Scientific Research Publishing

Journal of Cancer Therapy, 2024, 15, 287-302
https://www.scirp.org/journal/jct
ISSN Online: 2151-1942
ISSN Print: 2151-1934

Prediction of Lung Cancer Stage Using Tumor Gene Expression Data

Yadi Gu

School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China
Email: 202121030260@stu.zuel.edu.cn

How to cite this paper: Gu, Y.D. (2024) Prediction of Lung Cancer Stage Using Tumor Gene Expression Data. *Journal of Cancer Therapy*, **15**, 287-302.
https://doi.org/10.4236/jct.2024.158028

Received: July 26, 2024
Accepted: August 25, 2024
Published: August 28, 2024

Abstract

Lung cancer remains a significant global health challenge and identifying lung cancer at an early stage is essential for enhancing patient outcomes. The study focuses on developing and optimizing gene expression-based models for classifying cancer types using machine learning techniques. By applying Log2 normalization to gene expression data and conducting Wilcoxon rank sum tests, the researchers employed various classifiers and Incremental Feature Selection (IFS) strategies. The study culminated in two optimized models using the XGBoost classifier, comprising 10 and 74 genes respectively. The 10-gene model, due to its simplicity, is proposed for easier clinical implementation, whereas the 74-gene model exhibited superior performance in terms of Specificity, AUC (Area Under the Curve), and Precision. These models were evaluated based on their sensitivity, AUC, and specificity, aiming to achieve high sensitivity and AUC while maintaining reasonable specificity.

Keywords

Lung Cancer Detection, Stage Prediction, Gene Expression Data, Xgboost, Machine Learning

1. Introduction

Lung cancer remains a critical medical unmet need due to its high incidence and mortality rates, compounded by the challenges of early detection and accurate diagnosis [1]. Despite advancements in treatment, lung cancer continues to be diagnosed at advanced stages in a significant number of patients, which drastically reduces the efficacy of curative treatments. Current diagnostic methods, including systemic chemotherapy, local radiation therapy, and targeted therapy (including most recently immunotherapy), often lack the sensitivity and specificity necessary for early-stage detection [2]. Consequently, there is a pressing need for improved diagnostic tools that can detect lung cancer at an earlier stage and accurately characterize its subtypes. Innovations such as molecular and genetic profiling hold promise for addressing these challenges, potentially enabling personalized treatment approaches that could improve survival rates and quality of life for patients. The importance of early and precise diagnosis is underscored in studies like Field *et al.*, which highlight the benefits of early detection strategies in lung cancer screening [3]. Additionally, the National Comprehensive Cancer Network (NCCN) guidelines emphasize the need for enhanced diagnostic accuracy to guide effective treatment planning [4]. Furthermore, Hirsch *et al.* discuss current therapies and the potential of new targeted treatments, reinforcing the need for better diagnostic methodologies to optimize these therapeutic advancements [5].

Accurately distinguishing between early-stage and late-stage lung cancer in the clinical setting is crucial because it directly impacts the treatment strategy and prognosis for patients. Early-stage lung cancer, typically confined to the lungs, can often be treated with curative intent through surgical resection and localized therapies, significantly improving survival rates. In contrast, late-stage lung cancer, which has metastasized beyond the lungs, requires more aggressive systemic treatments such as chemotherapy, targeted therapy, or immunotherapy, aimed primarily at prolonging life and alleviating symptoms rather than achieving a cure. Misclassification of the cancer stage can lead to inappropriate treatment plans, potentially causing ineffective treatment or unnecessary side effects. Accurate staging also informs clinical decision-making and patient counseling, helping healthcare providers offer the most appropriate care and set realistic expectations for outcomes. Studies underscore the importance of precise staging; for instance, the National Comprehensive Cancer Network (NCCN) guidelines highlight that proper staging is essential for selecting the optimal therapeutic approach [4]. Moreover, research by Goldstraw *et al.* in the IASLC Lung Cancer Staging Project emphasizes the prognostic significance of accurate staging in guiding treatment decisions and improving patient outcomes [6].

The promise of gene expression-based diagnosis for lung cancer stages lies in its ability to enhance the precision and accuracy of cancer staging, ultimately leading to better patient outcomes. Traditional diagnostic methods, such as imaging and histopathology, often struggle to detect early-stage lung cancer and accurately characterize tumor subtypes. In contrast, gene expression profiling enables the identification of specific molecular signatures associated with different stages of lung cancer. This molecular approach facilitates earlier detection and more precise staging, which are critical for tailoring personalized treatment strategies. Early-stage lung cancer can be more effectively treated with localized therapies, while advanced-stage cancer requires systemic treatments. Accurate staging through gene expression profiling ensures that patients receive the most appropriate therapies, potentially improving survival rates and reducing treatment-related morbidity. Studies such as Shedden *et al.* have demonstrated the utility of gene expression-based models in predicting survival and disease progression in lung cancer patients [7]. Kratz *et al.* validated the practical application of molecular assays in predicting outcomes in non-small cell lung cancer [8]. Moreover, Chen *et al.* highlighted the role of gene expression signatures in distinguishing histological subtypes and stages of lung cancer [9]. Additionally, Roepman *et al.* underscored the clinical relevance of gene expression profiling in predicting prognosis and guiding treatment decisions [10]. These advancements underscore the transformative potential of gene expression-based diagnostics in improving the accuracy of lung cancer staging and optimizing patient care.

Here in this paper, we intend to develop a machine learning model to distinguish between early and late-stage lung cancer based on gene expression profiles. Using count data for unified preprocessing and Log2 normalization for stability, we employed the Wilcoxon Rank Sum Test and Iterative Feature Selection (IFS) to identify key genes. Classifiers such as XGBoost, SVM, and Random Forest were used to optimize model performance. Our findings demonstrate the potential of gene expression-based diagnostics to enhance lung cancer staging accuracy, enabling more personalized and effective treatment strategies.

2. Results

We downloaded lung cancer tissue gene expression data from The Cancer Genome Atlas (TCGA) public data portal. The gene expression was measured by RNA sequencing, representing a whole-transcriptome-wide profiling of gene activity. RNA sequencing provides a comprehensive and detailed view of the gene activity within the cancer tissues, allowing for a thorough analysis of gene expression patterns. The count data was used, ensuring a unified preprocessing. The downloaded dataset includes 992 samples and 19,938 genes, comprising 774 early-stage lung cancer patients and 218 late-stage lung cancer patients. We then aimed to build a machine learning classification model to distinguish between early-stage and late-stage lung cancer patients based on their gene expression profiles.

