Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics e-ISSN: 2656-8624
Homepage: https://ijeeemi.org/; Vol. 8, No. 4, pp. 633-651, November 2026

RESEARCH PAPER **OPEN ACCESS**

A Comparative Study of Feature Selection in Machine Learning for Survey-Based Lung Cancer Classification Utilizing SHAP

Irwan Budiman, Nanda Putri Rahmawati, Annisa Rizqiana, and Rafi'e

Department of Computer Science, Faculty of Mathematics and Natural Science, Lambung Mangkurat University, Banjarbaru, Indonesia

Corresponding author: Irwan Budiman (e-mail: irwan.budiman@ulm.ac.id), **Author(s) Email**: Irwan Budiman (e-mail: irwan.budiman@ulm.ac.id), Nanda Putri Rahmawati (e-mail: nandaputri15406@gmail.com), Annisa Rizqiana (e-mail: annisarizqiana@ulm.ac.id), and Rafi'e (e-mail: rafiekom@ulm.ac.id)

Abstract **Lung cancer is a fatal malignancy requiring early detection. While machine learning aids in diagnosis using survey data, existing studies frequently suffer from extreme class imbalance and methodological flaws, specifically data leakage during feature selection, leading to optimistically biased performance. This study aims to determine the most robust feature selection and classification methodology for lung cancer prediction while strictly eliminating information leakage and addressing class imbalance. Duplicate entries were removed from a 309-record survey dataset, resulting in 276 unique patients (86.23% positive, 13.77% negative). An unbiased evaluation framework was established using Nested Stratified 10-Fold Cross-Validation. Within this architecture, nine classifiers were optimized across five dynamic feature selection techniques (ANOVA + Chi-Square, Forward Selection, Backward Elimination, Mutual Information Gain, and Recursive Feature Elimination) strictly within the inner training loops. The K-Nearest Neighbors algorithm paired with Recursive Feature Elimination (KNN-RFE) achieved the highest overall predictive accuracy (91.28% ± 4.91%) and F1-score (95.02% ± 2.81%). However, Logistic Regression optimized with Backward Elimination (LR-BE) demonstrated superior stability in the presence of skewed data, achieving the highest balanced accuracy (89.95% ± 8.42%) and precision (98.25% ± 2.85%). Additionally, SHAP (SHapley Additive exPlanations) analysis was integrated, transparently identifying 'Swallowing Difficulty' and 'Coughing' as the most critical predictive biomarkers. In practice, KNN-RFE is recommended for maximizing overall diagnostic discrimination, whereas LR-BE is strictly preferred for handling severe class imbalances to minimize false positives. The rigorous nested validation ensures these metrics reflect realistic clinical generalization rather than algorithmic overfitting. Furthermore, the SHAP integration provides necessary transparency, translating complex predictive pathways into interpretable, evidence-based screening insights for healthcare professionals.**

Keywords **Lung Cancer; Feature Selection; Machine Learning; Nested CV; SHAP**

I. Introduction

As a formidable challenge to global public health, lung cancer consistently ranks as one of the most frequently diagnosed and fatal malignancies worldwide. The pathology is fundamentally driven by the unchecked proliferation of pulmonary cells, a localized malignancy that frequently escalates into systemic disease through metastasis [1]. Ranked just behind breast cancer for women and prostate cancer for men, it holds the position of the second most widespread malignancy globally, underscored by 2020 statistical projections revealing roughly 2.2 million new diagnoses and 1.8 million associated deaths [2]. Completely averting pulmonary cancer remains an impossible task, and once a patient receives a diagnosis, the chances of living past the five-year mark are strikingly low, hovering around 15.5 percent [3]. Consequently, identifying early warning signs is critical; patients typically present with a multitude of surveyable symptomsincluding chest aches, severe coughing fits, unexplained weight loss, exhaustion, and loss of appetitewhich medical professionals use to determine if a patient needs to be moved forward for cancer treatment [4], [5]. To catch cancer early, predict life expectancy, and map out treatment success, the medical community relies on advanced screening and classification methods to spot tumors long before a person starts feeling severe symptoms [6]. In this context, ensemble methods represent a specialized branch of machine learning that enhances overall predictive performance by aggregating the outputs of several foundational models, rather than relying on just one [7]. Given the complexities of analyzing survey-based symptom data, both standalone machine learning and ensemble algorithms have become essential tools for conducting high-stakes diagnostic assessments [8].

Prior studies [9] analyzed an array of standard classifiers, such as K-Nearest Neighbors, Naïve Bayes, Support Vector Machines, Decision Trees, and Logistic Regression. Among these, the Random Forest model demonstrated superior predictive capabilities, recording both an accuracy and recall of 90.32%, alongside a precision of 89.82% and an F1-score of 87.09%. The current research utilizes that identical pulmonary oncology dataset, which is composed of 309 patient samples mapped across 15 attributes and one specific target variable. The dataset variables are described in

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 1. Highlighting the importance of feature selection, another contemporary study applied ANOVA, Chi-Square, and the mRMR method to extract the most highly correlated clinical attributes. The selection pipeline narrowed the data down to nine essential diagnostic markers, namely swallowing difficulty, peer pressure, gender, allergies, yellow fingers, anxiety, wheezing, alcohol consumption, and chronic disease. When trained exclusively on these selected variables using a 5-fold cross-validation scheme, the Naïve Bayes algorithm reached a peak classification accuracy of 82.81% [8].

In a related study, researchers [1] evaluated the performance of ensemble learning techniquesspecifically AdaBoost and Baggingon a lung cancer dataset without the implementation of any feature selection methods. Evaluated through a rigorous 10-fold cross-validation protocol, the AdaBoost model demonstrated superior predictive capabilities, achieving an accuracy of 90.70%, a precision of 92.53%, a recall of 90.26%, and an F1-score of 90.89%. Meanwhile, the Bagging algorithm also delivered strong results, recording an overall accuracy of 89.76%, alongside precision, recall, and F1-score metrics of 91.88%, 89.35%, and 90.00%, respectively. In another comparative analysis, researchers [10] utilized a 10-fold cross-validation approach to evaluate four foundational classifiers: Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), and K-Nearest Neighbors (KNN). Among the evaluated models, SVM achieved the highest predictive performance with an accuracy of 90.9%, followed by NB at 88.7%, DT at 87.4%, and KNN at 85.5%.

Despite the high standard accuracies reported in the aforementioned literature, numerous studies applying machine learning to this specific dataset suffer from critical, unresolved methodological vulnerabilities. Most notably, feature selection is frequently executed entirely outside the validation cycle, which introduces severe information leakage and yields optimistically biased performance estimates. Furthermore, models are often evaluated on small sample sizes characterized by extreme class imbalance without utilizing robust control mechanisms like Stratified K-Fold cross-validationa technique proven to provide significantly more reliable and unbiased estimations by strictly preserving class distribution ratios across all data partitions [11] successfully implemented this exact technique on a Finally, the inherent lack of interpretability remains a significant concern; relying on black-box models without implementing Explainable Artificial Intelligence (XAI) limits the ability to systematically understand the precise contribution of each survey-based feature to individual predictions [12], [13].

To address these gaps, this study proposes a Nested Stratified 10-Fold Cross-Validation framework that integrates hyperparameter tuning and dynamic feature selection within inner training folds to prevent data leakage and ensure robust evaluation of nine standard and ensemble classifiers. Additionally, SHAP is applied to the best-performing model to provide interpretable, patient-specific diagnostic insights from survey-based biomarkers.

The remainder of this paper is organized as follows: Section II presents the dataset and proposed methodology; Section III reports the experimental results; Section IV discusses the findings, comparisons with previous studies, and limitations; and Section V concludes the study and outlines future work.

II. Materials and Method

A. Dataset

Table 1. Description of the survey-based lung cancer dataset, including predictor variables, clinical characteristics, measurement scales, categorical encoding, and target class definitions.

Features	Description
Gender	The patient's biological sex (M, F)
AGE	The patient's age in years (21-87)
SMOKING	Indicates whether the patient smokes (1, 2)
YELLOW FINGERS	Presence of yellow stains on fingers (1, 2)
ANXIETY	Indicates if the patient frequently experiences anxiety (1, 2)
PEER PRESSURE	Indicates if the patient is subjected to peer pressure (1, 2)
CHRONIC DISEASE	Presence of any other chronic medical conditions (1, 2)
FATIGUE	Indicates if the patient suffers from chronic or unusual fatigue (1, 2)
ALLERGY	Indicates if the patient has any known allergies (1, 2)
WHEEZING	Presence of a whistling or wheezing sound while breathing (1, 2)
ALCOHOL	Indicates if the patient regularly consumes alcohol (1, 2)
COUGHING	Presence of a persistent or chronic cough (1, 2)
SHORTNESS OF BREATH	Indicates if the patient frequently experiences difficulty breathing (1, 2)
SWALLOWING DIFFICULTY	Indicates if the patient has trouble swallowing food or liquids (1,2)
CHEST PAIN	Presence of pain, tightness, or discomfort in the chest (1, 2)
LUNG CANCER	Target: lung cancer (No, Yes)

The empirical data utilized in this research are derived from the "Survey Lung Cancer" dataset available on the Kaggle repository. Architecturally, the dataset comprises 15 independent predictive variables and a single

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

diagnostic target label indicating the presence of lung cancer, with 309 rows. Similar to the work in [10], this study utilizes the identical dataset and evaluates the same four machine learning algorithms: Support Vector Machine, K-Nearest Neighbors, Naïve Bayes, and Decision Tree. The characteristics of the dataset's features are summarized in Table 1.

B. Preprocessing

During the data preprocessing phase, 33 duplicate rows were removed to prevent data leakage during model training. Subsequently, all categorical attributes were standardized into a binary numerical format. GENDER ('M'/'F') and LUNG_CANCER ('YES'/'NO') were converted such that Male and 'YES' were encoded as 1, while Female and 'NO' were encoded as 0. Additionally, the remaining features, which initially contained values of 1 ('No') and 2 ('Yes'), were re-encoded to 0 and 1, respectively. Table 2 illustrates the initial and final structure of the dataset following the elimination of these redundant records and the numerical conversion. To prepare the dataset prior to model training and evaluation, rigorous data preprocessing and normalization steps must be applied. Continuous variables are standardized using the Z-score method, which is a statistical metric that quantifies the number of standard deviations a specific data point deviates from the dataset's average value. The scaling transformation is formally expressed in Eq.(1) [14]:

$$Z = \frac{x-\mu}{\sigma} \quad (1)$$

where Z represents the standardized feature (Z-score), x denotes the raw numerical data point, μ is the mean of the dataset for the respective feature calculated strictly within the inner training fold, and σ represents its corresponding standard deviation [14].

