Gene expression of non-small cell lung cancer (RNAseq) classification, biomarker identification using feature selection embedded tree-based model and interpretability

Franklin V. M. Nunes[1]

[1]Laboratory of Structural Bioinformatics and Computational Biology, Institute of Informatics, Federal University of Rio Grande do Sul, Brazil

Author Correspondence: franklin.nunes@inf.ufrgs.br

Keywords: classification, ensemble, explainable, gene-expression, feature-selection, machine learning.

Abstract

High-throughput RNA sequencing (RNA-Seq) has emerged as a transformative tool in biological research, enabling detailed investigation into gene expression patterns and the molecular mechanisms underlying various diseases, including non-small cell lung cancer (NSCLC). Given that NSCLC represents approximately 85% of all lung cancer cases and continues to pose significant treatment challenges, understanding its pathogenesis through transcriptomics is critical. This study leverages RNA-Seq data to identify key gene expression signatures associated with NSCLC while employing advanced machine learning (ML) techniques for classification tasks. Specifically, we focus on tree-based models—such as Decision Trees, Random Forests, and LightGBM—to enhance biomarker discovery efforts. These models are favored for their interpretability and ability to handle high-dimensional data, thereby revealing the most influential features in classifying NSCLC patients. To further elucidate model predictions, we incorporate Local Interpretable Model-agnostic Explanations (LIME), providing insights into the local decision-making processes of the best-performing model. The findings from this study aim to identify potential biomarkers that can contribute to improved diagnosis and treatment strategies for NSCLC, facilitating a deeper understanding of the disease's molecular landscape.

1 Introduction

High-throughput RNA sequencing (RNA-Seq) has revolutionized biological research, offering a powerful tool for the quantitative measurement of transcription, one of the most dynamic processes within cells. As a technology, RNA-Seq stands at the forefront of transcriptomics, the field dedicated to studying the complete set of RNA transcripts (both coding and non-coding) expressed by an organism. By providing a comprehensive snapshot of gene expression, RNA-Seq enables researchers to delve deeply into the molecular mechanisms underlying a wide array of biological functions and diseases. The transcriptome is a crucial link between the genetic blueprint encoded in DNA and the functional proteins that drive cellular processes (Postel et al., 2022). Through RNA-Seq, the expression levels of genes can be accurately quantified, allowing for the identification of gene expression patterns associated with various physiological states and diseases. In particular, the ability to measure and compare gene expression across different conditions makes RNA-Seq an invaluable tool in medical science and cancer research (Williams et al., 2018).

Non-small cell lung cancer (NSCLC) is the most common type of lung cancer, accounting for approximately 85% of all lung cancer cases (Zappa & Mousa, 2016). Despite advances in treatment, the prognosis for NSCLC patients remains poor, especially in advanced stages, underscoring the need for a better understanding of the molecular mechanisms driving this disease. Recent studies have leveraged RNA-Seq to identify key gene expression signatures in NSCLC, offering new insights into its pathogenesis and potential therapeutic targets (Sultana et al., 2023).

Machine learning (ML) has become an indispensable tool in various fields, including bioinformatics, where it plays a crucial role in analyzing complex biological data. ML can be broadly categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning (Pugliese et al., 2021). Each of these types serves distinct purposes, with applications that range from pattern recognition to decision-making. Supervised learning, the most widely used form of ML, involves training a model on a labeled dataset, where the input features are paired with corresponding output labels. The goal is to learn a mapping function that can accurately predict the labels of unseen data. Supervised learning is typically divided into two main tasks: classification and regression. Classification involves predicting discrete labels, such as disease presence or absence, while regression deals with continuous output values, like predicting the progression of a disease (Géron, 2022). In bioinformatics, supervised classification is often employed for disease diagnosis, patient stratification, and biomarker discovery. Biomarkers are measurable indicators of a biological state or condition, and their identification is critical for early diagnosis, treatment planning, and understanding of disease mechanisms. Feature selection methods play a key role in this process by identifying the most relevant features (genes, proteins, etc.) that contribute to the classification task (Torres & Judson-Torres, 2019). These methods can be broadly classified into three categories: filter, wrapper, and embedded methods. Filter methods evaluate the relevance of features based on statistical measures, independent of any ML model. Wrapper methods, on the other hand, evaluate subsets of features by training and testing an ML model, making them more computationally expensive. Embedded methods combine the benefits of both by performing feature selection during the model training process (Stańczyk, 2015). Among embedded methods, those based on tree-based models, are particularly powerful due to their ability to handle complex, high-dimensional data while also providing insights into feature importance.