Given the complexity and variability of gene expression data, it is crucial to normalize the data effectively to ensure accurate and reliable analysis. The expression levels of different genes in the raw data varied greatly, with differences up to five orders of magnitude. Log2 normalization was employed to improve the stability of the model and reduce the impact of noise on the model parameters. This normalization method retains the relative relationships between samples under each feature and preserves the relative relationships between features. By limiting the size of parameters, the model becomes more robust to small changes in the input data, thereby reducing the risk of overfitting.

Next, the Wilcoxon Rank Sum Test was used to detect significant differences in the median gene expression levels between early-stage and late-stage samples for each gene. Genes with significant differences were identified based on p-values. This statistical method helps in identifying key genes that could potentially differentiate between early and late stages of lung cancer, providing a foundation for further analysis. Since the optimal subset of features should be as small as possible while ensuring metrics like sensitivity, we initially selected the top 500 genes with the smallest p-values. Subsequently, using Iterative Feature Selection (IFS) based on these 500 genes, we determined the optimal feature subset.

After splitting the data into training and testing sets with an 8:2 ratio, and further dividing the training set into training and validation sets also at an 8:2 ratio, we obtained 499 early-stage samples and 135 late-stage samples in the training set, 119 early-stage samples and 40 late-stage samples in the validation set, and 156 early-stage samples and 43 late-stage samples in the test set. This meticulous partitioning ensures that the models are trained, validated, and tested on separate subsets of data, thereby enhancing the reliability of the performance metrics. The dataset division is detailed in Table 1.

Table 1. Dataset used to train, validate and test the model performance.

	Early stage	Late stage	Total
Training set	499	135	634
Validation set	119	40	159
Test set	156	43	199
Total	774	218	992

After identifying all genes with significant p-values and selecting the 500 genes with the smallest p-values, we still needed to determine the optimal number of genes to use. To balance a small number of selected genes with excellent classification performance, we employed Incremental Feature Selection (IFS). In the IFS process, we constructed a series of feature sets, $F = \left[f_1, f_2, \cdots, f_N \right]$, where N ranged from 1 to 500. For each feature set containing the first n genes with the lowest p-values, we built corresponding classifiers, including XGBoost, SVM, and Random Forest, using default parameters in the training set and determined the best decision threshold using the validation set. This thorough approach ensures that the selected features and classifiers are optimized for the best performance. Finally, we evaluated the performance on the test set. This process provides various metrics corresponding to different feature sets, such as sensitivity, recall, and AUC. By analyzing the IFS curves, we balanced model complexity and classification performance. Optimal selection is achieved when the number of features is minimum and the performance score is highest. The whole process is shown in Figure 1.

Figure 2 displays the performance metrics obtained using three classifiers and the IFS strategy on each feature set, in which the x-axis indicates the number of features, while the y-axis represents different metrics values. Considering the nature of bioinformatics data, this study evaluates the model performance based on sensitivity, AUC, specificity, and the number of features. High sensitivity and

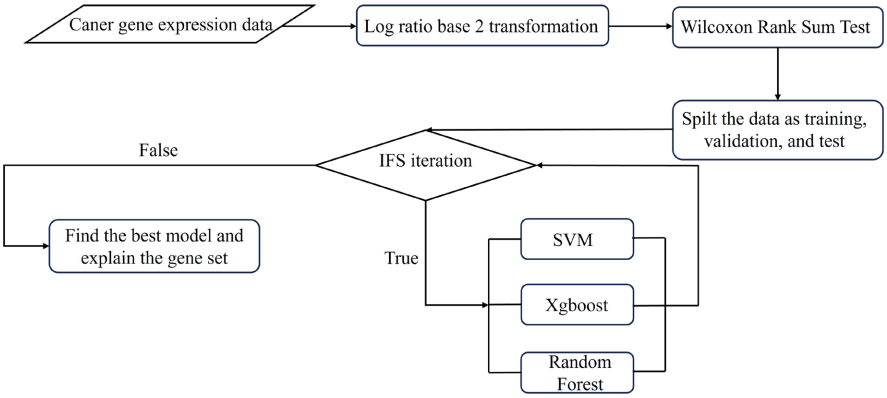
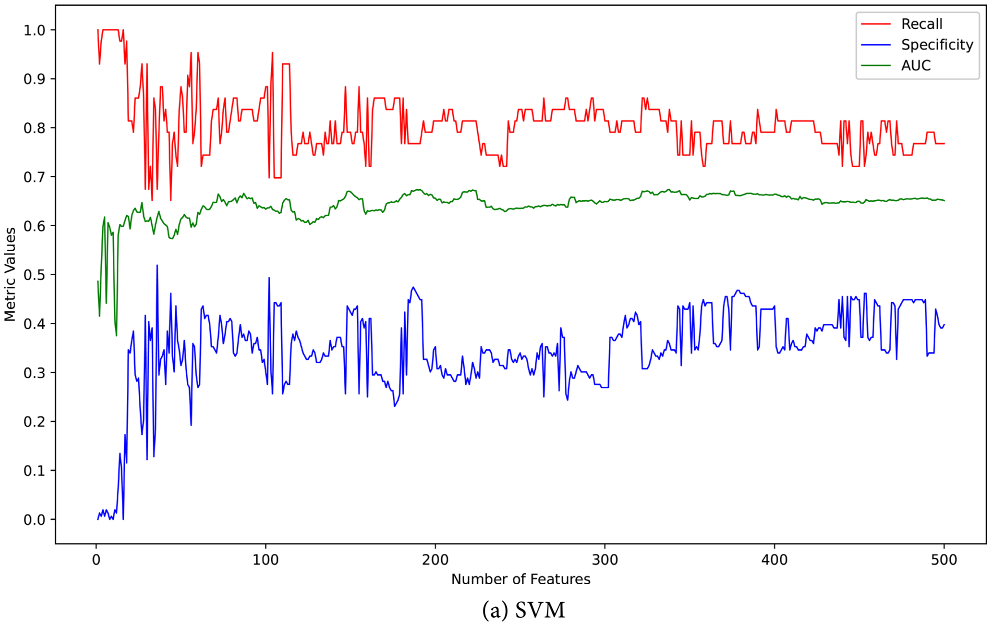
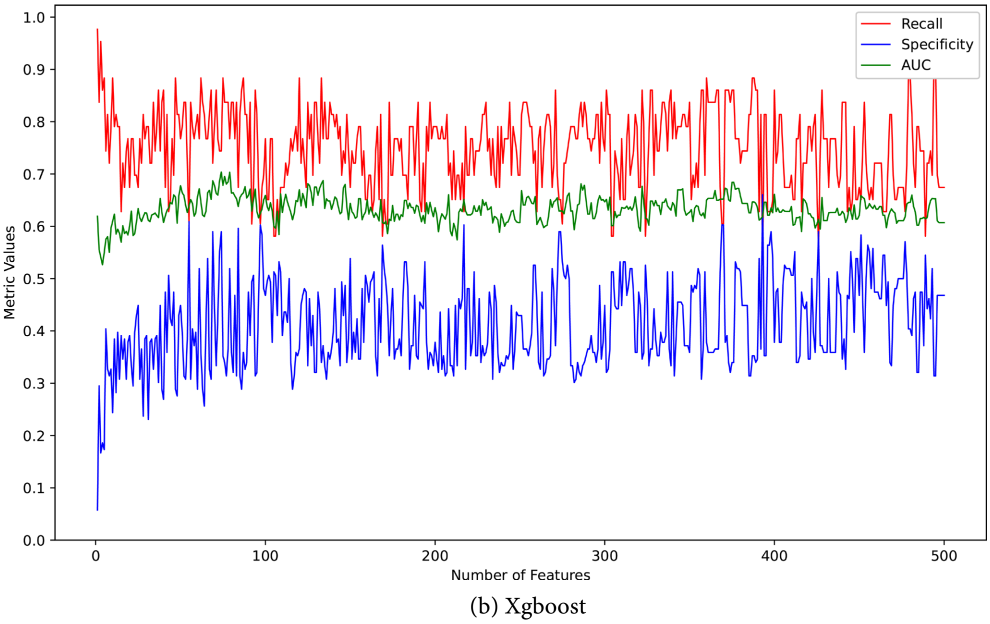