While Z-score standardization is applied to the majority of the classification models, algorithms requiring strictly non-negative input values, specifically the Naïve Bayes classifier and the Chi-Square feature selection test, utilize Min-Max normalization. This alternative scaling technique transforms the features into a fixed bounded range of [0, 1] to prevent numeric features with larger scales from dominating the model, thereby giving features equal initial importance. The Min-Max normalization is formulated mathematically in Eq.(2) [15]:

$$x'_{ij} = \frac{x_{ij}-\min(x_j)}{\max(x_j)-\min(x_j)} \quad (2)$$

where $min(x_j)$ and $max(x_j)$ denote the minimum and maximum observed valuea of feature j within the training partition, respectively [15].

C. Feature Selection and SHAP

Table 2. Comparison of dataset characteristics before and after duplicate removal, including sample size, predictor distributions, class proportions, and missing-value counts.

Variable	Before: 2 / YES / M (%)	Before: 1 / NO / F (%)	Missing before	After: 1 (%)	After: 0 (%)	Missing after
GENDER	52.4	47.6	0	51.4	48.6	0
AGE (21-87)	62.67 (mean)	8,21 (Std)	0	62.91 (mean)	8.38(Std)	0
SMOKING	56.3	43.7	0	54.3	45.7	0
YELLOWFINGERS	57.0	43.0	0	57.6	42.4	0
ANXIETY	49.8	50.2	0	49.6	50.4	0
PEER_PRESSURE	50.2	49.8	0	50.7	49.3	0
CHRONICDISEASE	50.5	49.5	0	52.2	47.8	0
FATIGUE	67.3	32.7	0	66.3	33.7	0
ALLERGY	55.7	44.3	0	54.7	45.3	0
WHEEZING	55.7	44.3	0	54.7	45.3	0
ALCOHOL CONSUMING	55.7	44.3	0	55.1	44.9	0
COUGHING	57.9	42.1	0	57.6	42.4	0
SHORTNESS OF BREATH	64.1	35.9	0	63.0	37.0	0
SWALLOWING DIFFICULTY	46.9	53.1	0	46.7	53.3	0
CHEST PAIN	55.7	44.3	0	55.8	44.2	0
LUNG_CANCER	87.38	12.62	0	86.23	13.77	0

In this study, we will apply three scenarios, using feature selection to compare them with the baseline model.

1. Analysis of Variance (ANOVA)

Feature selection can assist with data visualization and understanding and minimize measurement and storage needs, training and utilization times, and overcome the curse of dimensionality to enhance prediction performance. ANOVA is a well-known statistical approach for comparing several independent means; the ANOVA ranks features by calculating the ratio of variances between and within groups [16]. The ratio represents the degree of significance between the λ-th feature and the target variables. The equation below in Eq.(3) is utilized

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

to compute the F-value for the λ-th feature in the dataset [16]:

$$F(\lambda) = \frac{s_B^2(\lambda)}{s_W^2(\lambda)}, \quad (3)$$

where $s_B^2(\lambda)$ and $s_W^2(\lambda)$ denote the sample variance between groups (Mean Square Between, MSB) and within groups (Mean Square Within, MSW), respectively, as calculated in Eq.(4) and Eq.(5) [16]:

$$s_B^2(\lambda) = \sum_{i=1}^{K} n_i \frac{\left(\left(\sum_{j=1}^{n_i} f_{ij}(\lambda)/n_i\right) - \left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} f_{ij}(\lambda)/\sum_{i=1}^{K} n_i\right)\right)^2}{df_b}, \quad (4)$$

$$s_w{}^2(\lambda) = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(f_{ij}(\lambda) - \frac{\sum_{r=1}^{n_i} f_{ir}(\lambda)}{n_i}\right)^2}{df_w}, \quad (5)$$

Aligning with the methodology employed by [17], this study utilizes Analysis of Variance (ANOVA) to assess continuous variables, while categorical data are examined using the Chi-Square test.

2. Chi-Square

Isolating the most prognostic features within the datasets is a critical step to enhance the overall accuracy of the model. This selection process filters the data to retain only those attributes that provide significant discriminative value to the machine learning classifier, effectively discarding extraneous and noisy elements [18]. To accomplish this, the Chi-square (X^2) test can be applied as a nonparametric method to evaluate the underlying relationship between variables. In the context of feature selection, this approach measures the divergence between observed and expected frequencies to calculate a statistical score for each predictor relative to the target output, ensuring that only the highest-ranking features are selected for model training [19]. To quantify this deviation, the X^2 statistic is computed using the following formula in Eq.(6) [18]:

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (6)$$

3. Sequential Feature Selection

Sequential Feature Selection is a widely recognized wrapper method utilized for feature optimization in machine learning. Operating as an iterative algorithm, it begins with an empty set and systematically incorporates individual features based on the classifier's performance on the training data. During each iteration, the variable that yields the most substantial performance improvement is added to the subset. This incremental process continues until a predefined stopping criterion is satisfied , such as reaching a maximum feature count or a specific performance threshold. Ultimately, these curated variables form the definitive subset employed in the final classification model [13]. Algorithm 1 presents the corresponding pseudocode. Sequential Feature Selection encompasses two primary methodologies. In contrast to the previously detailed forward approach, in this study, Forward Selection (FS) and Backward Elimination (BE) are mentioned. Backward Elimination (BE) initiates with the complete feature space. During each iteration, the algorithm evaluates the model's performance by systematically excluding each feature. The feature whose removal results in the most negligible decrease in the evaluation metric, indicating minimal predictive value, is

Algorithm 1: Forward Selection for identifying optimal predictive features through iterative feature addition and cross-validation-based classification accuracy optimization.

Input: Feature matrix X, labels y, classifier M; k = 10 folds; metric = accuracy.
Output: Feature subset with the highest recorded accuracy.
Function ForwardSelection(X, y, M):
  S ← ∅; A ← all features; H ← empty history
  while A ≠ ∅ **do**
    best_score ← −∞; best_feature ← None
    for each feature f in A **do**
      T ← S ∪ {f}
      score ← mean accuracy of M on X[:, T], y using k-fold cross-validation
      if score > best_score **then**
        best_score ← score; best_feature ← f
      end
    end
    S ← S ∪ {best_feature}
    A ← A \ {best_feature}
    Append (S, best_score) to H
  end
  return the subset in H with the highest accuracy
End

Algorithm 2: Backward Elimination for identifying optimal predictive features through iterative feature removal and cross-validation-based classification accuracy optimization.

Input: Feature matrix X, labels y, classifier M; k = 10 folds; metric = accuracy.
Output: Feature subset with the highest recorded accuracy.
Function BackwardElimination(X, y, M):
  S ← all features; H ← empty history
  while |S| > 1 **do**
    best_score ← −∞; best_feature ← None
    for each feature f in S **do**
      T ← S \ {f}
      score ← mean accuracy of M on X[:, T], y using k-fold cross-validation
      if score > best_score **then**
        best_score ← score; best_feature ← f
      end
    end
    S ← S \ {best_feature}
    Append (S, best_score) to H
  end
  return the subset in H with the highest accuracy
End

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

permanently discarded [20]. Algorithm 2 presents the corresponding pseudocode.

The two pseudocodes in Algorithm 1 and Algorithm 2 highlight contrasting search directions in wrapper-based feature selection. Forward Selection employs a constructive approachstarting with an empty feature space and iteratively adding attributes that maximize model performance. Conversely, Backward Elimination utilizes a destructive approachinitializing with all available features and systematically pruning those that contribute the least to predictive accuracy.

4. Mutual Information Gain (MIG)

Mutual Information (MI) is a filter-based feature selection method that identifies informative variables by measuring their dependence on the target [21]. An MI value of zero indicates independence, while higher values indicate stronger dependence. MI uses marginal and joint probability distributions for discrete variables and extends to continuous variables by replacing summation with integration [19]. To mathematically quantify this mutual dependence between two discrete variables, X and Y, the MI is formulated in Eq.(7) as follows [19]:

$$I\,(Y:X) = \sum_i \sum_j P(X_i, Y_j) \log_2 \frac{P(X_i,Y_j)}{P(X_i).P(Y_j)} \quad (7)$$

5. Recursive Feature Elimination (RFE)

Algorithm 3: Recursive Feature Elimination for ranking predictive variables and identifying the optimal feature subset using estimator-based feature importance and cross-validation.

Input: Feature matrix X, labels y, classifier M; k = 10 folds; s = 1 feature per step.
Output: Ranking from least to most important and best feature subset.
Function RecursiveElimination(X, y, M):
  if M is SVC, KNN, or Naïve Bayes **then**
    E ← Random Forest importance estimator
  else
    E ← M
  end
  S ← all features; R ← []; H ← []
  while |S| > 0 **do**
    score ← mean accuracy of E on X[:, S], y using k-fold cross-validation
    Append (S, score) to H
    Fit E on X[:, S], y to obtain feature weights w
    Compute importance I(f) ← w(f)² **for each** f in S
    f_min ← feature with the lowest importance
    Append f_min to R
    S ← S \ {f_min}
  end
  S_best ← subset in H with the highest accuracy
  return R, S_best
End

As a wrapper-based feature selection technique, Recursive Feature Elimination (RFE) utilizes a designated classifier to evaluate the dataset. The algorithm operates by computing an importance score for every variable and systematically pruning the lowest-ranking feature. This elimination cycle is repeated iteratively until the feature subset is reduced to the predetermined target size [22].

Algorithm 3 summarizes the RFE workflow: in each iteration, the model is trained using the current set of predictors, and the feature with the lowest rankingdetermined by the square of its weight elementis removed due to its minimal contribution. This cycle continues until all traits are exhausted. Once all iterations are complete, the features are sorted in reverse order of elimination, successfully highlighting the most crucial predictors [23].

6. Shapley Additive Explanation

SHAP (SHapley Additive exPlanations) is a unified framework for local interpretability that is fundamentally rooted in the cooperative game theory concept of Shapley values. Within the context of machine learning, the individual input features are treated as "players" cooperating in a game, while the model's ultimate prediction represents the final "outcome" or reward [12], [24]. The framework provides a transparent method to interpret complex models by evaluating how each specific feature contributes, whether positively or negatively, to the prediction of a given instance [13]. The Shapley value Eq.(8) for feature X_j in a model is given by [24]:

$$Shapley\ Xj\ = \sum_{S \subseteq N \setminus \{j\}} \frac{k!(p-k-1)!}{p!} \big(f(S \cup \{j\}) - f(S)\big) \quad (8)$$

In this formulation, p denotes the total number of features. The term $N\backslash\{j\}$ represents the set of all possible feature combinations that omit the specific feature X_j while S refers to a single subset chosen from $N\backslash\{j\}$. Additionally, $f(S)$ indicates the model's predicted output when utilizing only the features within subset S, whereas $f(S \cup \{j\})$ reflects the model's prediction once feature X_j is incorporated into that subset.