Tree-based models, such as Decision Trees, Random Forests, and LightGBM, have gained widespread popularity due to their robustness, versatility, and interpretability. A Decision Tree is a simple, intuitive model that makes decisions by splitting data into branches based on feature values (Sun & Hu, 2017). Random Forest is an ensemble of Decision Trees that improve predictive accuracy and generalization by averaging the predictions of multiple trees (Breiman, 2001). LightGBM is a gradient-boosting framework that uses tree-based learning algorithms. It's an ensemble method that combines multiple weak models to create a predictive model (Hajihosseinlou, Maghsoudi e Ghezelbash, 2023). Interpretability is a critical aspect of tree-based models, particularly in fields like bioinformatics, where understanding the underlying biological processes is as important as making accurate predictions (Zhao et al., 2022). Tree-based models inherently provide a measure of feature importance, allowing researchers to identify which features are most influential in the classification task. This capability is particularly valuable in biomarker discovery, where the goal is to identify features that are not only predictive but also biologically meaningful.

In this study, we employ supervised classification techniques, with a focus on tree-based models, to identify potential biomarkers for non-small cell lung cancer (NSCLC). By leveraging the interpretability of the model with the best metrics and the efficiency of the embedded feature

selection method. Additionally, we use LIME (Local Interpretable Model-agnostic Explanations) to enhance our understanding of the model's predictions at a local level.

2 Materials and methods

2.1. Omics Data Obtainment

The samples of the GSE81089 (available on Gene Expression Omnibus) dataset were collected from fresh tumor tissues of 199 patients diagnosed with non-small cell lung cancer (NSCLC) and paired 18 normal lung tissues during surgical procedures performed at Uppsala University Hospital, Sweden, between 2006 and 2010. After collection, total RNA was extracted from each sample, and RNA quality and integrity assessments were conducted to ensure data reliability. The RNAseq libraries were sequenced on the Illumina HiSeq 2500 platform, generating millions of short sequences representative of the transcripts present in each sample. This approach enables a comprehensive characterization of the gene expression profile, including the identification of cancer-testis antigens (CTAs) in NSCLC samples, as well as comparison with the normal transcriptome of different organs.

2.2. Data preprocessing

The dataset was normalized using the MinMaxScaler, which scales the data features to a binary range (0-1). This step ensures that all features contribute equally to the analysis by removing the bias caused by different feature scales. Next, dimensionality reduction was performed using the SelectKBest method with the f_classif scoring function. The SelectKBest algorithm ranks all features by their ANOVA F-value and selects the top k features that contribute most significantly to the target variable. Originally, the dataset had 39975 features, for this analysis, k was set to half the total number of features, effectively reducing the dataset's dimensionality by 50% (19987 features). This approach aims to retain the most informative features, enhancing the performance and interpretability of subsequent machine learning models.

2.3. Machine Learning Approach

For each model, hyperparameter optimization was performed using Bayesian Optimization, a probabilistic model-based approach that efficiently searches the hyperparameter space to find the best combination of parameters. The hyperparameter spaces explored during optimization were specific to each model:

- LightGBM: The hyperparameters tuned included max_depth (sampled uniformly between 2 and 20), learning_rate (sampled log-uniformly between exp(-5) and exp(-2)), and subsample (sampled uniformly between 0.5 and 1).
- Random Forest: The hyperparameters included n_estimators (number of trees, sampled uniformly between 50 and 500), max_depth (sampled uniformly between 2 and 20), min_samples_split (sampled uniformly between 2 and 20), min_samples_leaf (sampled uniformly between 1 and 20), and bootstrap (a choice between using bootstrapping or not).
- Decision Tree: The hyperparameter space included max_depth (sampled uniformly between 2 and 20), allowing the model to determine the optimal depth for each decision tree.

Each model was trained on the preprocessed dataset using the optimized hyperparameters determined through Bayesian Optimization.

2.4. Cross Validation

The training process employed leave-one-out cross-validation (LOOCV), where each sample serves as a test case once. To evaluate model performance, the metrics of accuracy, F1-score, precision, recall, ROC AUC, sensitivity, and specificity were calculated. Each model was fitted 11 times, and the mean of these metrics was calculated to measure the model's overall performance. The results were then analyzed to determine the model with the best overall performance.

