. The target variable indicates the level of lung cancer risk, categorized as high, medium, or low. The dataset with more than two classes, beyond the binary classification of lung cancer positive and negative cases. This dataset included additional classes representing different subtypes or stages of lung cancer, adding complexity to the classification task. The dataset was collected from Kaggle repository [40]. Table 2 provides a brief overview of each attribute included in the dataset, offering insight into the factors potentially associated with lung cancer risk.

TABLE 2. Patient characteristics of lung cancer patient dataset.

Feature	Description
Age	Age of the individual
Gender	1 for Male, 2 for Female
Air Pollution	1 to 7, representing exposure level
Alcohol Use	1 to 8, representing frequency of alcohol use
Dust Allergy	1 to 8, representing severity
Occupational Hazards	1 to 8, representing exposure to occupational hazards
Genetic Risk	1 to 7, representing genetic predisposition
Chronic Lung Disease	1 to 7, representing the severity of lung disease
Balanced Diet	1 to 7, representing adherence to a balanced diet
Obesity	1 to 8, representing obesity levels
Smoking	1 to 8, representing frequency of smoking
Passive Smoker	1 to 8, representing exposure to passive smoking
Chest Pain	1 to 9, representing severity of chest pain
Coughing of Blood	1 to 8, representing frequency
Fatigue	1 to 8, representing severity of fatigue
Weight Loss	1 to 7, representing the extent of weight loss
Shortness of Breath	1 to 9, representing severity
Wheezing	1 to 8, representing severity
Swallowing Difficulty	1 to 9, representing severity
Clubbing of Finger Nails	1 to 8, representing severity
Frequent Cold	1 to 9, representing frequency of colds
Dry Cough	1 to 7, representing frequency of dry cough
Snoring	1 to 9, representing frequency of snoring
Level (Target)	Values: Low, Medium, High

Figure 3 illustrates the distribution of cases based on gender and the risk of developing lung cancer categorized as Low, Medium, and High. The figure reveals that 59.8% are male, whereas 40.2% are female. Furthermore, 30.3% of the cases are categorized as low risk, 33.2% as medium risk, and 36.5% as high risk. Specifically, out of 365 high-risk cases, 252 are male, while only 113 cases are female.

Figure 4 displays the distribution of samples by age, air pollution levels, and the risk of developing lung cancer. The figure illustrates that cases are distributed across various age groups corresponding to the three risk categories, with a higher potential for developing lung cancer observed in older individuals, particularly those aged 70 years and above. Additionally, there is a positive correlation between increased air pollution levels and a higher risk of developing lung cancer.

IV. METHODOLOGY

The proposed methodology offers an advanced predictive approach to lung cancer classification, integrating techniques

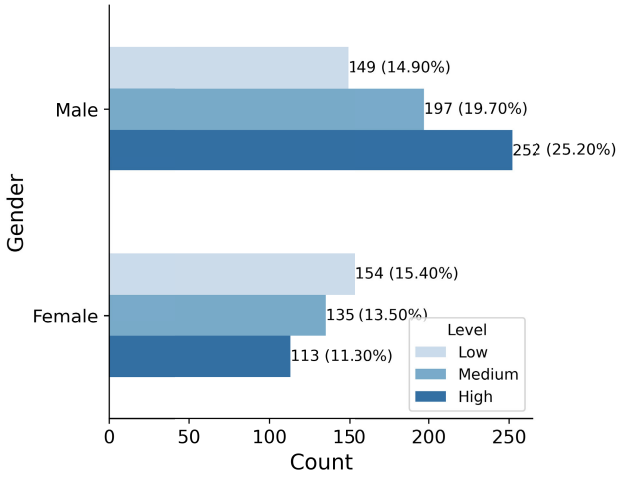


FIGURE 3. Samples distribution by gender and risk of developing lung cancer.

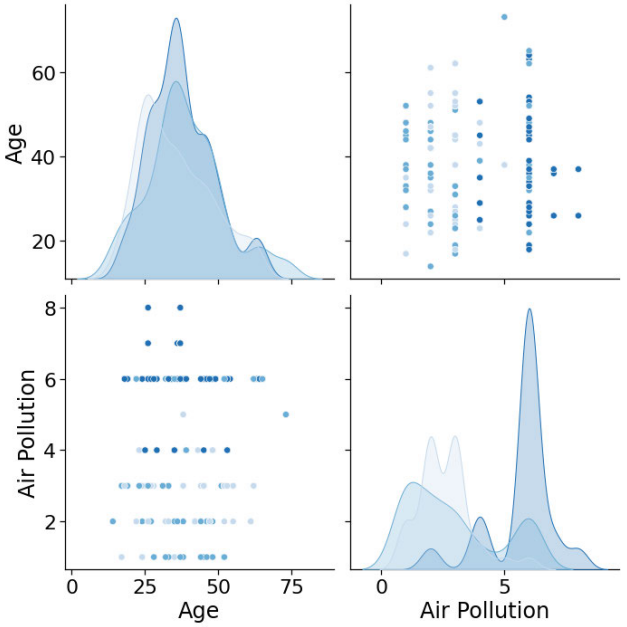


FIGURE 4. Samples distribution by age, air pollution, and risk of developing lung cancer.

for feature selection and model optimization. The combination of SVM-based RFE, XGBoost, Nand elder-Mead optimization strengthens the model's predictive accuracy, enhancing reliability in the early detection of lung cancer. The methodology is illustrated in Figure 5. Our methodology includes a data preprocessing phase, recognizing that missing and noisy data are common challenges in medical datasets. While the datasets utilized in this study are free of significant issues, we still implemented Encoding for categorical variables, such as gender and the target class, to ensure they are appropriately formatted for the modeling process. This paper considers data preprocessing as a key