Beyond individual local explanations, aggregating these SHAP values across all training instances provides a deeper understanding of global feature importance. This global aggregation is computed by taking the mean of the absolute SHAP values for each feature across the entire dataset, as formalized in Eq.(9) [25]:

$$I_i = \frac{1}{M} \sum_{j=1}^{M} |\phi_i^{(j)}| \quad (9)$$

where I_i represents the global importance of feature i, M is the total number of instances in the dataset, and $\phi_i^{(j)}$ is the SHAP value of feature i for the j-th instance. This aggregation helps identify the primary features that drive the classifier's overall behavior. By visually translating these complex interactions, SHAP makes the model's decision-making process highly transparent, making it significantly easier for health practitioners and policymakers to understand the results and make informed, data-driven medical decisions [13], [25].

D. Classification Model

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

The classification modeling phase in this study utilizes nine different supervised machine learning algorithms to evaluate and predict the target variable. The selected algorithms range from traditional baseline methods, such as Logistic Regression and K-Nearest Neighbors, to advanced ensemble techniques like Random Forest and XGBoost. By comparing these diverse models, this study aims to identify the most accurate and effective classifier for the given dataset.

1. Logistic Regression

Logistic Regression (LR) is a foundational binary classification algorithm designed to estimate the probability of a specific categorical outcome. Due to its interpretability and robust predictive capabilities, it is widely utilized across various machine learning domains, including medical diagnostics, sentiment analysis, and anomaly detection [26]. The core objective of this technique is to establish a highly accurate mapping between variables while adhering to the principle of parsimony utilizing the minimum number of predictors required for an optimal fit [27]. The probability function is formally defined in Eq.(10) [26]:

$$P(Y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta^T x)}} \quad (10)$$

where $P(Y=1|x)$ denotes the predicted probability of the positive class given the input feature vector $x = [x_1, x_2, \dots, x_p]^T$. The parameter β_0 represents the intercept term, and $\beta = [\beta_1, \beta_2, \dots, \beta_p]^T$ denotes the vector of learned coefficient weights indicating the strength and direction of the relationship for each predictor. The model estimates these parameters utilizing maximum likelihood estimation to find the specific set of parameters that maximizes the probability of the observed data [28].

2. Decision Tree

The Decision Tree is a supervised classification algorithm characterized by its hierarchical decision-making framework. Its major strength is the inherent ability to partition data by applying different subsets of features and sequential conditional rules at various stages of the modeling pipeline [29]. The Decision Tree is a classification method in the form of a tree structure with nodes representing decisions or predictions, at each node, the algorithm divides the data based on the most informative input variables [30]. The Decision Tree partitions feature space hierarchically based on impurity reduction. The Gini Impurity index at a node is calculated via Eq.(11) [31]:

$$G(t) = 1 - \sum_{k=1}^{C} p_k^2 \quad (11)$$

where $G(t)$ is the node impurity score, C represents the total number of target classes $(C = 2)$, and p_k is the proportion of samples at node t belonging to class k [31].

3. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a fundamental supervised machine learning algorithm primarily applied to classification tasks. As an instance-based learning method, KNN categorizes an unlabeled data point by evaluating its proximity to existing labeled instances in the training set. Specifically, the algorithm identifies the k - closest data points (neighbors) to the target query and assigns a final class label based on a majority voting mechanism among those neighbors. Due to its structural simplicity, high adaptability, and intuitive design, KNN is widely implemented across various domainsincluding medical diagnostics to resolve both classification and regression challenges involving diverse data sizes, dimensionalities, and noise levels [32]. To mathematically quantify this proximity between a query vector x_i and a reference training vector x_j across p selected predictors, the Euclidean distance metric is formulated in Eq. (12) [32]:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{p}(x_{il} - x_{jl})^2} \quad (12)$$

where $d(x_i, x_j)$represents the calculated distance, and x_{il} and x_{jl} denote the specific attribute values for instances i and j on the l-th dimension [33].

4. Naïve Bayes

The Naive Bayes classifier is a probabilistic algorithm derived from Bayes' theorem. It leverages statistical probability calculations learned during the training phase to categorize new instances, consistently assigning data points to the class with the maximum calculated probability [33]. The fundamental bayes formula is given in Eq.(13) [33]:

$$P(A|B) = \frac{P(B|A)\times P(A)}{P(B)} \quad (13)$$

where $P(A|B)$ represents the probability of occurrence of event A when event B occurs, and $P(B|A)$ is the probability of occurrence of event B when event A occurs. Furthermore, $P(A)$ and $P(B)$ are the preliminary probabilities of A and B events. In the context of this classification task, event A corresponds to the target diagnostic class (e.g., the presence of lung cancer), while event B represents the observed survey-based feature vector.

5. Support Vector Machine

The primary advantage of a Support Vector Machine (SVM) classifier is its capacity to construct an optimal decision boundary that maximizes the margin between different classes. This optimal hyperplane is defined exclusively by a critical subset of training instances known as support vectors (SVs). Consequently, the model's decision function is entirely dependent on these pivotal points; any training samples that do not function as support vectors can be discarded without altering the geometry of the final classification boundary. To find this optimal hyperplane in real-world scenarios where datasets may not be completely linearly separable, the algorithm solves a soft margin optimization problem. This approach relies on a quadratic programming model that minimizes the weight vector while introducing a penalty parameter C and non-negative slack variables ξ_i to measure and tolerate misclassification errors. Furthermore, to address nonlinear classification

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

challenges, the input vectors are mapped into a higher-dimensional feature space using a mapping function $\phi(x_i)$, a technique widely known as the "kernel trick". The mathematical formulation to determine this optimal separating hyperplane is expressed as minimizing the objective function in Eq. (14), subject to the constraints in Eq. (15) and Eq. (16) [34]:

$$min\frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i \quad (14)$$

$$y_i(w^T\phi(x_i) + b) - 1 + \xi_i \geq 0,\ i = 1, \dots, l \quad (15)$$

$$\xi_i \geq 0 \quad i = 1, \dots, l \quad (16)$$

Where w represents the weight vector normal to the hyperplane, b denotes the bias term and l corresponds to the total number of instances in the training dataset [34].

6. Random Forest

Random Forest is an ensemble classification method that mitigates overfitting by combining numerous decision trees. Characterized by a computational complexity of O(n) and the capacity for parallel processing, RF is highly efficient for large-scale data. The algorithm ensures diversity among its base learners by employing a dual-randomization approach: it utilizes random bootstrap sampling for the training instances and selects a random subset of predictors for tree construction. This strategy effectively reduces inter-tree correlation, enhancing the stability and overall accuracy of the final predictive model [35]. During the classification phase, the RF algorithm determines the final prediction for a new test sample x^{new} by aggregating the outcomes of all individual trees and selecting the most voted class. This majority voting mechanism is formally expressed in Eq. (17) [35]:

$$C_{rf}^{n_{tree}}(x^{new}) = majorityVote\{C_i(x^{new})\}_1^{n_tree} \quad (17)$$

where $C_{rf}^{n_{tree}}(x^{new})$ represent the final predicted class label by Random Forest, n_tree is the total number of decision trees in the ensemble and $C_i(x^{new})$ denotes the class label predicted by the i-th decision tree [19].

7. XGBoost

eXtreme Gradient Boosting (XGBoost) is a highly optimized ensemble learning technique that serves as an advanced extension of the traditional Gradient Boosting framework. Applicable to both classification and regression tasks, the algorithm constructs a robust predictive model by aggregating multiple weak base learners, typically decision trees. Through a sequential training process, XGBoost iteratively generates new trees specifically designed to model and correct the residuals the discrepancies between the predicted and actual observed values left by the preceding trees, ultimately converging into a highly accurate strong learner. To optimize these sequential predictions, the model is trained in an additive manner by minimizing a specific objective function at each iteration t, which is formulated in Eq. (18) [36]:

$$Obj^{(t)} = \sum_{i=1}^{n} L\left(p_i, \hat{p}_i^{(t-1)} + f_t(x_i)\right) + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \quad (18)$$

where p_i represents the actual target value, $\hat{p}_i^{(t-1)}$ denotes the cumulative predicted value from the previous iteration $t-1$, $f_t(x_i)$ represents the newly added tree, L is a differentiable loss function measuring the estimation error between p_i and the predicted values, γ is a complexity parameter for each leaf, T represents the total number of leaves in the tree, λ is the L2 weight regularization term, and w_j denotes the continuous scores assigned to the j-th leaf [36].

Table 3. Hyperparameter search spaces for the nine machine learning classifiers evaluated in the nested cross-validation framework.

Classification Model	Parameter
Logistic Regression	penalty: l2, None; C: 0.001, 0.01, 0.1, 1, 10, 100; max_iter: 50, 100, 500, 1000
Decision Tree	max_depth: None, 5, 10, 20; min_samples_leaf: 1, 2, 4;
K-Nearest Neighbors	n_neighbors: 1, 3, 5, 7, 9;
Naïve Bayes	var_smoothing: 1e-9, 1e-8, 1e-7;
Support Vector Machine	C: 0.001, 0.01, 0.1, 1, 10
Random Forest	n_estimator: 50, 100; max_depth: None, 10, 20; criterion: gini, entropy, log_loss;
XGBoost	learning_rate: 0.01, 0.1, 0.2, 0.3; max_depth: 3, 5, 7;
AdaBoost	learning_rate: 0.01, 0.1, 1.0; n_estimators: 50, 100;
CatBoost	iterations: 50, 100; depth: 4, 6;

8. AdaBoost

Adaptive Boosting (AdaBoost) is a prominent ensemble learning algorithm primarily utilized for classification tasks. It operates by aggregating multiple weak base classifiers to construct a highly accurate, strong predictive model. The fundamental mechanism of AdaBoost relies on an iterative reweighting process: during each sequential training round, the algorithm dynamically increases the weights of previously misclassified instances. By forcing subsequent base learners to focus heavily on these challenging data points, the algorithm progressively minimizes classification errors and optimizes the final predictive boundary [37].

9. CatBoost

Categorical Boosting (CatBoost) is an advanced Gradient Boosting Decision Tree (GBDT) framework renowned for its native and efficient processing of categorical features. By utilizing symmetric decision trees as its base learners, the model operates with fewer parameters, ensuring computational efficiency. Moreover, CatBoost resolves standard gradient bias and prediction shift issues by

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

implementing an ordered boosting technique, which minimizes the potential for overfitting and yields highly accurate classification results [38].

To ensure each classification model achieves optimal performance and robust generalization capabilities, a series of key hyperparameters were configured. Details regarding the parameter space and the specific values evaluated for each algorithm are summarized in Table 3.

E. Stratified 10-Fold Cross-Validation

Stratified K-Fold Cross-Validation is an advanced variant of the standard cross-validation method that utilizes stratified rather than random sampling. This approach divides the dataset into k mutually exclusive folds of approximately equal size. The primary objective of this method is to ensure that relative class frequencies are effectively maintained and kept constant across every training and validation fold [11]. Consequently, every partition possesses nearly the same percentage of samples from each target class as the complete dataset. During the iterative evaluation process, one nested fold is selected for testing, while the remaining folds are utilized for model training.