2.5. Embedded feature selection and interpretability

The model with the best performance metrics was then used for feature selection through an embedded method. The top 10 features identified by this model were further reviewed in the literature to explore the relevance of the selected genes. Additionally, The dataset was divided into 80% training and 20% testing subsets. For model interpretability, the 40 tumor samples from the test dataset were analyzed using LIME, which illustrates the contribution of each feature to the prediction for these samples. A comparison of the features selected from LIME with the top 10 features identified through Random Forest was realized, and posterior association with NSCLC.

2.6. Literature Review

The Ensembl database was used to retrieve the official gene names corresponding to the 10 selected features by Random Forest feature_importances_ and LIME local sample explainer. This step involved mapping the gene identifiers (Ensembl IDs) used in the RNA-seq dataset to their corresponding gene names to ensure accurate identification of the genes. Products, protein products, and biological functions present on the tables were collected from UniProt (UniProt Consortium, 2023).

A comprehensive literature review was conducted to assess the association of the selected genes with cancer. This involved searching major scientific databases such as Scopus and PubMed. For each gene, specific search queries were formulated using combinations of the gene name and keywords: "cancer," "oncogene," "NSCLC," and "lung cancer."

2.7. Libraries

The proposed approach was implemented in Python 3 using Scikit-Learn (Pedregosa et al., 2012). The LightGBM model was implemented using the lightgbm library (Shi et al., 2024). The model’s hyperparameter tuning was realized using the Hyperopt (Bergstra, Yamins and Cox, 2013) library. LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro, Singh and Guestrin, 2016) library was used to explain local predictions.

3 Results

3.1. Study Design

The study design comprises eight steps, depicted in a flowchart in Figure 1. First, the GSE81089 dataset was preprocessed. The classes (tumor and normal) are converted to a binary format and then normalized to ensure that the features are on the same scale. Then, a filter-based feature selection method is applied to reduce the dimensionality of the dataset by selecting the most relevant genes (features) that contribute to class differentiation. After creating the models with

hyperparameters optimization, they are evaluated using leave-one-out cross-validation (LOOCV) to calculate performance metrics. The model with the best metrics is selected for further analysis.

The selected model undergoes an embedded feature selection process to identify the top 10 most important genes. These genes are further investigated through a literature review to assess their association with cancer. LIME was used to explain the contribution of the features in the context of tumor samples of the test dataset, providing insights into how these genes contribute to the model's predictions and if they are shared with genes selected by random forest.

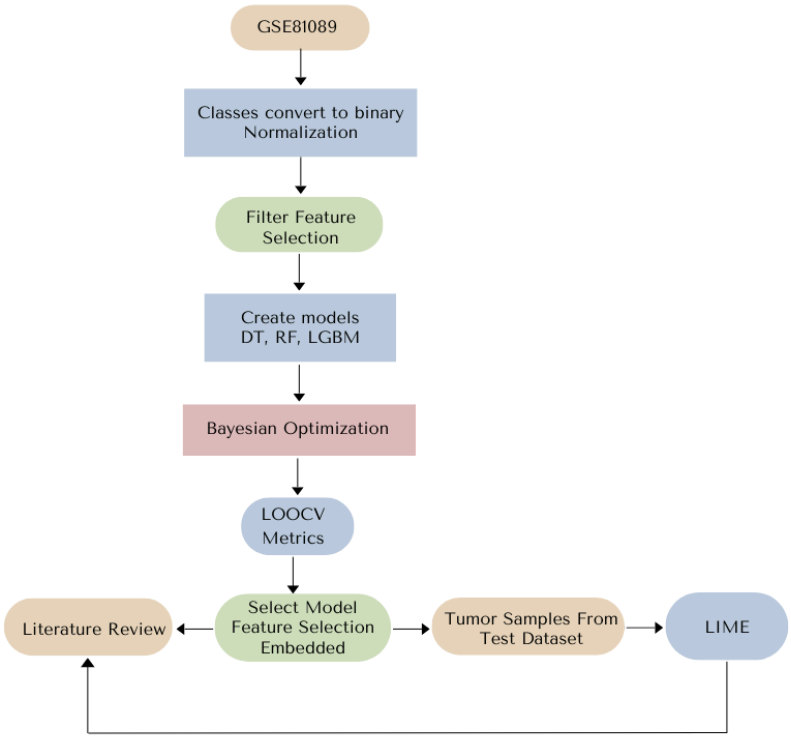

Figure 1. A Flowchart description of the study design.