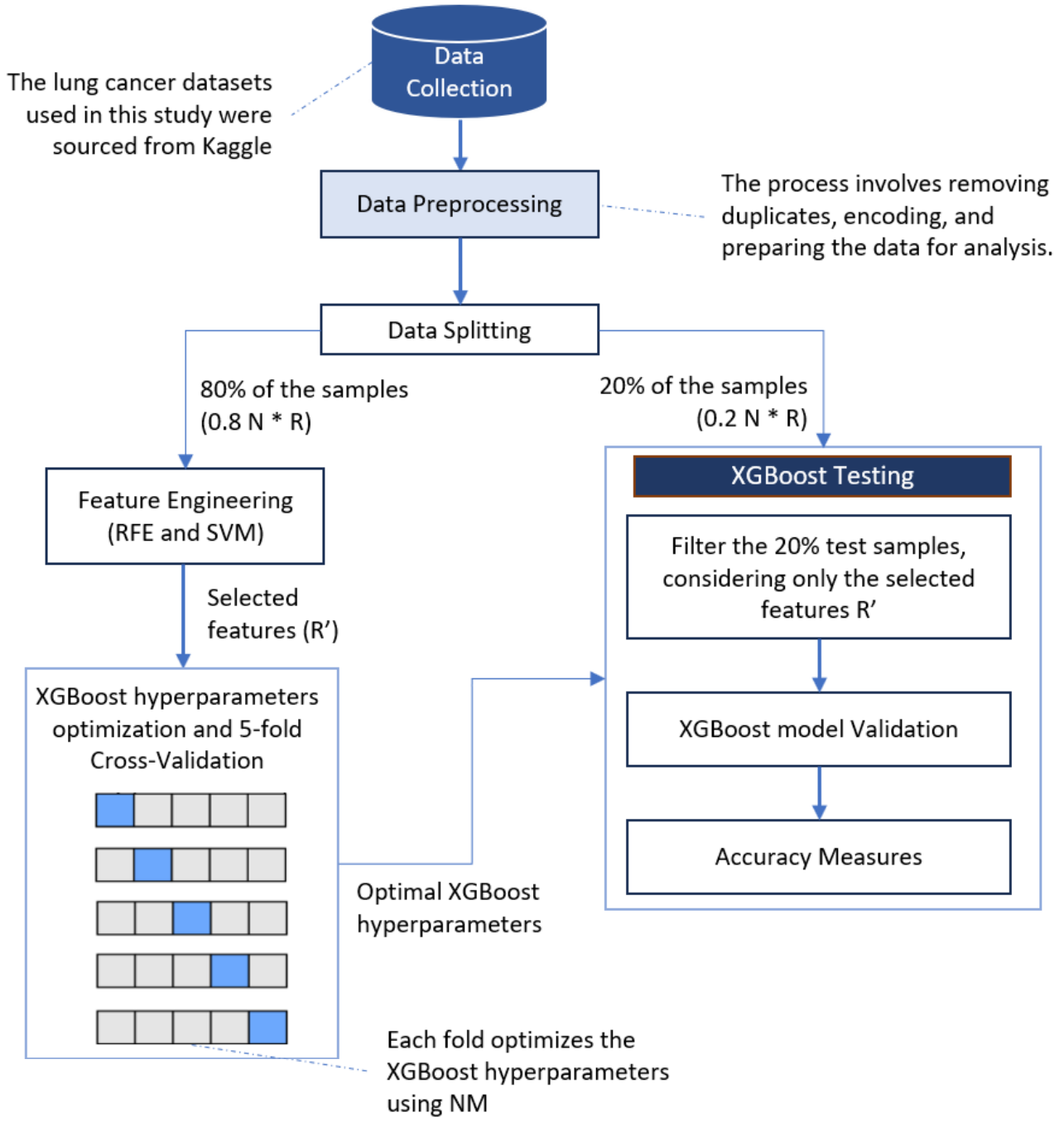


FIGURE 5. The proposed SVM-based RFE, XGBoost, and Nelder-Mead optimization methodology.

step to enhance generalization and comprehensiveness of the proposed methodology.

A. FEATURE ENGINEERING USING RFE WITH SVM

The combination of SVM and RFE assists in identifying and selecting relevant features for building an effective predictive model [41]. It also enhances the interpretability of the predictive model which would help the clinicians gain insights into which features are most influential. RFE iteratively removes the least important features based on the SVM weights until reaching the optimal number of features. In addition, RFE helps in reducing overfitting, especially when dealing with a limited amount of data which is the case of the investigated datasets [42]

The feature engineering process in this paper works as follows. The SVM weight indicates how much a feature contributes to the overall model. A feature that has a higher SVM weight has a greater influence on the decision boundary

of the model. On the other hand, the RFE rank evaluates how effectively the model functions in the absence of the feature. As a result, the model may have a low RFE rank if eliminating a feature with a high SVM weight has no significant effect on its performance. Conversely, a feature with a low SVM weight could have a high RFE rank if its absence lowers the model's accuracy. SVM weight and RFE rank, SVM weight and RFE rank provide complementary insights, with one emphasizing the feature's impact within the model and the other its significance when other features are absent.

B. XGBoost CLASSIFIER

The XGBoost algorithm is an ensemble learning method based on gradient boosting [43]. It demonstrates advantageous traits, such as fast computation, robustness, and accurate prediction in imbalanced datasets [44], [45]. Owing to its efficiency, XGBoost has been extensively applied in the prediction of various diseases [46], [47], [48]. XGBoost effectively mitigates overfitting through regularization objective function and second-order Taylor expansion of the loss function. It integrates both L1 (LASSO) and L2 (Ridge) regularization into its objective function, penalizing complexity and encouraging a simpler, more generalized model. Assume that we have a dataset denoted as D shown in Equation 1.

$$D = \{(x_i, y_i) \mid i = 1, 2, \ldots, n\} \tag{1}$$

In our study, D represents the dataset consisting of n data points. Each data point is denoted by (x_i, y_i) where x_i represents the input data and y_i represents the corresponding target or output data.