Algorithm 4 presents the pseudocode for the Stratified 10-Fold Cross Validation algorithm used in this study. This pseudocode outlines the step-by-step programmatic execution required to achieve the aforementioned proportional allocation. It specifies how the procedure iterates through each target class and randomly draws samples based on their original distribution to construct the robust evaluation partitions [39].

Algorithm 4: Stratified 10-Fold Cross-Validation for generating class-balanced training and validation partitions while preserving the original target class proportions.

Input: Feature matrix X, labels y; k = 10 folds.
Output: Training and validation indices **for each** fold.
Function StratifiedFolds(X, y, k):
  Initialize empty folds F[1], …, F[k]
  C ← unique classes in y
  for each class c in C **do**
    A ← indices of samples with label c
    t ← floor(|A| / k)
    for j = 1, …, k **do**
      S ← random subset of t indices from A
      F[j] ← F[j] ∪ S
      A ← A \ S
    end
    Distribute remaining indices in A across folds
  end
  for j = 1, …, k **do**
    Validation indices ← F[j]
    Training indices ← all indices outside F[j]
  end
  return the k training–validation partitions
End

Algorithm 5: Nested Stratified Cross-Validation for integrating feature selection, hyperparameter optimization, model training, and unbiased performance evaluation using independent inner and outer validation loops.

Input: X, y, classifier, hyperparameter search space; M = 10 outer folds; N = 10 inner folds.
Output: Mean performance across outer test folds.
Function NestedCV(X, y, classifier, search_space):
  Create M stratified outer folds; H ← []
  for each outer fold i **do**
    V_outer ← fold i; T_outer ← remaining folds
    Define pipeline: scaling → feature selection → classifier
    for each configuration θ from randomized search **do**
      L ← []; create N inner folds from T_outer
      for each inner fold j **do**
        V_inner ← fold j; T_inner ← remaining inner folds
        Fit the pipeline with θ on T_inner only
        Evaluate on V_inner; append loss to L
      end
      Record mean(L) **for** θ
    end
    θ_best ← configuration with lowest mean inner loss
    Refit the selected pipeline on all of T_outer
    Evaluate on V_outer; append metrics to H
  end
  return mean outer-fold metrics in H
End

F. Nested Cross-Validation

To further ensure robust model evaluation and prevent overoptimistic performance estimates, a Nested Cross-Validation (Nested CV) architecture was implemented. Nested CV is particularly advantageous for biomedical datasets characterized by a limited sample size but a significantly large number of predictors. The approach partitions the dataset into two distinct loops: inner folds and outer folds. At each outer iteration, the inner cross-validation loop is strictly utilized for tuning optimal model hyperparameters and conducting embedded feature selection, ensuring these critical steps are performed exclusively on the training data of that specific fold. Subsequently, the selected features and the tuned model are evaluated on the unseen hold-out data within the respective outer cross-validation loop. By independently nesting the feature filtering and parameter tuning processes within each training split, this methodology effectively eliminates information leakage and provides an unbiased determination of the overall model accuracy [40].

Furthermore, Algorithm 5 details the pseudocode for the Nested Cross Validation algorithms used in this study.

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

The provided algorithm delineates the exact procedural execution of the two-layer architecture. It systematically translates the theoretical separation of the loops into computational steps, explicitly demonstrating how the inner fold iterations are algorithmically isolated from the final generalization assessment on the outer folds [41].

G. Evaluation Metrics

The evaluation framework employs Accuracy, Precision, Recall, and F1-Score. Assessing model reliability from these multiple perspectives is vital in medical data classification, particularly because the clinical cost of false negative predictions is exceedingly high [42]. To provide a more robust and honest performance assessment, especially given the challenges of imbalanced data, this study also incorporates Specificity, Balanced Accuracy, and the Matthews Correlation Coefficient (MCC). Specificity measures the proportion of actual negatives that are correctly identified. When dealing with skewed datasets, standard Accuracy can be misleading as it tends to favor the majority class; thus, Balanced Accuracy (the arithmetic mean of sensitivity and specificity) is recommended as it treats performance on both classes with equal weight. Furthermore, the MCC is considered a highly reliable metric as it utilizes all four quadrants of the confusion matrix (TP, TN, FP, FN) and yields a high score only if the model performs well on both classes. These metrics were computed using the following formulas Eq.(19), Eq.(20), Eq.(21), Eq.(22), Eq.(23), Eq.(24) and Eq.(25) [42], [43], [44], [45].

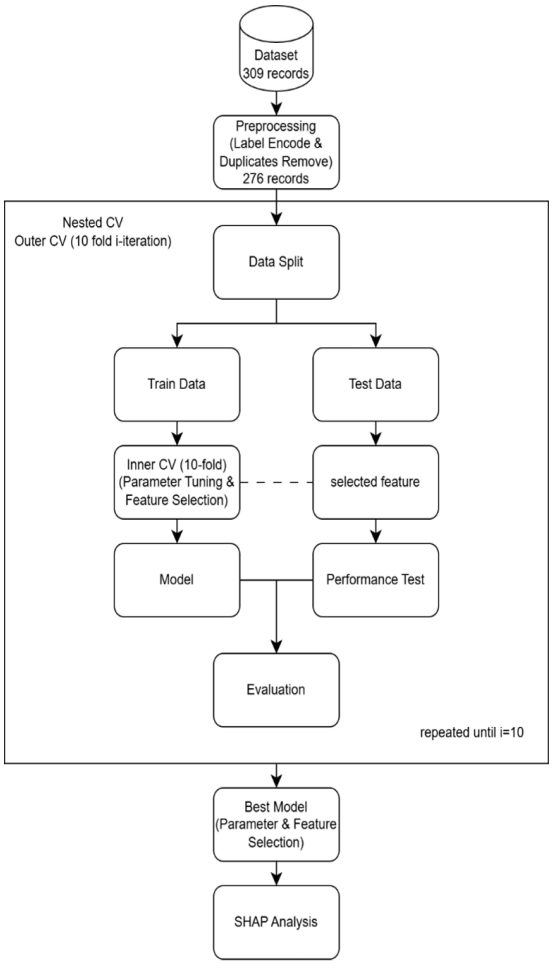


Fig. 1. Proposed machine learning workflow for survey-based lung cancer classification using nested stratified 10-fold cross-validation, incorporating data preprocessing, feature selection, hyperparameter optimization, model evaluation, and SHAP-based interpretation.

1. Accuracy (Eq. (19)):

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad \textbf{(19)}$$

2. Precision (Eq. (20)):

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad \textbf{(20)}$$

3. Sensitivity (Eq. (21)):

$$\boldsymbol{Sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad \textbf{(21)}$$

4. Specificity (Eq. (22)):

$$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN+FP}} \quad \textbf{(22)}$$

5. F1-Score (Eq. (23)):

$$\boldsymbol{F1} = \frac{\boldsymbol{2 \times Precision \times Sensitivity}}{\boldsymbol{Precision+Sensitivity}} \quad \textbf{(23)}$$

6. Balanced Accuracy (Eq. (24)):

$$\boldsymbol{Balanced\ Acc} = \frac{\boldsymbol{Sensitivity+Specificity}}{\boldsymbol{2}} \quad \textbf{(24)}$$

7. Matthews correlation coefficient (MCC) (Eq. (25)) [44]:

$$\boldsymbol{MCC} = \frac{\boldsymbol{TP\,.\,TN-FP\,.\,FN}}{\sqrt{\boldsymbol{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}}} \quad \textbf{(25)}$$

True Positives (TP) and True Negatives (TN) denote the accurate classification of positive and negative instances, respectively. In contrast, False Positives (FP) refer to negative instances that the model incorrectly labels as positive, while False Negatives (FN) represent positive instances that are erroneously classified as negative [46]. Fig. 1 summarizes the proposed workflow.

In addition to threshold-based metrics, model performance is evaluated using ROC AUC and PR AUC. The ROC AUC measures the integral of the True Positive Rate against the False Positive Rate across all decision thresholds and is valued for its invariance to class imbalance. Conversely, the PR AUC plots Precision against Recall; because it is highly sensitive to the class distribution, it serves as a more informative metric when the research objective is to prioritize performance on a rare minority class. Utilizing a combination of these metrics is essential to avoid potential interpretative biases that may arise from relying on a single performance measure [47].

To quantify the stability and variation of these performance metrics across the cross-validation folds, the Standard Deviation (SD) is calculated. The SD measures the dispersion of the evaluation scores from their mean, providing a clear indicator of the model's consistency

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

when reported as Mean ± SD. Mathematically, the SD for a specific metric across K folds is defined in Eq.(26) [48]:

$$\sigma_i = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(y_j - \bar{y}_i)^2} \quad (26)$$

where y_j is the value of the j-th point, $\bar{y}_i$ is the mean value of the attribute i, and m is the total number of points [48].

Furthermore, to rigorously evaluate the reliability of these estimates, 95% Confidence Intervals (CIs) are determined. A CI quantifies the uncertainty around the mean estimate due to sampling error, acting as a crucial bridge between statistical significance and clinical importance. For instances where the sample size is relatively small, and the exact population standard deviation remains unknown, the 95% CI is computed utilizing the critical value from the Student's t-distribution with $m-1$ degrees of freedom, as expressed in Eq. (27) [49]:

$$95\%CI = \bar{y}_i \pm t_{(0.025,m-1)} \times \left(\frac{\sigma_i}{\sqrt{m}}\right) \quad (27)$$

where $t_{(0.025,m-1)}$ represents the critical t-value and $\left(\frac{\sigma_i}{\sqrt{m}}\right)$ denotes the standard error of the mean. A narrower SD and CI reflect higher precision, ensuring that the model's predictive capability is consistently robust [49].

III. Results

This section presents a comprehensive comparative analysis of various feature selection techniques applied to the lung cancer prediction dataset. To ensure robust performance estimation and prevent data leakage, the entire evaluation is conducted within a Nested CV framework. In this approach, feature selection is performed independently within each inner loop, allowing for dynamic subset optimization based on the training data of each specific fold. The analysis initiates with a baseline model utilizing the full feature set, followed by a systematic comparison against models optimized through multiple feature selection algorithms, each rigorously evaluated across all outer folds. Based on Fig. 1, the proposed methodology employs a rigorous Nested Cross-Validation architecture to guarantee an unbiased evaluation of the machine learning models. Initially, the dataset undergoes preprocessing to eliminate duplicated records, ensuring a clean and unique set of observations. The data is then systematically partitioned into training and testing sets within the Outer CV loop. To strictly prevent any risk of data leakage, the feature scaling and feature selection processes are embedded directly into the Inner CV loop using a Pipeline approach. This ultra-strict configuration ensures that the hold-out test data remains completely isolated and unseen during the hyperparameter tuning phase.