Caner gene expression data
Log ratio base 2 transformation
Wilcoxon Rank Sum Test
Spilt the data as training, validation, and test
IFS iteration
False
True
Find the best model and explain the gene set
SVM
Xgboost
Random Forest

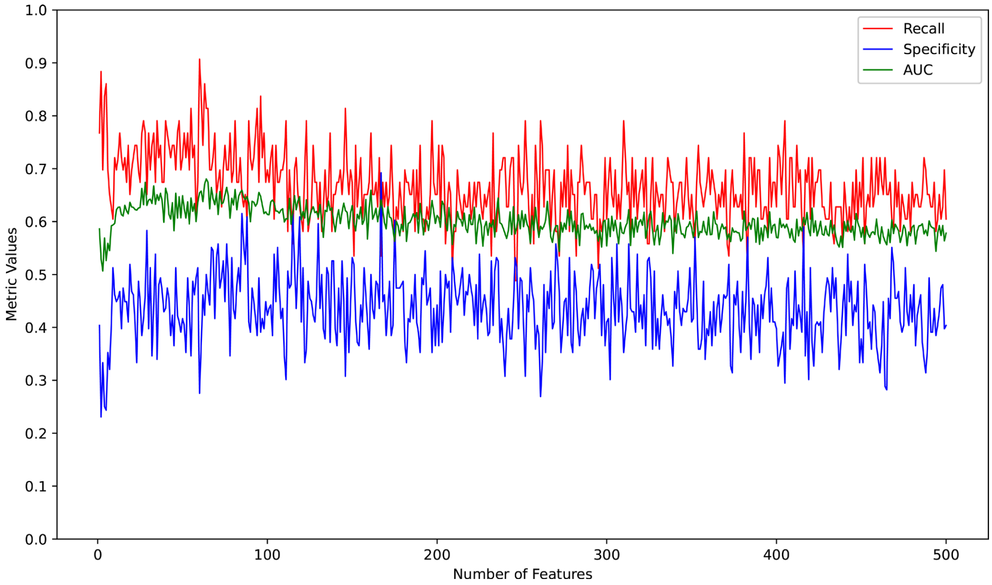
Figure 1. Schematic illustration of the classification model.



(a) SVM



(b) Xgboost



(c) Random Forest

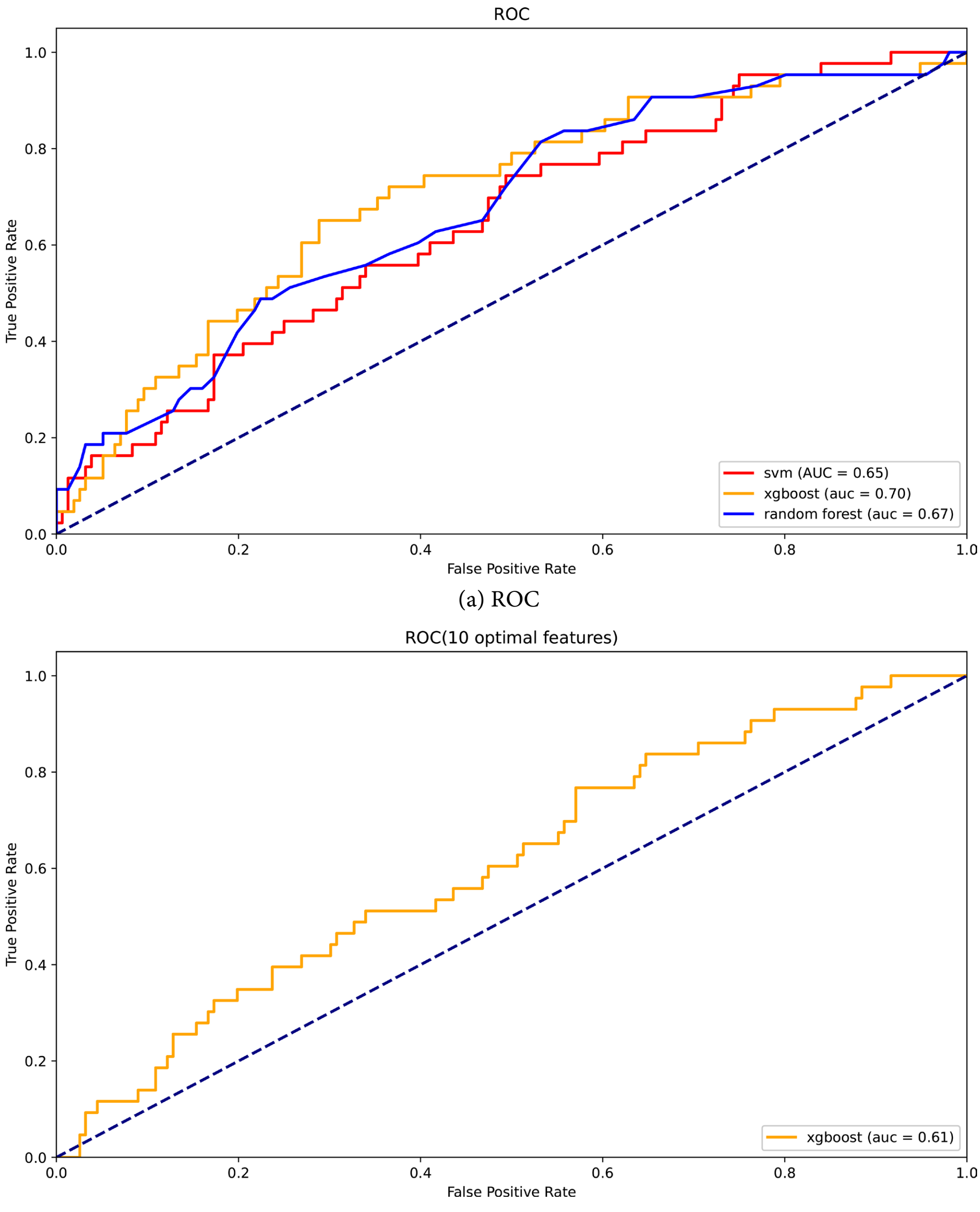
Figure 2. IFS iteration evaluation for the three models.