Similar to other ML and ensemble models, fine-tuning the hyperparameters of XGBoost is crucial for enhancing its accuracy and predictive performance. XGBoost provides a range of hyperparameters that can be optimized, each playing a distinct role in controlling the model's complexity, and generalization capabilities. Important hyperparameters to focus on for optimization include:

- *Number of estimators:* Determines the maximum number of weak learners in the XGBoost ensemble, affecting model complexity. Optimal values range from 50 to 1000, balancing underfitting and overfitting risks.
- *Maximum depth:* Controls the complexity of individual trees, capturing intricate feature relationships. Higher values may improve accuracy by 10% but increase overfitting risk by 15%. Typical depths range from 3 to 6, with shallow trees preferred to prevent overfitting.
- *Minimum child weight:* Sets the minimum sum of instance weights required for further tree splitting, influencing model conservatism. Values from 1 to 10 balance between model complexity and overfitting risk.
- *Colsample bylevel:* Controls feature selection randomness at each tree level, balancing model complexity and performance. Tuning this parameter finds the optimal balance for your data.
- *Subsampling ratio:* Injects randomness into training data during boosting iterations, improving generalization. Values from 0.5 to 1.0 balance randomness and model accuracy, optimized through grid or random search.

Fine-tuning the XGBoost hyperparameters allows for improved model performance and helps to mitigate overfitting [49]. In this work, we employed the Nelder-Mead algorithm to optimize the XGBoost classifier's hyperparameters [50].

C. HYPERPARAMETER OPTIMIZATION WITH NELDER-MEAD

The Nelder-Mead algorithm, a popular choice for complex optimization problems, excels at finding minima or maxima in multidimensional spaces, even when functions are nonlinear and lack derivatives [51]. It utilizes a geometric "simplex" to navigate the search space, making it ideal for problems where traditional gradient-based methods fail. Nevertheless, its effectiveness in diverse fields like engineering, science, and machine learning speaks volumes about its power for tackling challenging optimization tasks. The NM algorithm relies on four key maneuvers: expansion (χ) to push promising directions further, reflection (ρ) to bounce back from bad ones, contraction (γ) to tighten around a potential minimum, and shrinkage (α) to pull all points closer when stuck. While fixed values like 2 for expansion and 1 for reflection are sometimes used, research suggests exploring the effectiveness of these parameters for optimal performance [19].

D. MODEL DEVELOPMENT

Each dataset is divided into training and testing subsets, with a random allocation following a specific ratio of 4:1, where 80% of the data is designated for training purposes and the remaining 20% is earmarked for testing the classifier's performance. For feature selection, the training data (80%) with all variables (0.8 R * V) is used as input for the feature selection process, which identifies relevant features (V') for building an effective predictive model. The training samples with selected features (0.8 R * V') are then used to train the XGBoost model with 5-fold cross-validation. Each fold optimizes the XGBoost hyperparameters using the Nelder-Mead algorithm. Finally, the optimal hyperparameters and selected features from the SVM-based RFE process are applied in the testing phase. Here, the XGBoost classifier, using its tuned parameters, aims to maximize predictive accuracy. This integrated methodology optimizes the model's performance on both training and unseen data.

To address the class imbalance in the datasets, we fine-tuned XGBoost's `scale_pos_weight` parameter to assign higher weights to the minority class, improving the model's ability to learn from underrepresented cases and enhancing its prediction accuracy for lung cancer.

IEEE Proof

We evaluated the model using precision, recall, F1-score, AU-ROC, and MCC, providing a comprehensive assessment of its performance. These metrics helped ensure the model effectively handled the imbalance and delivered balanced predictions by minimizing false positives and considering both sensitivity and specificity.

E. PERFORMANCE EVALUATION

To evaluate the accuracy of the hybrid model on the test set, we employed various performance metrics. In addition to commonly used metrics such as accuracy, recall, and precision, we incorporated the Matthews Correlation Coefficient (MCC). This coefficient is considered a balanced measure suitable for classes of varying sizes. The performance metrics utilized are represented by the following equations:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1-Score} = \frac{Precision.Recall}{Precision + Recall} \tag{5}$$

$$\text{AUC} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FN}\right) \tag{6}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

V. RESULTS AND DISCUSSION

In the process of assessing the effectiveness of our newly introduced hybrid classifier, we used the two datasets for this purpose. The following subsections present the obtained results for the investigated datasets. In addition, we compare the NM-XGBoost performance with other models.

A. LUNG CANCER RISK DATASET

This dataset consists of 15 features in addition to the target variable. It was randomly divided into training and testing sets, with 247 samples used for training and 62 for testing. The training data was then input into the SVM-based Recursive Feature Elimination for feature selection, which identified nine features as the most relevant for predicting the target variable. Figure 6 presents the SVM weights and RFE ranks obtained through the feature selection process. Each feature has an associated weight, indicating its contribution to the SVM model. The weights are typically used to understand the influence of each feature on the model's decision-making. Each feature also has an RFE rank, which indicates its importance after the recursive feature elimination process. A lower rank generally implies higher importance. For instance, a feature with a high SVM weight, such as Fatigue, holds substantial influence in the overall model decision, as evidenced by its RFE rank of 1. The features of Smoking and Yellow fingers exhibit nearly identical SVM weights, despite their differing RFE ranks.

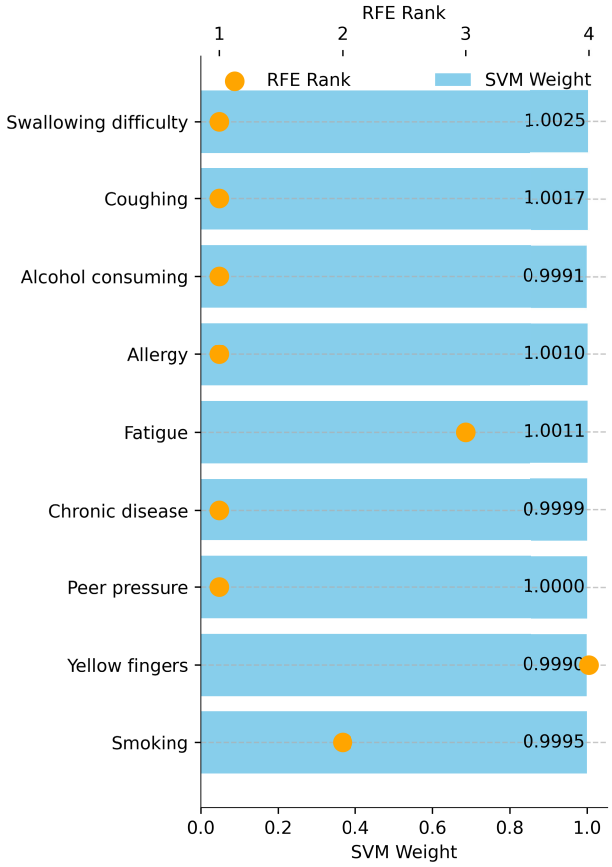


FIGURE 6. SVM weights and RFE ranks of various features - lung cancer risk dataset.