Within the Inner CV pipeline, the study evaluates multiple feature selection techniques in parallel to identify optimal predictive variables. The implemented methods consist of filter-based approaches, namely ANOVA combined with Chi-Square and Mutual Information, alongside wrapper-based techniques including Recursive Feature Elimination, Forward Selection, and Backward Elimination. To assess subset effectiveness, the predictive performance is evaluated on the Outer CV test sets using nine metrics: Accuracy, Precision, Recall, F1-Score, Balanced Accuracy, Specificity, MCC, ROC AUC, and PR AUC, reported as the mean with standard deviation.

Table 4 presents the evaluation results of models utilizing the nested CV and incorporating feature selection conducted within each inner loop. Once the best-performing model and feature subset combination is identified, the winning selection method is applied to the entire dataset to lock a uniform feature space. Ultimately, the model's decision-making process is interpreted using an Out-of-Fold SHAP analysis conducted via an independent Stratified 10-Fold cross-validation. This approach guarantees that SHAP values are extracted strictly from unbiased held-out predictions while maintaining dimensional stability, ensuring a transparent understanding of global feature contributions.

Table 4. Best classification accuracy achieved by each feature selection technique and its corresponding machine learning classifier under nested stratified cross-validation.

Feature Selection	Classification Model	Accuracy
Anova + Chi-Square	AdaBoost	90.97%
FS	Naïve Bayes	89.85%
BE	AdaBoost	90.57%
MIG	Random Forest	90.57%
RFE	KNN	91.28%

A brief overview of the top-performing models across the evaluated metrics reveals distinct strengths among the algorithms. For Accuracy, K-Nearest Neighbors (KNN) paired with Recursive Feature Elimination (RFE) achieves the highest overall score at 91.28% ± 4.91% (95% CI: 87.58% - 94.99%) compared to the baseline KNN with 88.37% ± 5.13%, the RFE integration provided an absolute accuracy increase of 2.91%. Followed closely by the baseline AdaBoost (ADAB) model without feature selection at 90.98% ± 5.13%. Precision is heavily dominated by Logistic Regression (LR) across almost all configurations, peaking at a remarkable 98.25% ± 2.85% when combined with Backward Elimination (BE). Conversely, ensemble models excel in Sensitivity, demonstrating superior capability in correctly identifying positive cases, with CatBoost (CATB) using Feature Selection (FS) reaching 97.50% ± 2.04% and the baseline ADAB scoring 97.05% ± 2.69%. The best F1-Score, representing a strong balance between precision and sensitivity, is achieved by KNN with RFE at 95.02% ± 2.81% and the baseline ADAB at 94.90%. Furthermore, LR stands out as the most robust model for handling class imbalances and detecting negative cases, leading the results in Balanced Accuracy at 89.95% ± 8.42% (95% CI: 83.60% - 96.29%), Specificity at 90.00% ± 16.58%, and the Matthews Correlation Coefficient or MCC at 67.85% ±

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

13.42%all of which are achieved under the BE

Table 5. Highest balanced accuracy achieved by each feature selection technique and its corresponding classifier for imbalanced lung cancer classification.

Feature Selection	Classification Model	Balanced Accuracy
Anova + Chi-Square	Logistic Regression	87.64%
FS	Logistic Regression	86.40%
BE	Logistic Regression	89.95%
MIG	Logistic Regression	82.44%
RFE	Logistic Regression	85.37%

configuration which give absolute improvement from its baseline of 3.14%, 5.00% and 5.95%, respectively. Regarding area under the curve metrics, LR with RFE displays the strongest ROC AUC for class separation at 94.04% ± 5.48%, marking a 1.14% absolute increase from its baseline, while also recording the highest PR AUC at 99.00% ± 0.96%, with 0.33% improvement from the baseline. Therefore, by securing the highest scores in four critical metricsPrecision, Balanced Accuracy, Specificity, and MCCLogistic Regression combined with Backward Elimination (LR + BE) solidifies its position as the best overall model in this study. To explicitly quantify the changes brought about by feature selection across the evaluated metrics, an independent analysis of the cross-validation results demonstrates both significant improvements and notable degradations. In terms of Accuracy, applying the ANOVA + Chi-Square method improved the K-Nearest Neighbors (KNN) model from its baseline of 88.37% ± 5.13% to 89.99% ± 5.47%, whereas the Mutual Information Gain (MIG) filter degraded the Decision Tree (DT) accuracy from 88.06% ± 5.76% down to 85.16% ± 6.66%. Similar contrasting numerical comparisons reflect this variance across class balance evaluations. For Balanced Accuracy, ANOVA + Chi-Square successfully raised the DT baseline from 76.01% ± 15.14% to 82.44% ± 13.36%, while Forward Selection (FS) severely reduced the CatBoost (CATB) score from 67.26% ± 10.56% to 59.58% ± 7.86%. Regarding Specificity, Backward Elimination (BE) effectively enhanced the Logistic Regression (LR) model from 85.00% ± 16.58% to 90.00% ± 16.58%; conversely, FS caused a drastic drop in CATB specificity from 37.50% ± 21.49% to a mere 21.67% ± 15.90%. The Matthews Correlation Coefficient (MCC) also highlights this dynamic: BE boosted the XGB baseline from 34.42% ± 33.50% to 45.86% ± 18.49%, whereas FS triggered a sharp decline in CATB from 43.05% ± 22.18% to 26.20% ± 22.07%. Finally, evaluating the area under the curve metrics, ANOVA + Chi-Square strengthened the Random

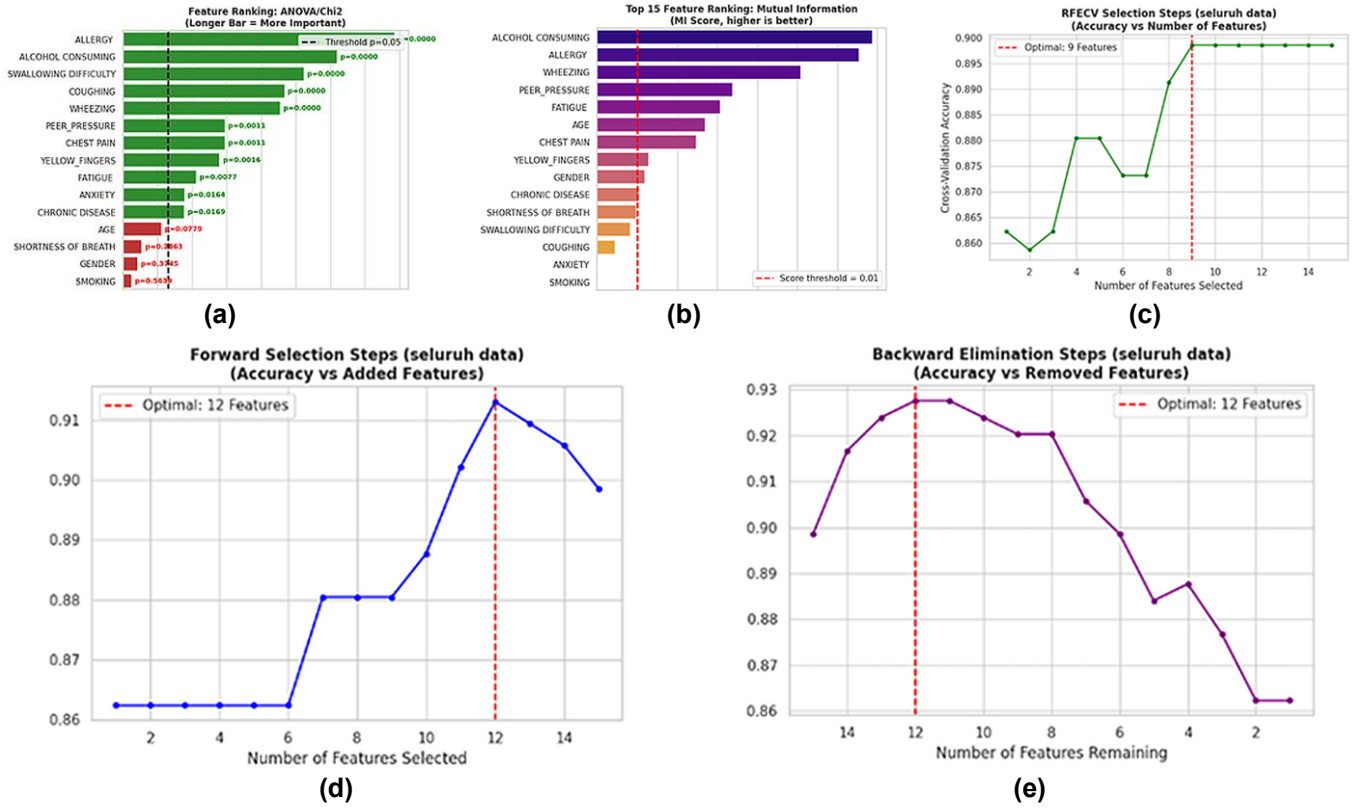


Fig. 2. Comparative analysis of feature selection techniques using Logistic Regression: (a) ANOVA and Chi-Square scores, (b) Mutual Information gain, (c) RFECV iteration scores, (d) Forward Selection iteration scores, (e) Backward Elimination iteration scores.

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Forest (RF) ROC AUC from 91.93% ± 6.36% to 93.67% ± 6.30% and its PR AUC from 98.48% ± 1.35% to 98.97% ± 1.11%. In contrast, the MIG method consistently degraded the classification probabilities of the DT model, dropping its ROC AUC from 76.01% ± 15.14% to 71.06% ± 16.22% and its PR AUC from 92.97% ± 4.40% to 91.77% ± 4.69%

Given that Logistic Regression emerged as the most robust and optimal classifier in this study, it is utilized as the primary base estimator to further investigate how each feature selection method operates in Fig. 2. The univariate filter charts (ANOVA/Chi-Square and Mutual Information) plot the exact significance values for each clinical predictor. Based on the predefined thresholds ($p \leq 0.05$ for ANOVA and MI > 0.01 for MI Gain), features such as 'ALLERGY', 'ALCOHOL CONSUMING', and 'WHEEZING' were retained, while variables including 'SMOKING' and 'SHORTNESS OF BREATH' were excluded. Concurrently, the sequential plots for the wrapper and recursive methods track the cross-validation accuracy across different subset sizes. The RFE curve identifies an optimal peak at a condensed subset of exactly 9 features, whereas the Forward Selection (FS) and Backward Elimination (BE) trajectories achieve their highest validation accuracy by retaining a subset of 12 features. Table 6 reports the best-performing models and their 95% confidence intervals.

As summarized in Table 4, which presents the optimal results achieved by each feature selection technique, the K-Nearest Neighbors (KNN) classifier paired with Recursive Feature Elimination (RFE) delivered the best general performance, yielding the highest accuracy (91.28%) and F1-score (95.02%) across all tested scenarios. Additionally, Logistic Regression combined with Backward Elimination (BE) achieved the highest precision (98.25%). Meanwhile, the highest recall metric (97.50%) was achieved by the CatBoost classifier paired with the FS method.