AUC is critical for ensuring that the model accurately identifies true positives, while specificity ensures that false positives are minimized. Our goal is to achieve high sensitivity and AUC while maintaining a certain level of specificity. It can be observed that XGBoost outperforms the other classifiers in general. On the top 74 gene set, XGBoost achieves a sensitivity of 0.8837, an AUC of 0.7038, and a specificity of 0.5897. Another notable performance is seen in the top 10 gene set, where XGBoost achieves a sensitivity of 0.8837 and a specificity of 0.3846, which is relatively higher among classifiers with fewer features, although its AUC score is slightly lower at 0.6091. In comparison, SVM and Random Forest exhibit AUC scores ranging from 0.6 to 0.65, although they demonstrate high sensitivity in certain feature sets. Additionally, the average specificity obtained by SVM is generally lower than that of XGBoost, while the average specificity of RandomForest is similar to that of XGBoost.

Figure 3 illustrates the ROC curves and AUC values obtained by each classifier for their respective best classification results. The SVM classifier achieves an AUC of 0.6477 using a gene set consisting of the top 69 genes, the XGBoost classifier selects the top 74 gene set with an AUC of 0.7038, and the Random Forest classifier obtains an AUC of 0.6746 using the top 65 gene set. The optimal number of gene combinations for all three classifiers is approximately 70.

Based on the IFS iteration plot of XGBoost, we obtained two models. One model consists of a larger set of 74 optimal features, and the functions of these 74 features are listed in Table 3. Additionally, we are particularly interested in the model with only the top 10 genes, as smaller gene sets are more practical for cancer stage classification. This focus on smaller gene sets is crucial for developing practical diagnostic tests that can be easily implemented in clinical settings. This allows us to perform specific gene expression level testing on potential patients. The weights and functions of each gene in both models are shown in Table 2 and Table 3. The weight is between 0 and 100, where 100 represents the largest weight possible, and 0 represents the gene that has no contribution to the model. Figure 4 shows the confusion matrices obtained by training and predicting with two different models after re-dividing the data.

It can be observed that these genes have diverse functions related to cellular regulation, metabolism, extracellular matrix, neuromuscular function, mitochondrial function, transcriptional regulation, and olfactory perception. Their abnormal expression may contribute to various aspects of lung cancer, including



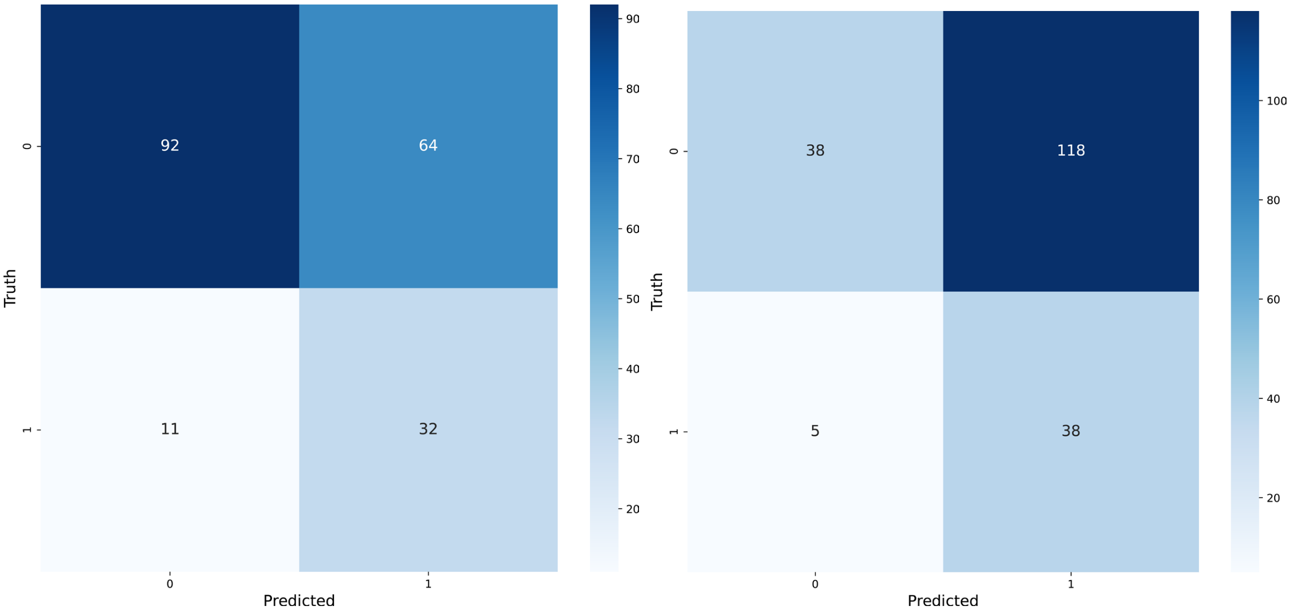
(a) ROC

(b) ROC (10 optimal feature)

Figure 3. ROC curve for evaluating models which incorporates 74 optimal features and 10 optimal features.