3.2. Metrics

The results of the model evaluations are summarized in Table 1. The table compares the performance of three machine learning models (Decision Tree, LightGBM, and Random Forest) across seven metrics: accuracy, F1-score, precision, recall, ROC AUC, sensitivity, and specificity. The two models that had the best metrics were chosen for the confusion matrix plot (Fig. 2).

The Random Forest algorithm consistently outperformed the other models across all metrics, demonstrating its superior ability to classify both positive and negative instances accurately. Its high accuracy, F1-score, precision, recall, and ROC AUC indicate that it not only classifies correctly but also effectively differentiates between classes with minimal errors. The high specificity of Random Forest suggests it is also effective at identifying true negatives, which is crucial for applications where avoiding false positives is important. LightGBM, while performing well, did not match Random Forest's performance but was better than the Decision Tree in terms of precision, recall, and ROC AUC. It achieved a slightly lower specificity compared to Random Forest, which might

indicate more false positives relative to true negatives. The Decision Tree, although simpler and easier to interpret, lagged behind both Random Forest and LightGBM in all metrics. It had the lowest precision, recall, and specificity, suggesting it may not be as reliable for this classification task, especially when precision and recall are critical.

	Decision Tree	LightGBM	Random Forest
Accuracy	0.975813	0.977064	0.994996
F1-score	0.986753	0.987531	0.997265
Precision	0.986768	0.980198	0.994998
Recall	0.986752	0.994975	0.999543
ROC AUC	0.923998	0.957683	0.999711
Sensitivity	0.986752	0.994975	0.999543
Specificity	0.861244	0.789474	0.947368

Table 1. Mean of metrics obtained with Leave-one-out cross-validation on the GSE81089 data.

In the confusion matrix (Fig. 2), random forest correctly classified 199 tumor samples as tumor (true positives) and 18 normal samples as normal (true negatives). It misclassified only one normal sample as a tumor (false positive) and had no false negatives, making it the most reliable model for this task. The LightGBM model correctly identified 198 tumor samples and 15 normal samples. However, it misclassified four normal samples as tumors (false positives) and one tumor sample as normal (false negatives).

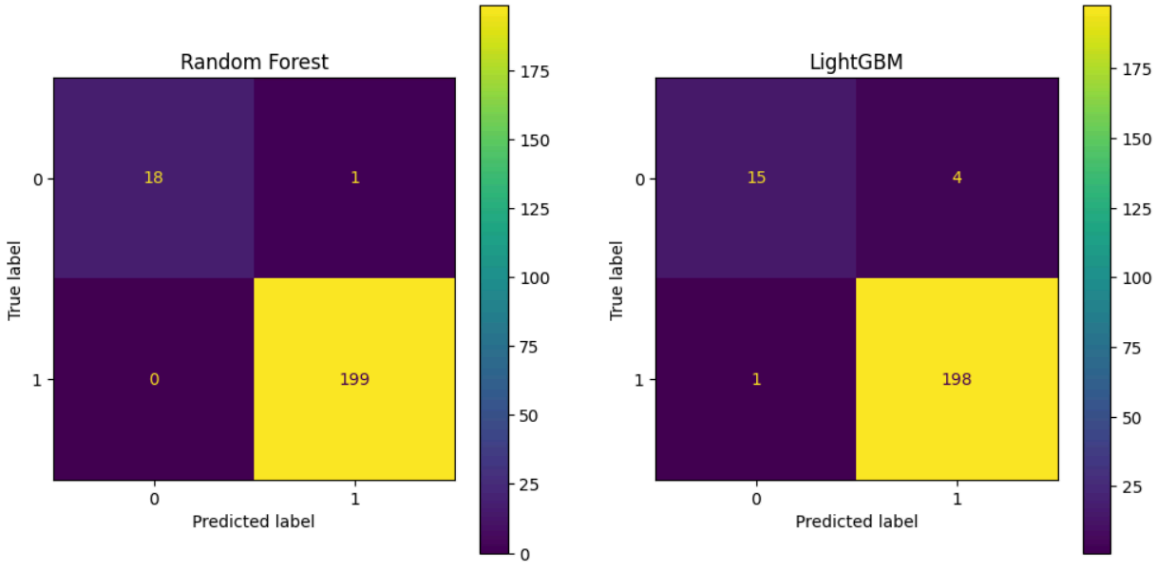

Figure 2. Confusion Matrix of the two best-performed models.