The developed model effectively identifies features that are significant indicators of lung cancer risk and symptoms from a physician's perspective. Key lifestyle factors, such as smoking and alcohol consumption, are crucial for assessing lung cancer risk, given their established links to the disease. Physical symptoms like yellow fingers and chronic fatigue provide important insights; yellow fingers indicate smoking-related damage, while fatigue is commonly reported among lung cancer patients. Swallowing difficulty may suggest advanced disease, and allergy symptoms can complicate diagnosis. Peer pressure is included to reflect its influence on smoking behavior, particularly among younger individuals. Coughing, a key symptom of lung cancer, is also integrated into the model, aligning with clinical presentations. Chronic diseases are considered to account for the overall health context, as they can exacerbate lung cancer effects. By encompassing lifestyle choices, physical symptoms, and social influences, the model enhances the ability to distinguish between benign conditions and significant health concerns, ultimately improving its accuracy in identifying at-risk individuals.

TABLE 3. Confusion matrices for lung cancer risk dataset.

Dataset	Actual / Predicted	No	Yes	Total
Training	No	24	11	35
	Yes	5	207	212
	Total	29	218	247
Testing	No	4	0	4
	Yes	0	58	58
	Total	4	58	62

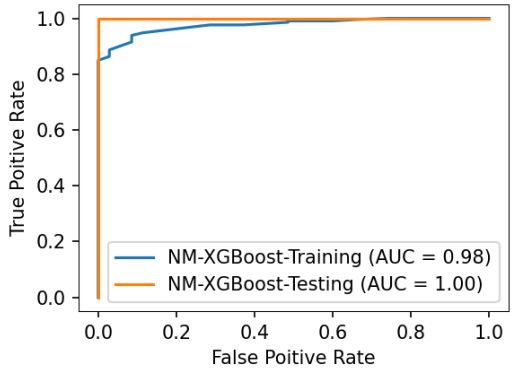


FIGURE 7. ROC analysis of NM-XGBoost on lung cancer risk dataset. ROC analysis of NM-XGBoost on lung cancer risk dataset.

The NM-XGBoost model has been trained and validated using the selected features, yielding highly promising results. The XGBoost model was optimized using the NM algorithm, resulting in optimal configuration that includes 120 estimators, a learning rate of 0.05, a maximum tree depth of 10, a subsample fraction of 0.7, and a colsample by tree value of 0.6. The optimized parameters collectively contribute to the model's performance by determining the number of boosting rounds, the step size during training, the depth of individual decision trees, and the randomness in sampling observations and features, respectively.

Table 3 demonstrates the confusion matrix that provides a comprehensive summary of both correct and incorrect predictions, indicating the count values for each class. This matrix offers insights into the types of errors originating from the classifier, encompassing the actual and predicted classes, which include positive and negative classifications. Figure 7 depicts the AUC obtained for each class during the model development phases. This observation underscores the model's adaptability and effectiveness in achieving reliable outcomes across varied datasets, thereby reinforcing its potential as an efficient tool in the domain of lung cancer classification. The model exhibited exceptional accuracy, correctly predicting all cases in the test dataset. This outstanding outcome underscores the effectiveness of the proposed approach in delivering reliable predictions for lung cancer classification based on selected features. The model's performance demonstrates its potential for further exploration in healthcare predictive modeling.

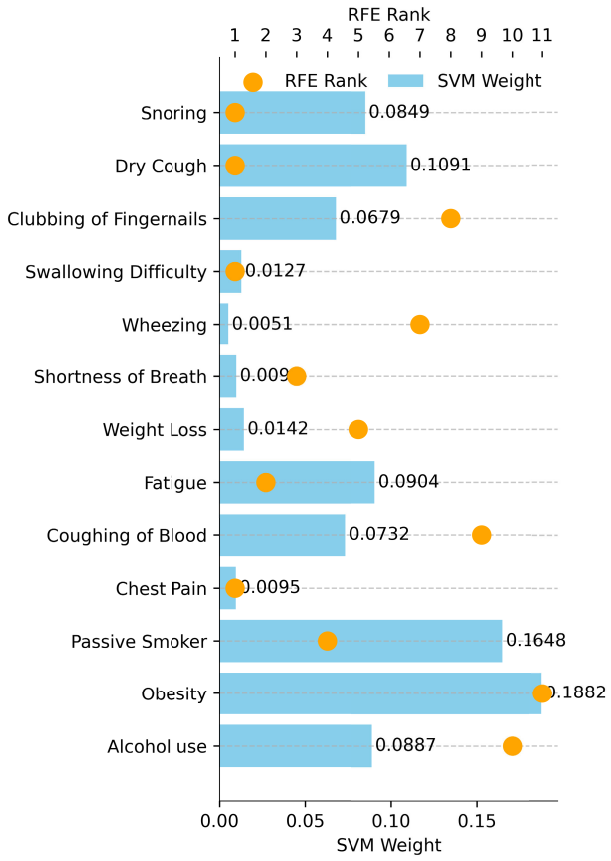


FIGURE 8. SVM weights and RFE ranks of various features - lung cancer patient dataset.

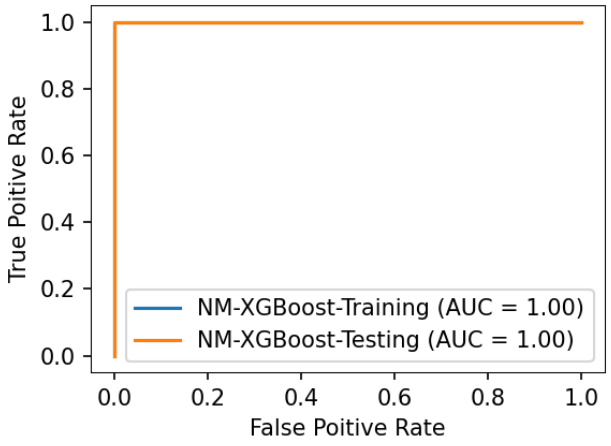


FIGURE 9. ROC analysis of NM-XGBoost on lung cancer risk dataset.