Table 2. The 10 optimal features with weight.

Feature	weight	Function
PCBD1	90.95	Encodes a multifunctional protein: a dehydratase essential for tetrahydrobiopterin biosynthesis and a cofactor for HNF1A-dependent transcription. Deficiency causes hyperphenylalaninemia.
IRAG2	90.00	Encodes a protein that is developmentally expressed in lymphoid cells and localized to the endoplasmic reticulum's cytoplasmic face.
ALDH8A1	83.81	Encodes an aldehyde dehydrogenase involved in 9-cis-retinoic acid synthesis, tryptophan breakdown, and kynurenine pathway.
SERPINE3	76.67	Predicted to encode a protein with serine-type endopeptidase inhibitor activity, involved in negatively regulating endopeptidase activity, and likely active in the extracellular spa
RAPSN	74.29	playing a critical role in synaptic function and potentially contributing to postsynaptic congenital myasthenic syndromes.
MRPS16	73.33	Encodes a highly conserved ribosomal protein that contributes to protein synthesis within the mitochondria, specifically as a component of the small 28S subunit of mammalian mitoribosomes.
CHCHD1	68.57	Enables RNA binding activity. Predicted to be involved in mitochondrial translation.
C9orf131	60.48	
POU2AF1	54.76	Facilitates transcription coactivator activity. Participates in the positive regulation of transcription by RNA polymerase II. Constitutes a component of the RNA polymerase II transcription regulator complex.
OR2A5	27.14	Encodes an olfactory receptor protein, a member of the G-protein-coupled receptor family, responsible for recognizing odorant molecules, initiating neuronal responses, and mediating the transduction of odorant signals, contributing to the perception of smell.



(a) Xgboost with 74 features

(b) Xgboost with 10 features

Figure 4. Confusion matrix for the test dataset for both 74-feature and 10-feature models.

Table 3. The 74 optimal features with weight and function.

Feature	weight	Function
AHSA1	95.82	Enables ATPase activation, Hsp90 protein binding, and chaperone binding, contributing to the positive regulation of ATPase activity and functioning in the cytosol.
PM20D1	92.52	Enables hydrolase activity on carbon-nitrogen bonds in linear amides, critical for amide biosynthesis, catabolism, and neuronal survival regulation.
DECR1	89.21	Involved in fatty acid beta-oxidation, performing 2,4-dienoyl-CoA reductase activity and participating in protein binding within the cytosol, mitochondrion, and nucleoplasm as part of a catalytic complex.
COLQ	85.91	Encodes a subunit of a collagen-like molecule that binds and anchors acetylcholinesterase in skeletal muscle, essential for the formation of the enzyme complex and associated with endplate acetylcholinesterase deficiency.
SULT2B1	82.60	Sulfotransferase enzymes sulfate hormones, neurotransmitters, drugs, and xenobiotics, differ in tissue distribution and substrate specificity. They sulfate dehydroepiandrosterone specifically and exhibit two spliced variants.
AMY2B	79.30	Encodes an amylase isoenzyme that is primarily produced by the pancreas and plays a crucial role in the initial breakdown of dietary starch and glycogen during digestion.
CRB2	72.69	Encodes a protein critical for embryonic cell polarity and development, mutations linked to kidney diseases.
MICU1	72.69	Encodes a key regulator of basal mitochondrial Ca^{2+} uptake. It interacts with the mitochondrial calcium uniporter, an inner membrane Ca^{2+} channel, and is essential in preventing mitochondrial Ca^{2+} overload.
GNPNAT1	72.69	Encodes a protein predicted to bind proteins, synthesize UDP-N-acetylglucosamine, and regulate cell response to leukemia inhibitory factor in various cell compartments.
URM1	72.69	Enables sulfur carrier activity and is involved in tRNA thio-modification and wobble uridine modification predicted to function in the cytosol and nucleus.
SPR	69.39	Encodes an aldo-keto reductase enzyme critical for the biosynthesis of tetrahydrobiopterin (BH4), and mutations in this gene cause DOPA-responsive dystonia due to sepiapterin reductase deficiency.
COMTD1	59.48	Predicted to enable S-adenosylmethionine-dependent methyltransferase activity, be involved in methylation, and be an integral component of the membrane.
GOT1	59.48	Encodes glutamic-oxaloacetic transaminase, an enzyme involved in amino acid metabolism, the urea cycle, and the tricarboxylic acid cycle,
CAPNS1	59.48	Encodes a small regulatory subunit essential for the stability and function of calpain heterodimers, calcium-dependent cysteine proteinases involved in diverse cellular processes such as apoptosis, proliferation, migration, adhesion, and autophagy, with implications in neurodegenerative disorders like myotonic dystrophy.
XKR6	59.48	Predicted to regulate apoptosis during development, including cell engulfment and phosphatidylserine exposure on cell surfaces. Integral membrane protein is active in the plasma membrane.
JAGN1	59.48	Encodes a transmembrane protein involved in the early secretory pathway, essential for neutrophil differentiation and survival, with mutations causing severe congenital neutropenia.
TRUB2	59.48	A prevalent component of rRNAs and tRNAs, produced enzymatically through the isomerization of uridine by pseudouridine synthase.
PHF19	56.17	Enables methylated histone binding activity. Involved in positive regulation of histone H3-K27 methylation.
LRIT3	56.17	Encodes a protein with fibronectin, leucine-rich repeat, and immunoglobulin domains, regulating fibroblast growth factor receptors. Mutations in this gene are linked to congenital stationary night blindness, type 1F.