3.3. Feature Selection and Functional Enrichment

Table 2 describes the top 10 genes identified through feature_importance_ using the Random Forest Classifier model. Each gene is associated with its Ensembl Gene ID, Gene Symbol, Protein Product, Biological Function, and importance score as determined by the model.

Ensembl Gene ID	Gene Symbol	Protein Product	Biological Function	Importance
ENSG00000118526	TCF21	Transcription factor 21	Epithelial-mesenchymal interactions in kidney and lung morphogenesis include epithelial differentiation and branching morphogenesis.	0.011364
ENSG00000129596	CDO1	Cysteine dioxygenase type 1	Catalyzes the oxidation of cysteine to cysteine sulfinic acid with addition of molecular dioxygen.	0.011358
ENSG00000169442	CD52	CAMPATH-1 antigen	May play a role in carrying and orienting carbohydrate, as well as having a more specific role.	0.010227
ENSG00000112367	FIG4	Polyphosphoinositide phosphatase	Play in a vital role in regulating both the synthesis and turnover of phosphatidylinositol 3,5-bisphosphate	0.009091
ENSG00000127329	PTPRB	Receptor-type tyrosine-protein phosphatase beta	Plays an important role in blood vessel remodeling and angiogenesis.	0.007955
ENSG00000253194	antisense to FAM184A	-	lncRNA	0.007955
ENSG00000196616	ADH1B	All-trans-retinol dehydrogenase [NAD(+)]	Catalyzes the NAD-dependent oxidation of all-trans-retinol and may participate in retinoid metabolism	0.007955
ENSG00000096063	SRPK1	SRSF protein kinase 1	Play in a regulatory network for splicing, controlling the intranuclear distribution of splicing factors in interphase cells and the reorganization of nuclear speckles during mitosis.	0.006818
ENSG00000091106	NLRC4	NLR family CARD domain-containing protein 4	Key component of inflammasomes that indirectly senses specific proteins from pathogenic bacteria and fungi and responds by assembling an inflammasome complex that promotes caspase-1 activation, cytokine production and macrophage pyroptosis	0.005691
ENSG00000078081	LAMP3	Lysosome-associated membrane glycoprotein 3	Plays a role in the unfolded protein response (UPR) that contributes to protein degradation and cell survival during proteasomal dysfunction	0.005690

Table 2. Functional enrichment of the top 10 selected features from Random Forest.

The analysis of the top 10 features identified through Random Forest feature selection revealed a diverse set of genes that may play crucial roles in non-small cell lung cancer (NSCLC). Each gene was evaluated based on its biological function, protein product, and importance score within the model, providing insights into their potential contributions to NSCLC pathology. Histograms of the distribution of expression levels are described on Figure 3. In summary, the genes

identified are involved in a wide array of biological functions, including transcription regulation, metabolism, immune response, and cell signaling. These genes varying importance scores reflect their potential roles in the complex biology of NSCLC, with some playing central roles in processes like angiogenesis, immune modulation, and cellular metabolism.