B. LUNG CANCER PATIENT DATASET

In order to conduct further validation for the proposed model and asses its generalization across diverse datasets, we extended our investigation by employing the NM-XGBoost-based classifier on an additional lung cancer dataset. The dataset comprises 1000 samples or patient

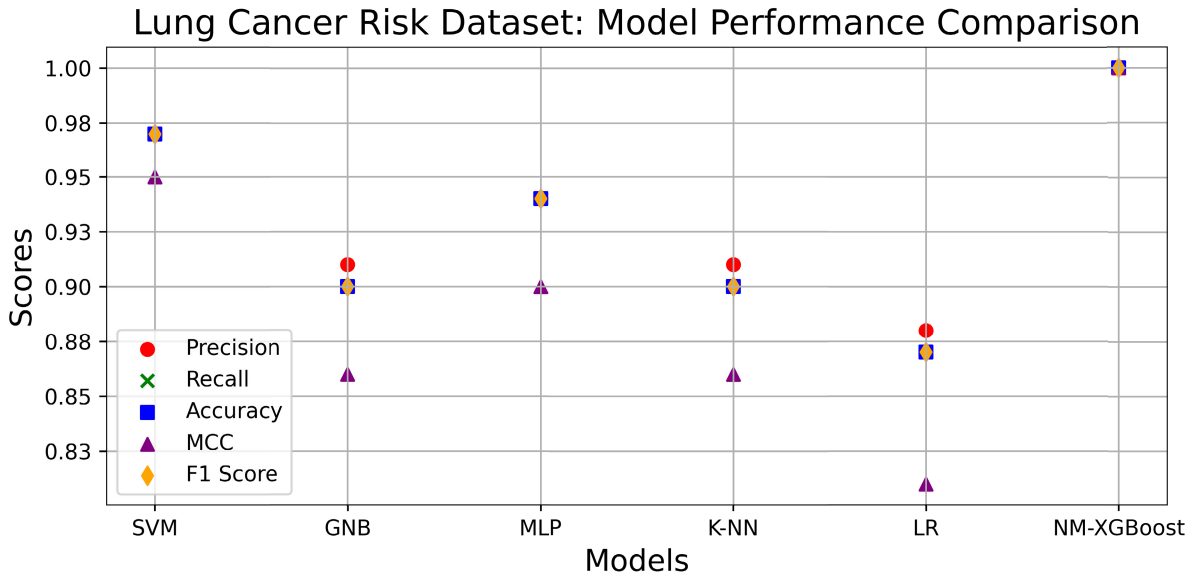


FIGURE 10. Performance comparison between the proposed model and ML models in the context of lung cancer risk dataset.

cases, each associated with 23 features, along with a target label that classifies the cases into three distinct classes. The label indicates the level of chronic lung disease of the patient (i.e., low, medium, or high). During the preprocessing phase, we identified the presence of duplicate records, which were determined by considering all columns in the dataset. We subsequently removed these duplicates, resulting in a final dataset consisting of 152 unique records. The dataset was then randomly split into 121 samples for training and 31 samples for testing.

The process of feature selection has identified 13 out of the original 23 features as relevant for predicting lung cancer. Figure 8 displays the SVM weights and RFE ranks resulting from this feature selection process for the lung cancer patient dataset. Each feature is assigned a weight, reflecting its importance in the SVM model. For the second dataset, the SVM-based RFE model selects key symptoms associated with lung cancer risk. Lifestyle factors such as alcohol use and obesity are critical, given their established links to cancer progression. Physical symptoms like chest pain, coughing of blood, fatigue, and weight loss are significant signs often reported by lung cancer patients. Shortness of breath and wheezing reflect respiratory distress, while swallowing difficulty may indicate advanced disease. Clubbing of fingernails serves as a notable physical sign of chronic respiratory conditions. Additionally, dry cough and snoring are included as they can provide insights into respiratory health, despite not being direct symptoms of lung cancer. Passive smoking is also considered, highlighting its role as a risk factor.

The XGBoost model was optimized using the NM algorithm, resulting in the optimized parameters including: 100 estimators, a learning rate of 0.1, a maximum tree depth of 5, a subsample fraction of 0.8, and a colsample by tree value of 0.8. The optimized parameters collectively contribute to the model's performance by determining the number of boosting rounds, the step size during training, the depth of individual decision trees, and the randomness in sampling observations and features. The NM-XGBoost classifier, as applied to this dataset, demonstrated its robust capacity for generating accurate predictions, as depicted in Figure 9. The figure illustrates the AUC achieved during the model development phases. Table 4 presents the confusion matrix from the classifier developed for the Lung Cancer Patient Dataset. The dataset contains 53 samples for the low-risk class, 47 for the medium-risk class, and 52 for the high-risk class.

TABLE 4. Confusion matrices for lung cancer patient dataset.

Dataset	Actual / Predicted	Low	Medium	High	Total
Training	Low	44	0	0	44
	Medium	0	37	0	37
	High	0	0	40	40
	Total	44	37	40	121
Testing	Low	9	0	0	9
	Medium	0	10	0	10
	High	0	0	12	12
	Total	9	10	12	31

C. COMPARATIVE ANALYSIS

In this section, we conducted additional validation for the proposed NM-XGBoost model by comparing its performance against several traditional ML algorithms, namely Support Vector Machine, Gaussian Naive Bayes (GNB), Multi-Layer Perceptron (MLP), k-Nearest Neighbors (K-NN), and Logistic Regression (LR), on the two investigated datasets. Figure 10 and Figure 11 provide visual representations of the