Continued

ANKHD1-EIF4EBP3	56.17	Remains undetermined, requiring further investigation to understand its significance and the function of the protein it encodes.
NSG1	52.87	Predicted to enable clathrin light chain binding activity. Involved in the apoptotic process.
AP3M1	52.87	Encodes the medium subunit of AP-3, an adaptor protein complex that facilitates vesicle budding from the Golgi and sorts proteins to the endosomal/lysosomal system. Mutations in AP-3 are linked to Hermansky-Pudlak syndrome
GHITM	52.87	Involved in inner mitochondrial membrane organization and negative regulation of release of cytochrome c from mitochondria.
RAPSN	52.87	Encodes a protein involved in clustering and anchoring nicotinic acetylcholine receptors at synaptic sites, playing a critical role in synaptic function and potentially contributing to postsynaptic congenital myasthenic syndromes.
RORA	49.56	Encodes a nuclear hormone receptor that regulates gene expression by binding to hormone response elements, interacts with NM23 proteins, and aids in transcriptional regulation of genes involved in circadian rhythm, organogenesis, and tumor metastasis.
CHCHD1	49.56	Enables RNA binding activity. Predicted to be involved in mitochondrial translation.
IRAG2	49.56	Encodes a protein that is developmentally expressed in lymphoid cells and localized to the endoplasmic reticulum's cytoplasmic face.
RAB30	49.56	Predicted to enable GTP binding activity and GTPase activity. Involved in Golgi organization.
LPAR6	49.56	Encodes a G-protein coupled receptor activated by adenosine and uridine nucleotides. Resides within an intron of the retinoblastoma susceptibility gene in the reverse orientation, and undergoes alternative splicing to produce multiple transcript variants.
VENTX	49.56	Encodes a Vent family homeodomain protein that likely acts as a transcriptional repressor involved in mesodermal patterning and maintenance of hemopoietic stem cells. Also associated with antigen production in certain melanomas.
ANXA7	49.56	Encodes a membrane-binding protein involved in calcium channel activity, ion selectivity, and membrane fusion, expressed mainly in brain, heart, and muscle.
CLEC17A	46.26	Enables fucose binding activity; identical protein binding activity; and mannose binding activity.
KRT8	46.26	Encodes a type II keratin protein that forms intermediate filaments with keratin 18, contributing to the structural integrity of epithelial cells, while also participating in signal transduction, cellular differentiation, and being associated with cryptogenic cirrhosis when mutated.
MSTN	46.26	Encodes a TGF-beta ligand regulating skeletal muscle cell growth, mutations linked to muscle hypertrophy.
ATP2B2	46.26	Encodes a plasma membrane calcium ATPase isoform 2, essential for regulating intracellular calcium levels by actively removing calcium ions from eukaryotic cells, and its expression and splicing are finely tuned to specific physiological needs across tissues and cell types.
MRPL15	46.26	Encodes a mitochondrial ribosomal protein involved in protein synthesis within the mitochondrion, specifically belonging to the EcoL15 ribosomal protein family.
CCDC180	46.26	Encodes a protein with a coiled-coil domain, involved in various isoforms through alternative splicing, and a specific single nucleotide polymorphism (SNP) in this gene is linked to heightened susceptibility to Behcet's Disease.
CARD17	46.26	
POU2AF1	46.26	Facilitates transcription coactivator activity. Participates in the positive regulation of transcription by RNA polymerase II. Constitutes a component of the RNA polymerase II transcription regulator complex.

Continued

ABCA10	46.26	Encodes a membrane-associated protein belonging to the ABC1 subfamily of ATP-binding cassette transporters, but its specific substrate and function are currently unknown.
KLRK1	42.95	Encodes a transmembrane protein belonging to the NKG2 gene family, which plays a crucial role in activating natural killer (NK) cells and T cells through interactions with ligands, thus contributing to immune responses against stressed cells and serving as a potential therapeutic target for immune diseases and cancers.
PNOC	42.95	Encodes a preproprotein that is processed to generate nociceptin/orphanin FQ, a neuropeptide that regulates pain sensitivity, body temperature, learning, memory, and hunger, as well as nocistatin, which may inhibit nociceptin's effects.
STARD5	39.65	Encodes a START-domain cholesterol transporter protein that facilitates the trafficking of cholesterol and sterol-derived molecules between intracellular membranes, with upregulated expression during endoplasmic reticulum stress.
BZW1	39.65	Enables RNA binding activity and cadherin binding activity.
MEI1	39.65	Implicated in meiosis I, gamete generation, meiotic spindle organization, meiotic telomere clustering, and gestational trophoblastic neoplasm.
FAM118A	36.35	Enables identical protein binding activity. Predicted to be an integral component of the membrane.
CHMP4A	36.35	Participates in the ESCRT-III complex, facilitating the degradation of surface receptor proteins, formation of endocytic multivesicular bodies (MVBs), and potentially contributing to the cell cycle regulation.
NEUROD2	36.35	Encodes a neurogenic basic helix-loop-helix protein that plays a crucial role in inducing and maintaining neuronal cell fates by promoting neurogenic differentiation and activating neuron-specific promoter.
FUCA2	33.04	Encodes Alpha-L-fucosidase which removes sugar from glycoproteins, essential for bacterial adhesion to gastric cancer cells.
PRR33	33.04	Predicted to act upstream of or within response to wounding.
AC114490.3	33.04	
PCBD1	33.04	Encodes a multifunctional protein: a dehydratase essential for tetrahydrobiopterin biosynthesis and a cofactor for HNF1A-dependent transcription. Deficiency causes hyperphenylalaninemia.
ZNHIT1	33.04	Involved in regulating histone deacetylation, nucleosome binding, and negative regulation of multiple cellular processes, primarily within the nucleoplasm.
CISD1	33.04	Encodes a protein with a CDGSH iron-sulfur domain that binds a redox-active [2Fe-2S] cluster, localized to the mitochondrial outer membrane, and involved in the regulation of oxidation.
SERPINE3	33.04	Predicted to encode a protein with serine-type endopeptidase inhibitor activity, involved in negatively regulating endopeptidase activity, and likely active in the extracellular spa.
TAS2R43	33.04	Expressed on the surface of taste receptor cells and mediate the perception of bitterness through a G protein-coupled second messenger pathway.
CLIC1	33.04	Encodes Chloride Intracellular Channel 1 (CLIC1), involved in regulating cellular processes, such as membrane potential stabilization, through its activity as a chloride ion channel localized in the cell nucleus and plasma membrane.
ECD	33.04	Facilitates histone acetyltransferase binding activity. Plays a role in the positive regulation of RNA polymerase II-mediated transcription.
ALDH8A1	29.74	Encodes an aldehyde dehydrogenase involved in 9-cis-retinoic acid synthesis, tryptophan breakdown, and kynurenine pathway.