Furthermore, Table 5 highlights the evaluation of balanced accuracy to account for potential class imbalances within the medical dataset. For this specific metric, Logistic Regression (LR) consistently dominated across all feature selection techniques, with its combination alongside Backward Elimination (BE) achieving the absolute highest score of 89.95%, demonstrating superior stability in classifying both positive and negative cases.

Although intended to improve performance, certain feature selection methods paradoxically degraded specific classifiers in this study. For example, Mutual Information Gain (MIG) reduced Logistic Regression's accuracy from 88.03% to 84.78%, while Forward Selection (FS) lowered AdaBoost's accuracy from 90.98%

Table 6. Classification performance and 95% confidence intervals for the best-performing models across feature selection techniques, including accuracy, precision, sensitivity, F1-score, balanced accuracy, specificity, Matthews correlation coefficient, ROC AUC, and PR AUC.

Metric	ADAB (Anova+Chi2)	NB (FS)	ADAB (BE)	RF (MIG)	KNN (RFE)	LR (BE)
Acc	90.97 (88.48-93.45)	89.85 (85.66-94.05)	90.57 (85.16-95.97)	90.57 (88.36-92.78)	91.28 (87.58-94.99)	89.84 (86.63-93.06)
Prec	93.25 (90.69-95.80)	93.77 (91.67-95.88)	93.27 (89.12-97.42)	93.29 (90.45-96.13)	93.91 (91.16-96.68)	98.25 (96.10-100)
Sens	96.63 (94.74-98.53)	94.49 (90.43-98.55)	96.18 (93.12-99.24)	96.20 (93.95-98.44)	96.21 (94.01-98.42)	89.89 (86.61-93.17)
F1	94.85 (93.44-96.27)	94.06 (91.47-96.65)	94.63 (91.53-97.72)	94.63 (93.39-95.86)	95.02 (92.90-97.13)	93.81 (91.77-95.85)
BAcc	76.65 (68.64-84.66)	77.66 (71.13-84.19)	76.42 (62.59-90.25)	76.01 (67.55-84.48)	78.52 (69.57-87.48)	89.95 (83.60-96.29)
Spec	56.67 (40.09-73.24)	60.83 (48.89-72.77)	56.67 (30.19-83.15)	55.83 (37.61-74.06)	60.83 (43.96-77.71)	90.00 (77.50-100)
MCC	59.50 (47.41-71.59)	59.25 (45.97-72.53)	55.96 (29.44-82.47)	57.40 (44.98-69.82)	60.80 (43.62-77.98)	67.85 (57.73-77.97)
ROC	92.48 (87.16-97.80)	88.63 (81.99-95.27)	91.80 (85.08-98.53)	90.44 (84.26-96.63)	90.96 (84.33-97.59)	93.09 (87.80-98.37)
PC	98.64 (97.51-99.76)	97.79 (96.13-99.46)	98.57 (97.33-99.81)	98.16 (96.73-99.59)	97.64 (95.64-99.64)	98.80 (97.77-99.82)

IV. Discussion

The primary objective of this study was to determine the most effective feature selection and classification methodology for a lung cancer dataset while strictly eliminating methodological bias and information leakage. Based on the evaluation of nine performance metrics, specific algorithmic strengths emerged depending on the classification objective.

to 88.07%. This decline can be attributed to three main factors. First, univariate filters like MIG may inadvertently discard variables containing complex, non-linear clinical interactions essential in medical datasets. Second, tree-based ensemble models already perform internal feature selection during node splitting; forcibly restricting their feature space beforehand limits their capacity to optimize data splits. Lastly, wrapper methods like FS can introduce optimization bias, where selected features overfit the

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

training folds and fail to generalize during nested validation. Ultimately, feature selection is not universally beneficial and must be carefully paired with the algorithm to avoid stripping essential structural complexity.

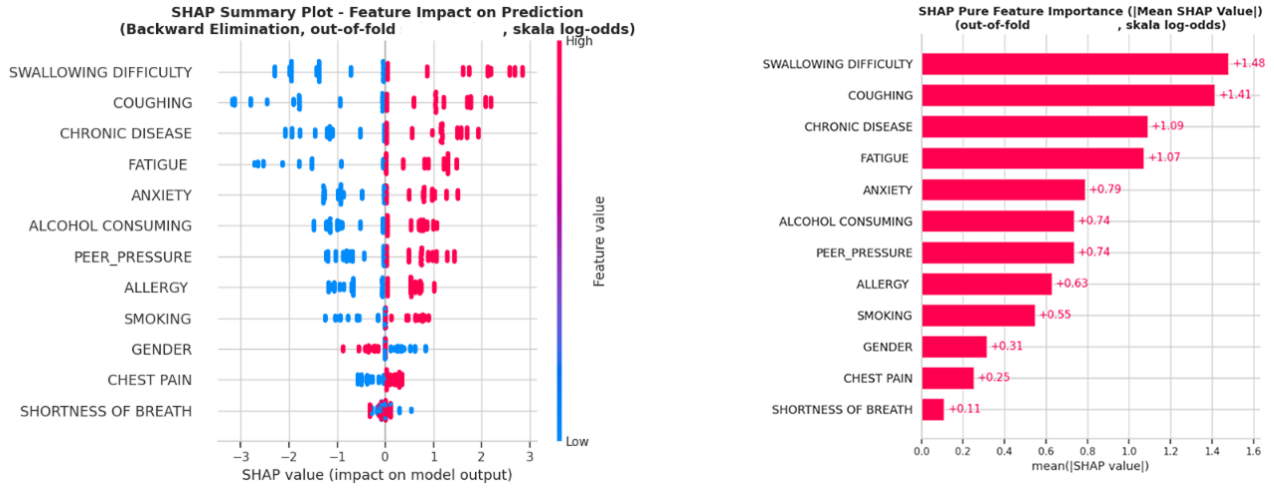
The superior performance of Logistic Regression (LR) combined with Backward Elimination (BE) is rooted in the synergy between the model's linear decision boundary and the iterative pruning strategy. Since LR is sensitive to redundant predictors, BE enhances robustness by initiating from the full feature space to capture complex interactions before systematically discarding non-informative variables. This iterative refinement allows the model to retain sufficient structural complexity while effectively pruning noise. As detailed in Table 6, which provides a comprehensive comparison of the best-performing models alongside their 95% confidence intervals (CI), this precise tuning enabled LR + BE to achieve superior specificity and balanced accuracy. These metrics confirm the model's enhanced ability to differentiate clinical manifestations and minimize false positives, establishing it as the most reliable approach for the classification task at hand.

Table 7 compares the proposed methodology with prior studies using the same survey-based lung cancer dataset. As observed in the literature, approaches vary significantly; some relied on standalone classification models, while others attempted optimization strategies. For instance, studies employing AdaBoost [1], Random Forest [9], and an Artificial Neural Network (ANN) [50] achieved accuracies of 90.70%, 90.32%, and 92.23%.

Potential methodological vulnerabilities in survey-based lung cancer classification include information leakage during resampling and feature selection. Applying SMOTE to the entire dataset before partitioning can allow information from validation observations to influence training. Similarly, selecting features before cross-validation can bias performance estimates. These risks should be assessed from the complete validation protocol of each study rather than inferred from its reported accuracy. Studies [1], [9], [51], [52], and [53] provide relevant comparisons, alongside feature-selection studies [8] and [51]. Study [50] reported an ANN accuracy of

Table 7. Comparison of the proposed machine learning and feature selection framework with previous lung cancer classification studies, including validation strategies, classification algorithms, and reported predictive performance.

Study / model	Evaluation and reported result
[1] Mamun et al. (2022) ensemble learning	10-fold CV; accuracy 90.70%
[1] Mamun et al. (2022) ensemble learning	10-fold CV; accuracy 89.76%
[9] Random Forest	Accuracy 90.32%; validation design not stated in manuscript
[8] Naïve Bayes	5-fold CV; accuracy 82.81%
[10] Vieira et al. (2021)	10-fold CV; accuracy 90.9% / 88.7% / 87.4% / 85.5%
[50] ANN	Accuracy 92.23%; evaluation design as reported in [50]
Proposed: KNN + RFE	Nested stratified 10-fold CV; accuracy 91.28% ± 4.91%
Proposed: LR + BE	Nested stratified 10-fold CV; balanced accuracy 89.95% ± 8.42%



(a) (b)

Fig. 3. SHAP-based interpretation of the Logistic Regression model optimized using Backward Elimination: (a) SHAP summary plot illustrating feature contributions to individual predictions and (b) global feature importance ranked by mean absolute SHAP values on the log-odds scale.

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

92.23%; this result should be interpreted in the context of its evaluation design.

Furthermore, many of these prior studies rely solely on standard accuracy and lack comprehensive reporting of standard deviations, making it difficult to evaluate the variance and stability of their models across different data folds. In contrast, the proposed methodology mitigates these vulnerabilities by employing a Nested Stratified 10-Fold Cross-Validation approach. By strictly encapsulating the feature selection and hyperparameter tuning processes within the inner loops, this study preserves the absolute integrity of the validation phase. Therefore, while the peak accuracy of 91.28% achieved by the KNN + RFE model, or the balanced accuracy of 89.95% achieved by the LR + BE model, may not dramatically exceed the reported results of previous studies, they represent a significantly more realistic, reliable, and unbiased estimation of the models' true generalization capabilities on unseen data.

To improve the interpretability of the final LR + BE model, SHAP analysis was performed using the LinearExplainer on held-out predictions obtained during stratified 10-fold cross-validation. Since the Logistic Regression (LR) model is inherently linear in the log-odds space, the LinearExplainer was utilized to derive SHAP values by directly leveraging the model's coefficients, ensuring an exact and efficient attribution of feature importance. The training dataset was employed as the background dataset to establish a reference baseline. All SHAP values were computed on the model output expressed in the log-odds scale, allowing for a precise interpretation of how each feature contributes to the model's final prediction.

As illustrated in the SHAP feature importance (Fig. 3b), Swallowing Difficulty and Coughing exhibited the largest mean absolute SHAP values (+1.48 and +1.41 on the log-odds scale, respectively), followed by Chronic Disease (+1.09) and Fatigue (+1.07), indicating that these variables contributed most strongly to the model's predictions. In the SHAP summary plot (Fig. 3a), given that the features are binary, the plot displays two distinct clusters representing the presence (red) and absence (blue) of each symptom. The horizontal dispersion of points within each cluster reflects instance-based variation in feature importance; although the input features are binary, their calculated SHAP values vary across individual samples due to the specific combination of features present in each record, demonstrating that the predictive weight of a symptom is not absolute but is modulated by the context of other indicators within the model. For the most influential features, higher feature values were generally associated with positive SHAP values, whereas lower feature values tended to produce negative SHAP contributions, reflecting their influence on the model's output. Conversely, predictors with SHAP values clustered around zero, such as Shortness of Breath, indicate a neutral impact on the model's output; for these instances, the presence or absence of the feature did not significantly deviate the prediction from the model's baseline, suggesting limited discriminatory power. In contrast, Gender, Chest Pain, and Shortness of Breath exhibited comparatively smaller SHAP value ranges (+0.31, +0.25, and +0.11, respectively) and greater overlap between high and low feature values, suggesting weaker and less consistent contributions to the model's predictions. These findings indicate that the retained predictors differ in their relative influence on the LR + BE model and provide an explanation of how the model generates its predictions.