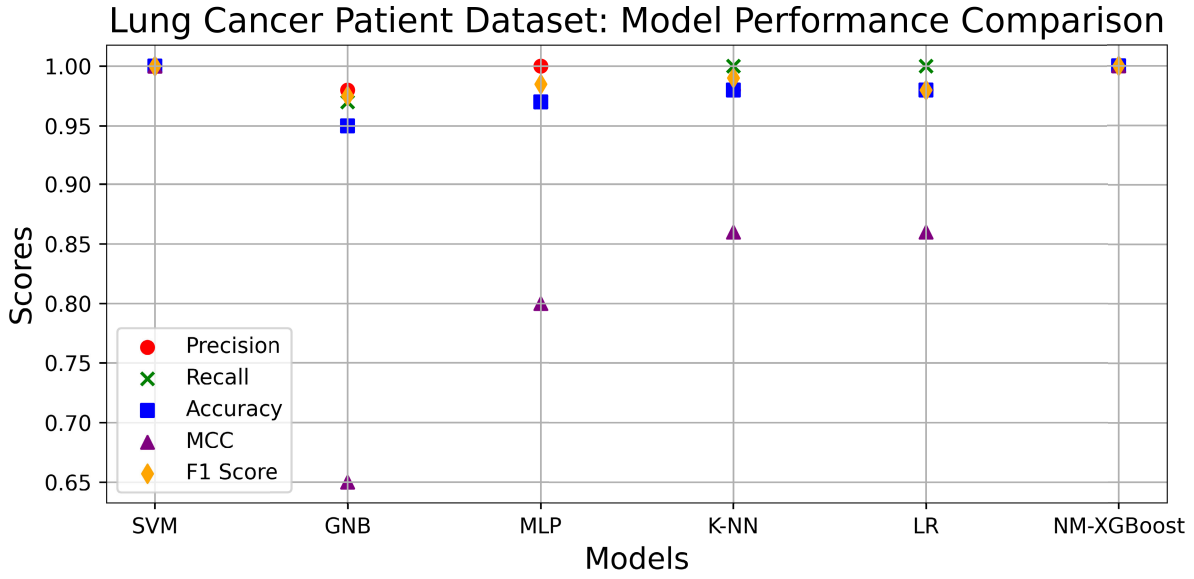


FIGURE 11. Performance comparison between the proposed model and ML models in the context of lung cancer patient dataset.

performance metrics comparisons. The performance metrics comparison revealed that the NM-XGBoost model consistently outperformed the employed traditional algorithms across various evaluation criteria. Specifically, it exhibited higher accuracy, precision, recall, and MCC than SVM, NB, MLP, K-NN, and LR. These findings underscore the NM-XGBoost model's effectiveness in handling the datasets' complexities and leveraging the optimization capabilities of the Nelder-Mead algorithm. Moreover, the superior performance of NM-XGBoost highlights its potential for application in real-world scenarios where accurate and reliable predictions are crucial. While the SVM model exhibited comparable performance in the second dataset with multiple classes, its performance varied when applied to the first imbalanced dataset. In the context of the imbalanced dataset, SVM struggled to effectively handle the class imbalance, leading to suboptimal performance compared to other models.

VI. CONCLUSION AND FUTURE WORK

Medical datasets often encompass numerous features, many of which may be irrelevant or noisy. Therefore, employing feature engineering can significantly enhance the performance of predictive models. In this paper, we explored the effectiveness of an integrated approach combining feature selection, ensemble learning, and model optimization. Specifically, Recursive Feature Elimination with SVM (RFE-SVM) was utilized to enhance feature selection by identifying the most relevant attributes. Meanwhile, the NM algorithm optimized the hyperparameters of XGBoost, resulting in a robust and accurate hybrid model.

This integrated approach was tested on two distinct lung cancer datasets, each presenting unique challenges. The first dataset, characterized by binary classes and imbalanced data, posed significant difficulties for conventional ML algorithms. In contrast, the second dataset involved multiclass classification, introducing different complexities. Compared to traditional ML methods, the proposed integrated approach demonstrated superior performance, effectively addressing the challenges of imbalanced data in medical diagnostics.

The implications of this study are significant for the field of healthcare. By demonstrating the efficacy of advanced computational techniques, we provide a pathway for improved diagnostic accuracy and decision-making in clinical settings. This integrated approach can potentially lead to earlier detection of lung cancer and better-targeted treatment strategies, ultimately enhancing patient outcomes. In future work, we plan to incorporate explainable AI (XAI) techniques to clarify the connection between clinical markers and predictions. Additionally, future research should explore applying this methodology to other chronic diseases and validating its effectiveness in real-world healthcare settings.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021.

[2] L. A. Torre, R. L. Siegel, and A. Jemal, "Lung cancer statistics," in *Lung Cancer and Personalized Medicine: Current Knowledge and Therapies*, 2016, pp. 1–19.

[3] H. Lemjabbar-Alaoui, O. U. Hassan, Y.-W. Yang, and P. Buchanan, "Lung cancer: Biology and treatment options," *Biochimica et Biophysica Acta (BBA)-Rev. Cancer*, vol. 1856, no. 2, pp. 189–210, Dec. 2015.

[4] C. R. Thomas Jr., *Lung Cancer: A Multidisciplinary Approach to Diagnosis and Management*. Demos Medical, 2010.

[5] C. Ladbury, A. Amini, A. Govindarajan, I. Mambetsariev, D. J. Raz, E. Massarelli, T. Williams, A. Rodin, and R. Salgia, "Integration of artificial intelligence in lung cancer: Rise of the machine," *Cell Rep. Med.*, vol. 4, no. 2, Feb. 2023, Art. no. 100933.

[6] S. M. Varnosfaderani and M. Forouzanfar, "The role of AI in hospitals and clinics: Transforming healthcare in the 21st century," *Bioengineering*, vol. 11, no. 4, p. 337, Mar. 2024.

[7] G. Chassagnon, C. D. Margerie-Mellon, M. Vakalopoulou, R. Marini, T.-N. Hoang-Thi, M.-P. Revel, and P. Soyer, "Artificial intelligence in lung cancer: Current applications and perspectives," *Jpn. J. Radiol.*, vol. 41, pp. 235–244, Nov. 2022.

[8] A. P. Zhao, S. Li, Z. Cao, P. J.-H. Hu, J. Wang, Y. Xiang, D. Xie, and X. Lu, "AI for science: Predicting infectious diseases," *J. Saf. Sci. Resilience*, vol. 5, no. 2, pp. 130–146, Jun. 2024.