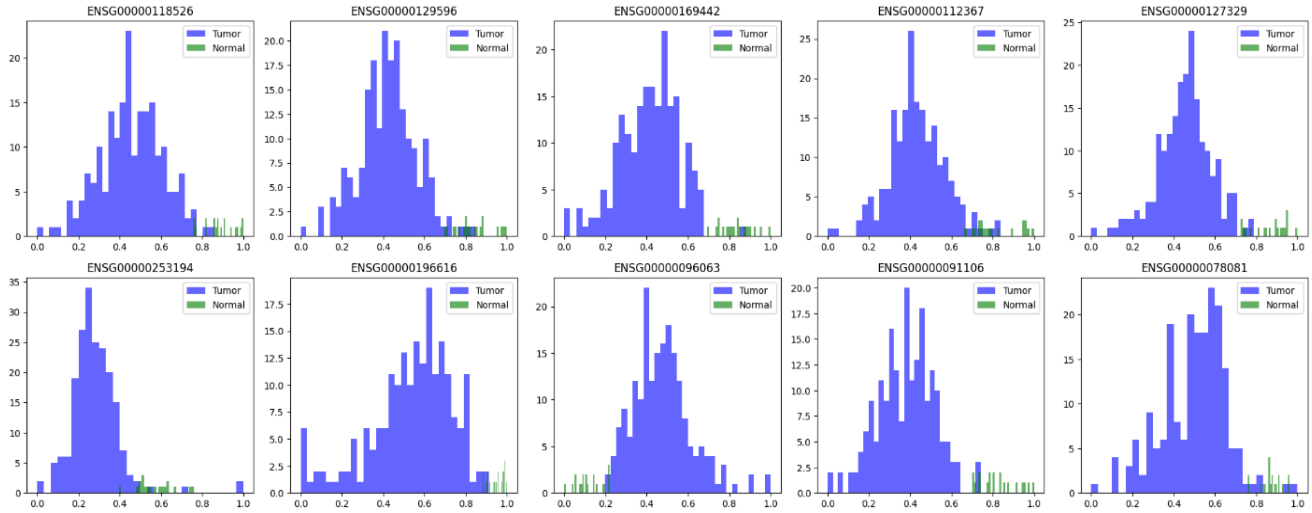

Figure 3. Distribution of expression levels for the top 10 genes selected by the Random Forest Embedded feature selection. Each plot compares the expression levels between tumor and normal samples.

The correlation matrix illustrates the pairwise correlations between the expression levels of the top 10 genes (Fig. 4). Strong positive correlations are observed between genes like ENSG00000118526 and ENSG00000196616 (0.88), and ENSG00000169442 and ENSG00000091106 (0.87). Negative correlations are seen between ENSG00000118526 and ENSG00000096063 (-0.58) and ENSG00000196616 and ENSG00000096063 (-0.53), suggesting these genes may have opposing roles in tumor development. The correlations indicate that many of these genes are co-expressed, potentially reflecting common regulatory mechanisms or involvement in related pathways.

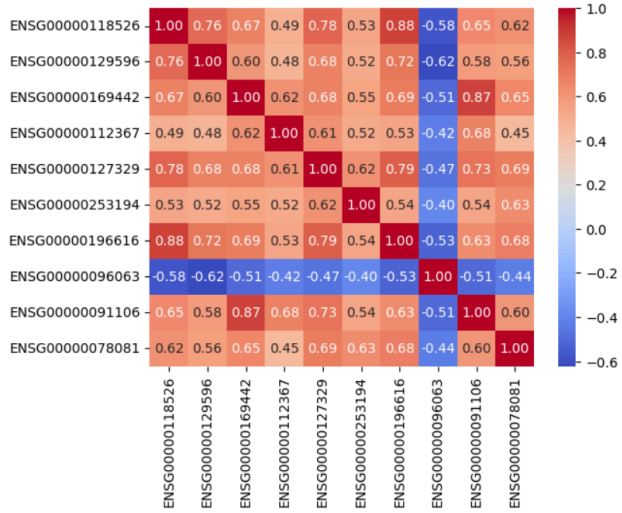

Figure 4. Correlation Matrix between the expression levels of the selected genes by Random Forest feature importance.

3.3. Interpretability

It was observed that out of the 40 tumor samples in the test dataset, there was only one feature (ENSG00000112367) that was shared between the top 10 features identified by the Random Forest feature importance and those listed by LIME for local explanations in each sample. This feature appeared in the local explanations for 3 of the 40 test dataset of the tumor samples, indicating that while the Random Forest model identified a diverse set of influential features, only this single feature consistently contributed to the model's predictions in the context of local explanations provided by LIME.