Despite the rigorous validation framework, this study acknowledges several critical limitations that must be addressed. First, the analysis relies entirely on a survey-based dataset, which is inherently dependent on self-reported patient symptoms that may be subjective. Second, there is a complete absence of external validation; the models were not tested on an independent, geographically or demographically distinct cohort, limiting the generalizability of the findings. Third, the dataset is characterized by a very small negative class, leading to severe class imbalance. Although stratified cross-validation was employed to maintain proportional distributions, this scarcity raises the possibility of sampling bias, where the available negative cases may not adequately represent the broader healthy population. Furthermore, the binary encoding of symptoms (e.g., simple presence or absence) oversimplifies the pathology by failing to capture symptom severity, frequency, or duration. The total lack of clinical imaging (such as CT scans or MRIs) and objective laboratory variables means the models are trained purely on superficial indicators. Due to these significant constraints, this study makes no claims regarding the diagnostic readiness of the models; they remain strictly experimental and are not prepared for deployment in actual clinical diagnostic settings.

V. Conclusion

This study aimed to determine the most effective machine learning classification and feature selection methods for predicting lung cancer while eliminating information leakage and methodological bias. Methodologically, the implementation of a Nested Stratified 10-Fold Cross-Validation architecture mitigated the risk of information leakage and provided a less biased performance estimation by strictly confining the feature selection and parameter optimization processes within the inner training loops. Based on the comparative analysis, the K-Nearest Neighbors (KNN) algorithm combined with Recursive Feature Elimination (RFE) achieved the highest overall performance, yielding an accuracy of 91.28% ± 4.91% and an F1-score of 95.02% ± 2.81%. However, to specifically address the severe class imbalance within the dataset, Logistic Regression (LR) optimized with Backward Elimination (BE) provided the highest level of stability, evidenced by achieving the highest balanced accuracy (89.95% ± 8.42%) and precision (98.25% ± 2.85%). The feature selection evaluation demonstrated that wrapper methods were more effective than univariate filters in retaining variables with synergistic predictive interactions. Furthermore, the integration of the SHAP

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

(SHapley Additive exPlanations) method transparently mapped the direction and magnitude of each feature's impact on a log-odds scale, significantly enhancing the model's interpretability. To further ascertain the diagnostic utility of this framework, future work must evaluate this approach through multicenter external validation using larger, balanced cohorts. Furthermore, future studies should explicitly incorporate calibration analysis, decision curve analysis, prospective clinical testing, and direct comparisons with clinically established screening approaches.

Acknowledgment

The authors would like to thank all individuals and parties who contributed to the development of this study. The authors also acknowledge the contributions of the participants whose survey responses formed the basis of the dataset analyzed in this research. The authors acknowledge Mysar Ahmad Bhat for making the Lung Cancer dataset available on Kaggle.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Ethical Approval

This study was a secondary analysis of a publicly available dataset. No new participants were recruited or contacted, and the downloaded dataset contained no direct personal identifiers. The publicly available dataset documentation does not report the ethics approval or exemption status of the original survey. Institutional ethical approval was not required for this secondary analysis, as confirmed by the authors in accordance with their institutional requirements.

Author Contributions

Irwan Budiman contributed to the conceptualization and supervision of the study. Nanda Putri Rahmawati, Annisa Rizqiana, and Rafi'e contributed to the research methodology, data analysis, interpretation of the results, and manuscript preparation. All authors reviewed and approved the final manuscript and agreed to be accountable for all aspects of the work.

Data Availability

The dataset analyzed in this study is publicly available on Kaggle as Lung Cancer at https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer (accessed 8 November 2026). The downloaded file contained 309 records and 16 variables. The dataset was uploaded by mysar ahmad bhat; the Kaggle release is version 1 and is listed under CC0: Public Domain. The analysis code and processed dataset are not provided with this article.

Consent To Participate

No new informed consent was obtained for this secondary analysis because no participants were recruited or contacted. The consent arrangements for the original survey are not reported in the publicly available dataset documentation.

Consent for Publication

Not applicable. This study does not present individually identifiable participant information or images.

Competing Interests

The authors declare that they have no financial or non-financial competing interests that could have influenced the conduct, analysis, interpretation, or reporting of this study.

Declaration of Generative Artificial Intelligence and AI-Assisted Technologies

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) to assist with language refinement, grammatical corrections, and improvements in clarity and readability. The AI tool was not used to generate experimental data, fabricate references, or independently determine the scientific findings. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the accuracy, originality, and integrity of the manuscript.

Research Integrity and Originality

The authors declare that this manuscript reports original research, has not been published previously, and is not under consideration for publication elsewhere. All authors approved the manuscript and take responsibility for the integrity of the analysis and reported findings.

References

[1] M. Mamun, A. Farjana, M. Al Mamun, and M. S. Ahammed, "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis," in *2022 IEEE World AI IoT Congress, AIIoT 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 187–193. doi: https://doi.org/10.1109/AIIoT54504.2022.9817326.

[2] C. Li *et al.*, "Global burden and trends of lung cancer incidence and mortality," *Chin. Med. J. (Engl).*, vol. 136, no. 13, pp. 1583–1590, Jul. 2023, doi: https://doi.org/10.1097/cm9.0000000000002529.

[3] L. A. G. R. Ries, J. L. Young, G. E. Keel, M. P. Eisner, Y. D. Lin, and M.-J. D. Horner, "Cancer Survival Among Adults: U.S. SEER Program, 1988-2001 (low res)," Bethesda, 2007. [Online]. Available: http://www.seer.cancer.gov

[4] A. Ruano-Raviña *et al.*, "Lung cancer symptoms at diagnosis: Results of a nationwide registry study,"

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

ESMO Open, vol. 5, no. 6, pp. 1–7, Nov. 2020, doi: https://doi.org/10.1136/esmoopen-2020-001021.

[5] J. Kim, H. Lee, and B. W. Huang, "Lung Cancer: Diagnosis, Treatment Principles, and Screening," *Am. Fam. Physician*, vol. 105, no. 5, pp. 487–494, Apr. 2022, [Online]. Available: https://www.aafp.org/afp/2022/0500/p487

[6] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015, doi: https://doi.org/10.1016/j.csbj.2014.11.005.

[7] C. Zhang and Y. Ma, *Ensemble machine learning: Methods and applications*, 1st ed., vol. 8. New York: Springer New York, 2012. doi: https://doi.org/10.1007/978-1-4419-9326-7.

[8] S. Pechprasarn, N. Suechoey, N. Pholtrakoolwong, P. Tanedvorapinyo, and Y. Toboonliang, "Optimizing Lung Cancer Diagnosis with Machine Learning and Feature Selection Methods," *J. Curr. Sci. Technol.*, vol. 14, no. 3, p. 55, Sep. 2024, doi: https://doi.org/10.59796/jcst.V14N3.2024.55.

[9] K. Mohan and B. Thayyil, "Machine Learning Techniques for Lung Cancer Risk Prediction using Text Dataset," *International Journal of Data Informatics and Intelligent Computing*, vol. 2, no. 3, pp. 47–56, Sep. 2023, doi: https://doi.org/10.59461/ijdiic.v2i3.73.

[10] E. Vieira, D. Ferreira, C. Neto, A. Abelha, and J. Machado, "Data Mining Approach to Classify Cases of Lung Cancer," in *Trends and Applications in Information Systems and Technologies*, Á. Rocha, H. Adeli, G. Dzemyda, F. Moreira, and A. M. Ramalho Correia, Eds., Cham: Springer International Publishing, 2021, pp. 511–521. doi: https://doi.org/10.1007/978-3-030-72657-7_49.

[11] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Frontiers in Nanotechnology*, vol. 4, p. 972421, Aug. 2022, doi: https://doi.org/10.3389/fnano.2022.972421.

[12] Z. Li, "Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost," *Comput. Environ. Urban Syst.*, vol. 96, p. 101845, Sep. 2022, doi: https://doi.org/10.1016/j.compenvurbsys.2022.101845.

[13] A. M. Ali *et al.*, "Explainable Machine Learning Approach for Hepatitis C Diagnosis Using SFS Feature Selection," *Machines*, vol. 11, no. 3, p. 391, Mar. 2023, doi: https://doi.org/10.3390/machines11030391.

[14] J. Lv and L. Wang, "Hybrid modeling of adsorption process using mass transfer and machine learning techniques for concentration prediction," *Journal of Saudi Chemical Society*, vol. 29, no. 4, p. 12, Jul. 2025, doi: https://doi.org/10.1007/s44442-025-00016-y.

[15] M. Shantal, Z. Othman, and A. A. Bakar, "A Novel Approach for Data Feature Weighting Using Correlation Coefficients and Min–Max Normalization," *Symmetry (Basel).*, vol. 15, no. 12, p. 18, Dec. 2023, doi: https://doi.org/10.3390/sym15122185.

[16] H. Nasiri and S. A. Alavi, "A Novel Framework Based on Deep Learning and ANOVA Feature Selection Method for Diagnosis of COVID-19 Cases from Chest X-Ray Images," *Comput. Intell. Neurosci.*, vol. 2022, p. 4694567, 2022, doi: https://doi.org/10.1155/2022/4694567.

[17] H. Mamdouh Farghaly and T. Abd El-Hafeez, "A high-quality feature selection method based on frequent and correlated items for text classification," *Soft comput.*, vol. 27, no. 16, pp. 11259–11274, Aug. 2023, doi: https://doi.org/10.1007/s00500-023-08587-x.

[18] H. Nayl, E. S. Aly, A. Rezk, and M. E. Fares, "Improved chi-square feature selection for robust heart disease data classification," *AIMS Mathematics*, vol. 11, no. 1, pp. 2682–2701, Jan. 2026, doi: https://doi.org/10.3934/math.2026108.

[19] S. Williamson, K. Vijayakumar, and V. J. Kadam, "Predicting breast cancer biopsy outcomes from BI-RADS findings using random forests with chi-square and MI features," *Multimed. Tools Appl.*, vol. 81, no. 26, pp. 36869–36889, Nov. 2022, doi: https://doi.org/10.1007/s11042-021-11114-5.

[20] M. J. Siraj, T. Ahmad, and R. M. Ijtihadie, "Analyzing ANOVA F-test and Sequential Feature Selection for Intrusion Detection Systems," *International Journal of Advances in Soft Computing and its Applications*, vol. 14, no. 2, pp. 185–194, 2022, doi: https://doi.org/10.15849/IJASCA.220720.13.