Continued

OR2A5	29.74	Encodes an olfactory receptor protein, a member of the G-protein-coupled receptor family, responsible for recognizing odorant molecules, initiating neuronal responses, and mediating the transduction of odorant signals, contributing to the perception of smell.
STIMATE-MUSTN1	29.74	Generates a read-through transcript between TMEM110 and MUSTN1 genes, resulting in a fusion protein containing sequences from both genes.
SIGMAR1	29.74	Encodes a receptor protein involved in cellular functions of multiple systems, interacting with psychotomimetic drugs, and associated with juvenile amyotrophic lateral sclerosis 16, despite its initial misclassification as an opioid receptor.
ANKK1	26.43	Encodes a Ser/Thr protein kinase involved in signal transduction pathways and is closely linked to the DRD2 gene, harboring a well-studied polymorphism known as TaqIA, potentially implicated in neuropsychiatric disorders.
DDX50	26.43	Encodes a DEAD box enzyme, a putative RNA helicase, implicated in various cellular processes such as RNA structure alteration, ribosomal RNA synthesis, and processing, with potential involvement in embryogenesis, spermatogenesis, and cellular growth and division.
ASCL4	26.43	ASIC helix-loop-helix transcription factors, like ASCL4, are crucial for cell fate determination and development/differentiation of multiple tissues.
MRPS16	26.43	Encodes a highly conserved ribosomal protein that contributes to protein synthesis within the mitochondria, specifically as a component of the small 28S subunit of mammalian mitoribosomes.
GSG1L2	23.13	Predicted to be an integral component of the membrane and be active in the plasma membrane
TNFRSF13C	23.13	Encodes a receptor for B cell-activating factor (BAFF), crucial for the survival of mature B cells and potentially involved in the pathogenesis of autoimmune diseases by regulating autoreactive B cell survival.
FCRL5	23.13	Encodes a membrane protein belonging to the immunoglobulin receptor superfamily, involved in B cell development and lymphomagenesis, and contains immunoglobulin-like domains, playing a role in immune responses and signaling.
KRT18	23.13	Encodes keratin 18 which is widely expressed in single-layer epithelial tissues and is associated with cryptogenic cirrhosis when mutated.
C9orf131	19.83	
IRF4	19.83	Encodes protein belonging to IRF family members negatively regulate TLR signaling and may contribute to multiple myeloma via chromosomal translocation.
GPR55	9.91	Implicated in a range of physiological and pathological processes through the activation of various signal transduction pathways.
OR56B1	3.30	Encodes olfactory receptor proteins that detect odorant molecules, initiating neuronal responses and facilitating the perception of smell, as part of the largest gene family in the genome.

development, progression, angiogenesis, metastasis, muscle invasion, mitochondrial dysfunction, and cell differentiation. Understanding these functions helps in elucidating the biological mechanisms underlying lung cancer and identifying potential therapeutic targets. Further research is required to fully elucidate the specific roles of these genes in lung cancer and their potential as diagnostic markers or therapeutic targets.

3. Discussion

Current lung cancer diagnostic methods have several issues. They often detect cancer late due to non-specific early symptoms and have limited sensitivity and specificity, leading to false results. Invasive procedures like biopsies carry risks, and imaging exposes patients to radiation. High costs and limited access to advanced tools create disparities, and some methods are time-consuming. Accurate interpretation of results requires specialized expertise, and the accuracy varies by technique and provider experience. Additionally, early detection programs are insufficient, reducing early diagnosis chances. There is a need for more accurate, non-invasive, cost-effective, and widely accessible diagnostic methods.

Gene expression-based cancer diagnosis offers several key features. It provides high precision and accuracy by identifying specific molecular signatures associated with different cancer stages, facilitating accurate staging. This method also enables early detection of cancer, which is crucial for effective treatment and improved patient outcomes. Additionally, gene expression profiling supports personalized treatment strategies, increasing the likelihood of successful therapy and minimizing unnecessary side effects. RNA sequencing offers a comprehensive view of gene activity within cancer tissues, allowing for detailed analysis of gene expression patterns. However, clinical implementation requires consideration of cost, complexity, and validation across diverse populations. Overall, gene expression-based diagnosis has significant potential to improve the accuracy of cancer staging and enhance patient care.

This study primarily develops machine learning models to classify lung cancer stages based on gene expression profiles. The research used gene expression data from The Cancer Genome Atlas (TCGA), comprising 992 samples and 19,938 genes, with 774 early-stage and 218 late-stage lung cancer patients. Through the Wilcoxon rank sum test, the study identified significant differences in gene expression between early and late-stage samples, initially selecting the 500 genes with the smallest p-values. Incremental Feature Selection (IFS) was then employed to determine the optimal feature subset.

Three classifiers like XGBoost, Support Vector Machine (SVM), and Random Forest are used with data divided into training (64%), validation (16%), and testing (20%) sets. Performance metrics included sensitivity, AUC (Area Under the Curve), and specificity.

Two optimized models were developed: one XGBoost classifier with 10 genes and another with 74 genes. The 10-gene model, due to its simplicity, was proposed for clinical implementation, while the 74-gene model showed superior performance in specificity, AUC, and precision. Specifically, the 10-gene model achieved a sensitivity of 0.8837, a specificity of 0.3846, and an AUC of 0.6091; the 74-gene model achieved a sensitivity of 0.8837, a specificity of 0.5897, and an AUC of 0.7038. Overall, XGBoost outperformed other classifiers in terms of performance.

However, classifying cancer types based on gene expression levels is a complex task that involves extensive data processing and modeling steps. There are several areas in this study that could benefit from further improvement. For example, in the feature selection process, methods such as hyper-Lasso, which incorporates L1 regularization in regression to automatically select important features, could be explored. Additionally, tuning classifier hyperparameters through grid search or Bayesian optimization may identify the optimal parameter combinations. Finally, the model's generalizability could be validated using more external datasets.

The future of ML in cancer prediction is bright, with significant potential to transform cancer diagnosis, prognosis, and treatment. Continuous advancements in data integration, algorithm development, and clinical implementation will drive progress, ultimately improving patient outcomes and advancing personalized medicine. Collaboration across disciplines and careful consideration of ethical, regulatory, and technical challenges will be key to realizing the full potential of ML in cancer prediction.