[9] H. A. Al-Jamimi, "Synergistic feature engineering and ensemble learning for early chronic disease prediction," *IEEE Access*, vol. 12, pp. 62215–62233, 2024.

[10] Ş. Ay, E. Ekinci, and Z. Garip, "A comparative analysis of meta-heuristic optimization algorithms for feature selection on ML-based classification of heart-related diseases," *J. Supercomput.*, vol. 79, no. 11, pp. 11797–11826, Jul. 2023.

[11] S. Zhao, P. Wang, A. A. Heidari, H. Chen, W. He, and S. Xu, "Performance optimization of salp swarm algorithm for multi-threshold image segmentation: Comprehensive study of breast cancer microscopy," *Comput. Biol. Med.*, vol. 139, Dec. 2021, Art. no. 105015.

[12] H.-Y. Chiu, H.-S. Chao, and Y.-M. Chen, "Application of artificial intelligence in lung cancer," *Cancers*, vol. 14, no. 6, p. 1370, Mar. 2022.

[13] S. Huang, J. Yang, N. Shen, Q. Xu, and Q. Zhao, "Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective," in *Seminars in Cancer Biology*, vol. 89. Amsterdam, The Netherlands: Elsevier, 2023, pp. 30–37.

[14] M. Liu, J. Wu, N. Wang, X. Zhang, Y. Bai, J. Guo, L. Zhang, S. Liu, and K. Tao, "The value of artificial intelligence in the diagnosis of lung cancer: A systematic review and meta-analysis," *PLoS ONE*, vol. 18, no. 3, Mar. 2023, Art. no. e0273445.

[15] X.-W. Chen and J. C. Jeong, "Enhanced recursive feature elimination," in *Proc. 6th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2007, pp. 429–435.

[16] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Apr. 1998.

[17] W. S. Noble, "What is a support vector machine?" *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006.

[18] X. Zhou, Y. Chen, Z. Wu, A. A. Heidari, H. Chen, E. Alabdulkreem, J. Escorcia-Gutierrez, and X. Wang, "Boosted local dimensional mutation and all-dimensional neighborhood slime mould algorithm for feature selection," *Neurocomputing*, vol. 551, Sep. 2023, Art. no. 126467.

[19] F. Gao and L. Han, "Implementing the Nelder–Mead simplex algorithm with adaptive parameters," *Comput. Optim. Appl.*, vol. 51, no. 1, pp. 259–277, Jan. 2012.

[20] C. De Margerie-Mellon and G. Chassagnon, "Artificial intelligence: A critical review of applications for lung nodule and lung cancer," *Diagnostic Interventional Imag.*, vol. 104, no. 1, pp. 11–17, Jan. 2023.

[21] H. Park, J. Yun, S. M. Lee, H. J. Hwang, J. B. Seo, Y. J. Jung, J. Hwang, S. H. Lee, S. W. Lee, and N. Kim, "Deep learning—Based approach to predict pulmonary function at chest CT," *Radiology*, vol. 307, no. 2, Apr. 2023, Art. no. 221488.

[22] T. I. A. Mohamed and A. E.-S. Ezugwu, "Enhancing lung cancer classification and prediction with deep learning and multi-omics data," *IEEE Access*, vol. 12, pp. 59880–59892, 2024.

[23] S. K. Bhatt and S. Srinivasan, "Lung cancer detection using ai and different techniques of machine learning," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 8, pp. 630–638, 2024.

[24] P. Sathe, A. Mahajan, D. Patkar, and M. Verma, "End-to-end fully automated lung cancer screening system," *IEEE Access*, vol. 12, pp. 108515–108532, 2024.

[25] A. Heidari, D. Javaheri, S. Toumaj, N. J. Navimipour, M. Rezaei, and M. Unal, "A new lung cancer detection method based on the chest CT images using federated learning and blockchain systems," *Artif. Intell. Med.*, vol. 141, Jul. 2023, Art. no. 102572.

[26] M. G. Lanjewar, K. G. Panchbhai, and P. Charanarur, "Lung cancer detection from CT scans using modified DenseNet with feature selection methods and ML classifiers," *Expert Syst. Appl.*, vol. 224, Aug. 2023, Art. no. 119961.

[27] M. Mamun, M. I. Mahmud, M. Meherin, and A. Abdelgawad, "LCD-ctCNN: Lung cancer diagnosis of CT scan images using CNN based model," in *Proc. 10th Int. Conf. Signal Process. Integr. Netw. (SPIN)*, Mar. 2023, pp. 205–212.

[28] S. Wankhade and S. Vigneshwari, "A novel hybrid deep learning method for early detection of lung cancer using neural networks," *Healthcare Analytics*, vol. 3, Nov. 2023, Art. no. 100195.

[29] A. Seth and V. D. Kaushik, "Lung and colon classification using improved local Fisher discriminant analysis with ANFIS," *Int. J. Inf. Technol.*, vol. 16, no. 8, pp. 4845–4853, Dec. 2024.

[30] B. Mostafa, M. Sakr, and A. Keshk, "Employing the capabilities of LSTM and Bi-LSTM for lung cancer detection and classification," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 5, 2024.

[31] K. Priyadarshini, M. Alagarsamy, K. Sangeetha, and D. Thangaraju, "Hybrid RNN-FFBPNN optimized with glowworm swarm algorithm for lung cancer prediction," *IETE J. Res.*, vol. 70, no. 5, pp. 4453–4468, May 2024.

[32] N. A. Wani, R. Kumar, and J. Bedi, "DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence," *Comput. Methods Programs Biomed.*, vol. 243, Jan. 2024, Art. no. 107879.

[33] S. T. Rikta, K. M. M. Uddin, N. Biswas, R. Mostafiz, F. Sharmin, and S. K. Dey, "XML-GBM lung: An explainable machine learning-based application for the diagnosis of lung cancer," *J. Pathol. Informat.*, vol. 14, Apr. 2023, Art. no. 100307.

[34] Y. Li, X. Wu, P. Yang, G. Jiang, and Y. Luo, "Machine learning for lung cancer diagnosis, treatment, and prognosis," *Genomics, Proteomics Bioinf.*, vol. 20, no. 5, pp. 850–866, Oct. 2022.