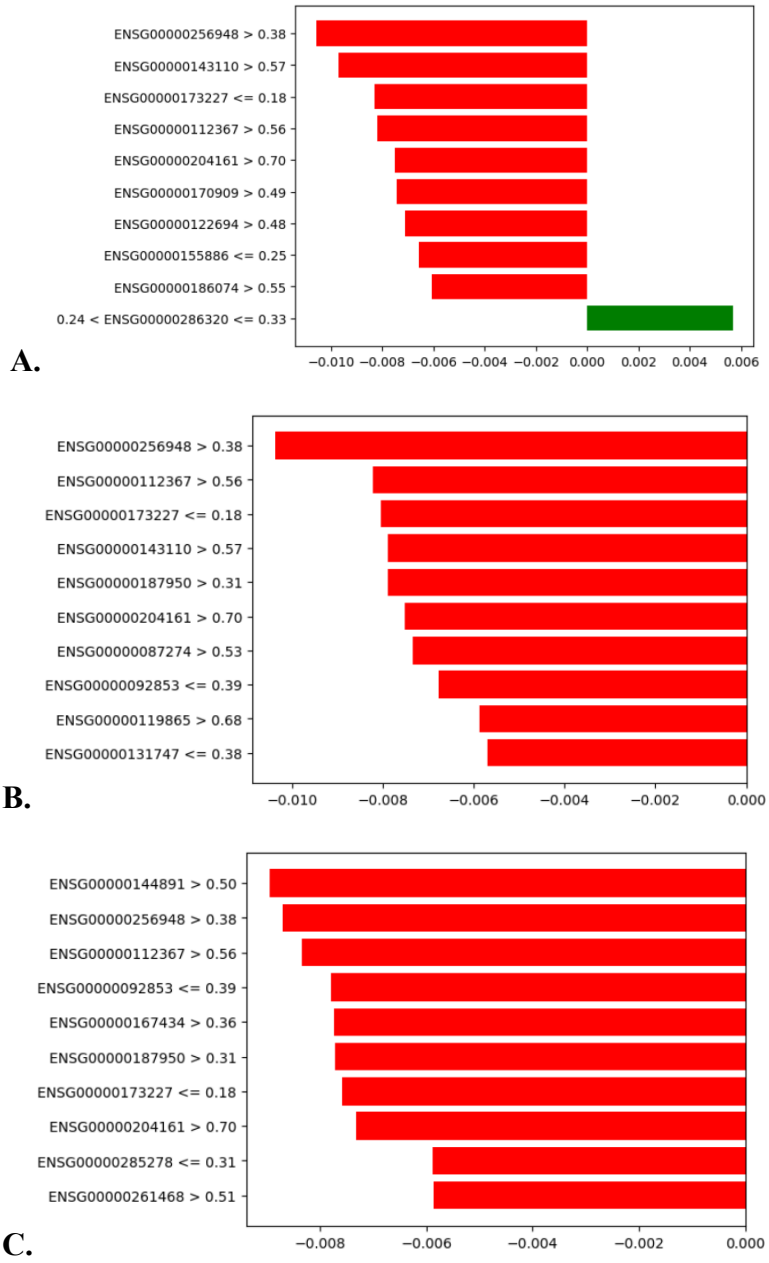

Figure 5. Local explanation of the test dataset tumor samples. A, B, and C are runs where the feature ENSG00000112367 found by random forest appears in the analysis by LIME.

Table 3 presents a detailed analysis of genes that appeared most frequently in the local explanations for tumor classification using LIME (Local Interpretable Model-agnostic Explanations). It identifies genes and their association with the prediction of tumor status. Genes such as IQSEC3 antisense RNA 2, C1orf162, TMEM273, OVCH1, AGTR1, and AGER were associated positively, indicating their frequent role in supporting tumor classification. In contrast, genes like ITM2A, antisense to MAML3, and PPAT were associated negatively, reflecting their tendency to detract from the prediction of tumors and possibly pointing to their relevance in non-tumor contexts. The table shows the functional roles of these genes, with long non-coding RNAs (LncRNAs) and transmembrane proteins or receptors being prominent.

Ensembl Gene ID	Gene Symbol	Product	Association
ENSG00000078596	ITM2A	Integral membrane protein 2A	Negative
ENSG00000286320	antisense to MAML3	LncRNA	Negative
ENSG00000256948	IQSEC3 antisense RNA 2	LncRNA	Positive
ENSG00000143110	C1orf162	Transmembrane protein C1orf162	Positive
ENSG00000128059	PPAT	Amidophosphoribosyltransferase	Negative
ENSG00000268926	-	LncRNA	Positive
ENSG00000204161	TMEM273	Transmembrane protein 273	Positive
ENSG00000187950	OVCH1	Ovochymase-1	Positive
ENSG00000144891	AGTR1	Type-1 angiotensin II receptor	Positive
ENSG00000204305	AGER	Advanced glycosylation end product-specific receptor	Positive

Table 3. Genes that appeared most frequently in the local explanations of test tumor data were analyzed with LIME. It includes their Ensembl Gene ID, gene symbol, the product they encode, and their association with tumor classification.

4 Discussion

The Random Forest model was shown as the most effective classifier in this analysis, surpassing other models in all evaluated metrics. Its superior performance is highlighted by its high accuracy, F1-score, precision, recall, and ROC AUC, demonstrating its robust ability to correctly classify both tumor and normal instances with minimal errors. The model's high specificity is particularly noteworthy, as it effectively identifies true negatives, which is crucial in minimizing false positives in cancer diagnosis. In comparison, while the LightGBM model also performed well, it did not reach the same level of accuracy as Random Forest. Although it outperformed the Decision Tree in precision, recall, and ROC AUC, its slightly lower specificity suggests a higher rate of false positives, indicating potential trade-offs between precision and recall and a greater likelihood of misclassifying normal samples as tumors.