[21] R. Lamba, T. Gulati, H. F. Alharbi, and A. Jain, "A hybrid system for Parkinson's disease diagnosis using machine learning techniques," *Int. J. Speech Technol.*, vol. 25, no. 3, pp. 583–593, Sep. 2022, doi: https://doi.org/10.1007/s10772-021-09837-9.

[22] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *J. Reliab. Intell. Environ.*, vol. 7, no. 3, pp. 263–275, Sep. 2021, doi: https://doi.org/10.1007/s40860-021-00133-6.

[23] P. Theerthagiri, "Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique," *Intelligent Systems with Applications*, vol. 16, p. 200121, Nov. 2022, doi: https://doi.org/10.1016/j.iswa.2022.200121.

[24] E. Mosca, F. Szigeti, S. Tragianni, D. Gallagher, and G. Groh, "SHAP-Based Explanation Methods: A Review for NLP Interpretability," in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K.

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Lee, E. Santus, F. Bond, and S.-H. Na, Eds., Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 4593–4603. Accessed: Jul. 10, 2026. [Online]. Available: https://aclanthology.org/2022.coling-1.406/

[25] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: https://doi.org/10.1038/s42256-019-0138-9.

[26] F. H. Awad, M. M. Hamad, and L. Alzubaidi, "Robust Classification and Detection of Big Medical Data Using Advanced Parallel K-Means Clustering, YOLOv4, and Logistic Regression," *Life 2023*, vol. 13, no. 3, p. 34, Mar. 2023, doi: https://doi.org/10.3390/life13030691.

[27] O. Karasoy and S. Ballı, "Spam SMS Detection for Turkish Language with Deep Text Analysis and Deep Learning Methods," *Arab. J. Sci. Eng.*, vol. 47, no. 8, pp. 9361–9377, Aug. 2022, doi: https://doi.org/10.1007/s13369-021-06187-1.

[28] D. Dey *et al.*, "The proper application of logistic regression model in complex survey data: a systematic review," *BMC Med. Res. Methodol.*, vol. 25, no. 1, p. 18, Dec. 2025, doi: https://doi.org/10.1186/s12874-024-02454-5.

[29] R. A. Khan *et al.*, "A novel framework for classification of two-class motor imagery EEG signals using logistic regression classification algorithm," *PLoS One*, vol. 18, no. 9 September, p. e0276133, Sep. 2023, doi: https://doi.org/10.1371/journal.pone.0276133.

[30] R. Guo, D. Fu, and G. Sollazzo, "An ensemble learning model for asphalt pavement performance prediction based on gradient boosting decision tree," *International Journal of Pavement Engineering*, vol. 23, no. 10, pp. 3633–3646, 2022, doi: https://doi.org/10.1080/10298436.2021.1910825.

[31] K. Gajowniczek and M. Dudziński, "Influence of Explanatory Variable Distributions on the Behavior of the Impurity Measures Used in Classification Tree Learning," *Entropy*, vol. 26, no. 12, p. 35, Dec. 2024, doi: https://doi.org/10.3390/e26121020.

[32] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Sci. Rep.*, vol. 12, no. 1, p. 6256, Dec. 2022, doi: https://doi.org/10.1038/s41598-022-10358-x.

[33] S. Ballı and E. Özdemir, "A novel method for prediction of EuroLeague game results using hybrid feature extraction and machine learning techniques," *Chaos Solitons Fractals*, vol. 150, p. 111119, Sep. 2021, doi: https://doi.org/10.1016/j.chaos.2021.111119.

[34] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," *Information (Switzerland)*, vol. 15, no. 4, p. 235, Apr. 2024, doi: https://doi.org/10.3390/info15040235.

[35] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Dec. 2024, doi: https://doi.org/10.58496/BJML/2024/007.

[36] M. Niazkar *et al.*, "Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023)," *Environmental Modelling & Software*, vol. 174, p. 105971, 2024, doi: https://doi.org/10.1016/j.envsoft.2024.105971.

[37] S. Gamil, F. Zeng, M. Alrifaey, M. Asim, and N. Ahmad, "An Efficient AdaBoost Algorithm for Enhancing Skin Cancer Detection and Classification," *Algorithms*, vol. 17, no. 8, p. 353, Aug. 2024, doi: https://doi.org/10.3390/a17080353.

[38] W. Chang, X. Wang, J. Yang, and T. Qin, "An Improved CatBoost-Based Classification Model for Ecological Suitability of Blueberries," *Sensors*, vol. 23, no. 4, p. 1811, Feb. 2023, doi: https://doi.org/10.3390/s23041811.

[39] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors*, vol. 23, no. 4, p. 2333, Feb. 2023, doi: https://doi.org/10.3390/s23042333.

[40] M. J. Lewis, A. Spiliopoulou, K. Goldmann, C. Pitzalis, P. McKeigue, and M. R. Barnes, "Nestedcv: An R package for fast implementation of nested cross-validation with embedded feature selection designed for transcriptomics and high-dimensional data," *Bioinformatics Advances*, vol. 3, no. 1, p. 5, 2023, doi: https://doi.org/10.1093/bioadv/vbad048.

[41] M. Tahkola and Z. Guangrong, "ATSC-NEX: Automated Time Series Classification With Sequential Model-Based Optimization and Nested Cross-Validation," *IEEE Access*, vol. 10, pp. 39299–39312, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3166525.

[42] A. Toufah, M. A. Kadim, and M. Y. El Hafidi, "Overcoming SVM limitations in lung cancer classification with a quantum feature map," *BMC Artificial Intelligence*, vol. 1, no. 15, p. 13, Dec. 2025, doi: https://doi.org/10.1186/s44398-025-00016-3.

[43] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3151048.

[44] D. Chicco and G. Jurman, "A statistical comparison between Matthews correlation coefficient (MCC), prevalence threshold, and Fowlkes–Mallows index," *J. Biomed. Inform.*, vol. 144, p. 7, Aug. 2023, doi: https://doi.org/10.1016/j.jbi.2023.104426.

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

[45] P. Thölke *et al.*, "Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data," *Neuroimage*, vol. 277, p. 14, Aug. 2023, doi: https://doi.org/10.1016/j.neuroimage.2023.120253.

[46] F. M. J. M. Shamrat *et al.*, "AlzheimerNet: An Effective Deep Learning Based Proposition for Alzheimer's Disease Stages Classification From Functional Brain Changes in Magnetic Resonance Images," *IEEE Access*, vol. 11, pp. 16376–16395, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3244952.

[47] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, no. 6, p. 11, Jun. 2024, doi: https://doi.org/10.1016/j.patter.2024.100994.

[48] O. Ismael, "The Standard Deviation Score: a novel similarity metric for data analysis," *J. Big Data*, vol. 12, no. 1, p. 42, Dec. 2025, doi: https://doi.org/10.1186/s40537-025-01091-z.

[49] H. S. Darling, "Are you confident about your confidence in confidence intervals?," *Cancer Research, Statistics, and Treatment*, vol. 5, no. 1, pp. 139–144, Jan. 2022, doi: https://doi.org/10.4103/crst.crst_75_22.

[50] D. Rawat, Meenakshi, L. Pawar, G. Bathla, and R. Kant, "Optimized Deep Learning Model for Lung Cancer Prediction Using ANN Algorithm," in 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), IEEE, 2022, pp. 889–894. doi: https://doi.org/10.1109/icesc54411.2022.9885607.

[51] J. Viji Cripsy, and T. Divya, "Lung Cancer Disease Prediction and Classification based on Feature Selection method using Bayesian Network, Logistic Regression, J48, Random Forest, and Naïve Bayes Algorithms," in 2023 3rd International Conference on Smart Data Intelligence (ICSMDI), IEEE, 2023, pp. 335–342. doi: https://doi.org/10.1109/icsmdi57622.2023.00066.

[52] A. Y. Prity, A. Hossain, and S. Sultana, "FT-HELC: A Fine-tuned Hybrid Ensemble Technique for Lung Cancer Classification and Prediction Using a Clinical Dataset," in 2025 2nd International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM), IEEE, 2025, pp. 1–6. doi: https://doi.org/10.1109/ncim65934.2025.11160004.

[53] S. M. Ganie, P. K. Dutta Pramanik, and Z. Zhao, "Enhanced and Interpretable Prediction of Multiple Cancer Types Using a Stacking Ensemble Approach with SHAP Analysis," Bioengineering, vol. 12, no. 5, Art. no. 472, 2025. doi: https://doi.org/10.3390/bioengineering12050472.

Author Biography



Irwan Budiman is a lecturer at Lambung Mangkurat University. He earned his Bachelor's Degree in Informatics Engineering from Islam Indonesia University, Yogyakarta. Subsequently, he completed his Master's studies in information systems at Diponegoro University, Semarang. His research interests include data mining, human-computer interaction, applied business intelligence, and e-government. Email: irwan.budiman@ulm.ac.id. Orcid ID: 0000-0002-0514-7429.



Nanda Putri Rahmawati is a computer science student at the Faculty of Mathematics and Natural Sciences, Universitas Lambung Mangkurat, having commenced her studies in 2020. Her research interests lie in the field of medical data classification and machine learning performance evaluation. Her final project, titled "Comparison Between K-Fold Cross Validation and Percentage Split in Decision Tree Algorithms for Anemia Classification" focuses on analyzing the effectiveness of different data testing methods to optimize the accuracy of anemia detection.



Annisa Rizqiana is a lecturer in the Department of Computer Science at Universitas Lambung Mangkurat (ULM), South Kalimantan, Indonesia. She earned her Bachelor's degree in Computer Science from ULM and her Master's degree in Computer Science from Universitas Gadjah Mada (UGM). Her research interests focus on bioinformatics, data science, and machine learning, particularly in applying computational methods to biological data and complex datasets. In teaching activities, she teaches courses that focus on mathematical concepts as the foundation of programming, including mathematical logic, discrete mathematics, relations and functions, and basic algebra to support algorithm development and computational problem solving. She is also actively involved in supervising undergraduate students and academic research. She can be contacted via email at annisarizqiana@ulm.ac.id.

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).



Rafi'e was born in Pulang Pisau, Indonesia, in 1993. He received a Bachelor's degree in Informatics Engineering from Universitas Islam Kalimantan Muhammad Arsyad Al Banjari (Uniska MAB), Banjarmasin, Indonesia, and a Master's degree in Computer Science from Universitas Budi Luhur (UBL), Jakarta, Indonesia. His area of expertise is Applied Computational Engineering, with research interests in Artificial Intelligence, Machine Learning, Computer Vision, and Edge AI. His research focuses on the development of intelligent computational models for detection, classification, prediction, and real-time data processing. In addition to his research and scientific publication activities, he is actively involved in the development and implementation of digital technology solutions to support digital transformation and address various challenges in industrial sectors. He can be contacted via email at rafiekom@ulm.ac.id.

Manuscript received 17 June 2026; Accepted 20 August 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.367
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).