4. Methods

4.1. Xgboost

XGBoost is a highly scalable machine learning system proposed by Tianqi Chen. It is widely used in various machine learning tasks and has achieved nearly perfect results. The advantages of XGBoost include performing second-order Taylor expansion on the loss function to increase accuracy, adding a regularization term to the objective function to prevent overfitting, and effectively handling missing data. Additionally, it supports parallel computing, which improves training speed. Its objective function is shown in Equation (1):

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \Omega\left(f_t\right) \tag{1}$$

where $l\left(y_i, \hat{y}_i^{(t)}\right)$ is the loss function; n is the total number of samples; $\hat{y}_i^{(t)}$ is the predicted value of the ith sample; y_i is the true value of the ith sample; t is the number of iterations; and $\Omega\left(f_t\right)$ is the regularization term, representing the complexity of the tree model. Different loss functions are chosen for different tasks.

The formula $\Omega\left(f_t\right)$ is shown in Equation (2), where T is the number of leaves; γ is the shrinkage coefficient; λ is the L2 norm coefficient; and ω is the score of the leaf nodes.

$$\Omega\left(f_t\right) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \tag{2}$$

4.2. SVM

SVM is a type of generalized linear classifier that performs binary classification of data in a supervised learning manner. The core idea is to find the optimal classification hyperplane that separates two classes of samples. When the samples are linearly separable, SVM finds the optimal classification hyperplane in the original space. For linearly inseparable samples, it first uses a kernel function to transform the samples from a low-dimensional space to a high-dimensional space, where it can then find the optimal classification hyperplane in this feature space. In the sample space, the hyperplane can be represented as shown in Equation (3):

$$\omega^{\mathrm{T}}\boldsymbol{x}+b=0 \tag{3}$$

In Equation (3), ω is the normal vector, b is the bias term, and $\boldsymbol{x}$ represents any point in the space. The objective function is defined as follows Equation (4).

$$\begin{cases} \min \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t. } y_i\left(\omega^{\mathrm{T}}\boldsymbol{x}_i+b\right)\geq 1-\xi_i,\, \xi_i\geq 0,\, i=1,2,\cdots,n \end{cases} \tag{4}$$

where n is the number of test samples; C is the penalty parameter ($C>0$); ξ_i is the slack variable ($\xi_i=\max\left(0,1-y_i\left(\omega^{\mathrm{T}}\boldsymbol{x}_i+b\right)\right)$), which is the hinge loss function; both ω and b need to be obtained through model training; $\boldsymbol{x}_i$ is the ith training sample; and y_i is the class corresponding to the ith sample.

Using the Lagrange multiplier method, the constrained objective function can be transformed into an unconstrained one, and the kernel function can be used to convert the nonlinear classification problem into a linear classification problem in some feature space, as shown in Equation (5). Here, α_i and α_j are the Lagrange multipliers corresponding to the ith and jth samples in the objective function, and $\boldsymbol{\kappa}$ is the kernel function.

$$\begin{cases} \min \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{\kappa}\left(x_i,x_j\right)-\sum_{i=1}^{n}\alpha_i \\ \text{s.t. } \sum_{i=1}^{n}\alpha_i y_i=0,\, 0\leq\alpha_i\leq C,\, i=1,2,\cdots,n \end{cases} \tag{5}$$

4.3. Random Forest

Random forest is an ensemble algorithm that enhances Bagging by introducing random attribute selection in decision tree training. A decision tree, a common machine learning method, classifies events by analyzing and inferring from data. It consists of a root node, leaf nodes, and internal nodes. The root represents all samples, leaf nodes show classification results, and internal nodes test attributes. The goal of decision tree learning is to produce a tree with strong generalization.

Unlike traditional decision trees that select the best attribute from all at a node, random forests randomly choose a subset of attributes and then select the best from this subset. A random forest consists of many decision trees, each constructed from different sample subsets using the bootstrap resampling method. The final prediction is determined by aggregating the predictions of these trees. This algorithm is simple and improves generalization performance by adding both sample and attribute perturbation.

Data Availability

The dataset analyzed during the current study are available in the Cancer Genome Atlas (TCGA) public data portal.

Conflicts of Interest

The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] Ning, J., Ge, T., Jiang, M., Jia, K., Wang, L., Li, W., *et al.* (2021) Early Diagnosis of Lung Cancer: Which Is the Optimal Choice? *Aging*, **13**, 6214-6227. https://doi.org/10.18632/aging.202504

[2] Chen, X., Guo, J. and Guo, M. (2020) Advances in the Treatment of Advanced Non-Small Cell Lung Cancer. *Advances in Lung Cancer*, **9**, 30-40. https://doi.org/10.4236/alc.2020.92004

[3] Field, J.K., *et al.* (2013) Screening and Early Detection of Lung Cancer. *Cancer Prevention Research*, **6**, 4-7. https://cancerpreventionresearch.aacrjournals.org/content/6/1/4

[4] Ettinger, D.S., *et al.* (2017) NCCN Guidelines Insights: Non–Small Cell Lung Cancer, Version 4.2016. *Journal of the National Comprehensive Cancer Network*, **15**, 504-535. https://jnccn.org/view/journals/jnccn/15/4/article-p504.xml

[5] Hirsch, F.R., *et al.* (2017) Lung Cancer: Current Therapies and New Targeted Treatments. *The Lancet*, **389**, 299-311. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)30958-8/fulltext

[6] Goldstraw, P., *et al.* (2016) The IASLC Lung Cancer Staging Project: Proposals for the Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. *Journal of Thoracic Oncology*, **11**, 39-51. https://www.jto.org/article/S1556-0864(15)33615-1/fulltext

[7] Shedden, K., *et al.* (2008) Gene Expression-Based Survival Prediction in Lung Adenocarcinoma: A Multi-Site, Blinded Validation Study. *Nature Medicine*, **14**, 822-827. https://www.nature.com/articles/nm.1790

[8] Kratz, J.R., *et al.* (2012) A Practical Molecular Assay to Predict Survival in Resected Non-Squamous, Non-Small-Cell Lung Cancer: Development and International Validation Studies. *The Lancet*, **379**, 823-832. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(11)61941-7/fulltext

[9] Chen, H.-Y., Yu, S.-L., Chen, C.-H., Chang, G.-C., Chen, C.-Y., Yuan, A., *et al.* (2007) A Five-Gene Signature and Clinical Outcome in Non-Small-Cell Lung Cancer. *New England Journal of Medicine*, **356**, 11-20. https://www.nejm.org/doi/full/10.1056/NEJMoa060096

[10] Roepman, P., *et al.* (2009) An Immune Response Enriched 72-Gene Prognostic Profile for Early-Stage Non-Small-Cell Lung Cancer. *Clinical Cancer Research*, **15**, 284-290. https://clincancerres.aacrjournals.org/content/15/1/284