[35] K. Jabir and A. Thirumurthi Raja, "Prediction of lung cancer from electronic health records using CNN supported nlp," in *Computational Intelligence for Clinical Diagnosis*. Cham, Switzerland: Springer, 2023, pp. 549–560.

[36] I. M. Nasser and S. S. Abu-Naser, "Lung cancer detection using artificial neural network," *Int. J. Eng. Inf. Syst.*, vol. 3, pp. 17–23, Mar. 2019.

[37] E. Dritsas and M. Trigka, "Lung cancer risk prediction with machine learning models," *Big Data Cognit. Comput.*, vol. 6, no. 4, p. 139, Nov. 2022.

[38] M. Mamun, A. Farjana, M. A. Mamun, and M. S. Ahammed, "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis," in *Proc. IEEE World AI IoT Congr. (AIIoT)*, Jun. 2022, pp. 187–193.

[39] (2024). *Lung Cancer Risk Dataset*. [Online]. Available: https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer

[40] (2024). *Lung Cancer Patient Dataset*. [Online]. Available: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

[41] Pamir, N. Javaid, A. Almogren, M. Adil, M. U. Javed, and M. Zuair, "RFE based feature selection and KNNOR based data balancing for electricity theft detection using BiLSTM-LogitBoost stacking ensemble model," *IEEE Access*, vol. 10, pp. 112948–112963, 2022.

[42] P. Theerthagiri and S. D. Siddalingaiah, "RG-SVM: Recursive Gaussian support vector machine based feature selection algorithm for liver disease classification," *Multimedia Tools Appl.*, vol. 83, no. 20, pp. 59021–59042, Dec. 2023.

[43] T. Chen, "XGBoost: Extreme gradient boosting," *R Package Version 0.4-2*, vol. 1, no. 4, 2015.

[44] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *Int. J. Distrib. Sensor Netw.*, vol. 18, no. 6, Jun. 2022, Art. no. 155013292211069.

[45] T.-T.-H. Le, Y. E. Oktian, and H. Kim, "XGBoost for imbalanced multiclass classification-based industrial Internet of Things intrusion detection systems," *Sustainability*, vol. 14, no. 14, p. 8707, Jul. 2022.

[46] A. Ogunleye and Q.-G. Wang, "XGBoost model for chronic kidney disease diagnosis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 6, pp. 2131–2140, Nov. 2020.

[47] A. Paleczek, D. Grochala, and A. Rydosz, "Artificial breath classification using XGBoost algorithm for diabetes detection," *Sensors*, vol. 21, no. 12, p. 4187, Jun. 2021.

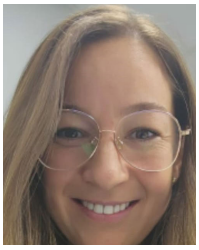
[48] V. Jain and M. Agrawal, "Heart failure prediction using XGB classifier, logistic regression and support vector classifier," in *Proc. Int. Conf. Advancement Comput. Comput. Technol. (InCACCT)*, May 2023, pp. 1–5.

[49] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, Nov. 2020.

IEEE Proof

[50] E. Kusuma, G. Shidik, and R. Pramunendar, "Optimization of neural network using Nelder Mead in breast cancer classification," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 6, pp. 330–337, Dec. 2020.
[51] M. A. Luersen and R. L. Riche, "Globalized Nelder–Mead method for engineering optimization," *Comput. Struct.*, vol. 82, no. 26, pp. 2251–2260, Aug. 2004.

HAMDI A. AL-JAMIMI received the Ph.D. degree in computer science and engineering, in 2015. He is currently a Distinguished Scientist and an Academician with the King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia. He has since built an impressive publication portfolio in renowned journals and premier conferences. Specializing in the intersection of artificial intelligence and data science, he explores cutting-edge applications and methodologies, with a particular emphasis on the healthcare, petrochemicals, and engineering sectors. His research interests include advancing knowledge and fostering technological innovation in these crucial domains. Committed to collaborative research, he actively cultivates interdisciplinary partnerships that yield meaningful societal impact. His unwavering dedication to excellence and insatiable quest for knowledge underscore his substantial contributions to the scientific community and broader academic landscape.



SARAH AYAD was born in Lebanon, in 1987. She received the M.S. degree in computer science from Université de Versailles Saint-Quentin-en-Yvelines (UVSQ), Versailles, France, in 2009, and the Ph.D. degree in semantic quality of business process models from Conservatoire National des Arts et Métiers (CNAM), Paris, France, in 2013. From 2013 to 2019, she was a Research Assistant with Lebanese University, Lebanon. Since 2019, she has been an Assistant Professor with Arab Open University, Saudi Arabia. Her recent research focuses on using machine learning and artificial intelligence (AI) to enhance the semantic quality of domain ontologies and business process models. Her work has been published in three Q1 journals and presented at four international conferences. Her research interests include the application of machine learning algorithms and large language models to improve the semantic quality of domain ontologies. She is also dedicated to enhancing the syntactic, semantic, and pragmatic quality of business process models, integrating advanced AI techniques to address complex challenges in these areas.

AMMAR EL KHEIR received the Diploma degree in immunology and allergic diseases from Université Libre de Bruxelles (ULB), Belgium. He is a highly skilled and specialized pediatric consultant with expertise in pediatric respiratory and allergic diseases. Currently, he is a full-time Consultant with Liege University Hospital Center, Liège, specializing in pediatric respiratory and allergic conditions. He holds a general medicine diploma followed by a pediatric diploma from the Lebanese University, Beirut. He further advanced his training with a two-year fellowship in pediatric pulmonary and allergic diseases with HUDERF Pediatric Hospital, Brussels, complemented by a one-year tenure at the Pediatric Hospital of Jeanne de Flandre, Lille, France. In addition to his clinical training, he is certified in pediatric respiratory medicine through the European Respiratory Society (ERS). He is also an Active Member of the Primary Ciliary Dyskinesia Diagnostic Center, Liège, contributing to advanced diagnostic and therapeutic approaches in this rare condition. With a commitment to evidence-based care and a strong background in both clinical and academic settings across Europe, he is dedicated to improving pediatric patient outcomes through cutting-edge respiratory and allergic disease management.

IEEE Proof