In the study, TCF21 (Transcription Factor 21) emerged as the top-selected gene through Random Forest feature selection, highlighting its significant role in the classification of NSCLC. The expression of TCF21 is frequently downregulated in NSCLC, a phenomenon that has been

extensively documented in the literature. This downregulation is often associated with increased promoter methylation, which is a common epigenetic modification observed in the early stages of NSCLC. The high levels of TCF21 promoter methylation correlate with advanced tumor stages, increased metastatic potential, and greater invasion capacity of lung cancer cells (Chen et al., 2018). Recent studies have highlighted the significant role of CDO1 (Cysteine Dioxygenase 1) in non-small cell lung cancer (NSCLC) through its regulation by the NRF2 pathway. Reintroducing CDO1 into NSCLC cell lines restores levels comparable to those in physiological conditions observed in mouse tissues. The antiproliferative effect of CDO1 is closely linked to cysteine (CYS) levels, with reduced CYS availability or inhibited uptake leading to diminished CDO1 expression and its effects. NRF2 activation, which supports CYS accumulation, is essential for maintaining CDO1 stability. In experiments using NRF2 knockout A549 cells, CDO1 inhibited proliferation only in cells expressing NRF2, indicating its influence on cell growth through NADPH availability. The correlation between increased CDO1 expression and higher CYS levels contrasts with mRNA levels from an inducible promoter system, suggesting post-translational regulation. These findings suggest that CDO1 could serve as a potential biomarker and therapeutic target for NSCLC (Kang et al., 2019; Cai et al., 2024).

In NSCLC, PTPRB is notably down-regulated compared to adjacent normal tissues, with this decreased expression linked to poorer overall survival (Qi et al., 2016). This suggests that PTPRB may act as a tumor suppressor, where its low levels contribute to cancer progression by enhancing Src activation and promoting oncogenic signaling pathways. PTPRB's role as an independent prognostic biomarker is supported by its correlation with adverse factors like advanced tumor stage and lymph node metastasis. Understanding PTPRB's mechanisms in NSCLC could lead to targeted therapies for cases with reduced PTPRB expression, emphasizing its significance in cancer progression. The investigation into ADH1B (alcohol dehydrogenase 1B) reveals a nuanced perspective on its potential role as a prognostic marker. Survival analysis of ADH1B expression in NSCLC yielded a hazard ratio (HR) of 0.99 with a 95% confidence interval of 0.78–1.25 and a log-rank P-value of 0.912, indicating that there is no statistically significant relationship between ADH1B levels and overall survival outcomes (Wang et al., 2018). This suggests that, overall, ADH1B does not significantly impact survival in NSCLC. SRPK1, the seventh gene listed in order of importance, plays a crucial role in NSCLC prognosis and treatment resistance, with high expression levels linked to poorer progression-free survival (PFS) in patients receiving EGFR-TKIs. Overexpression of SRPK1 is associated with increased resistance to EGFR-TKI therapy, as it enhances tEGFR membrane expression. Clinical studies show that high SRPK1 levels correlate with shorter PFS and elevated tEGFR levels, positioning SRPK1 as an independent prognostic factor. Additionally, SRPK1 upregulation in gefitinib-resistant NSCLC cells suggests it as a potential target to improve therapeutic outcomes in NSCLC (Huang et al., 2023).

The NLRC4 gene plays a critical role in NSCLC by regulating apoptosis and inflammation through its CARD domain, which is essential for cell death and NF-κB signaling (Christgen, Place and Kanneganti, 2020). Notably, NLRC4 is downregulated in NSCLC, potentially affecting cancer initiation and progression by altering key signaling pathways. Its interaction with TP53 and caspase I further highlights its importance in modulating cellular processes related to the cell cycle and apoptosis. This downregulation suggests that NLRC4 could be a valuable therapeutic target or prognostic marker in NSCLC. (Välk et al., 2010). The LAMP3 gene plays a pivotal role in modulating the immune environment in NSCLC, particularly through its expression in mature dendritic cells (DCs) found within tumors. These LAMP3+ DCs are linked to tumor-infiltrating lymphocytes, suggesting their involvement in initiating and regulating anti-tumor immune responses. Their prevalence increases in patients undergoing neoadjuvant chemotherapy, highlighting their role in modulating immune responses during treatment. Additionally, LAMP3+ DCs overexpress

interleukin-15 (IL-15), enhancing the activation of key immune cells, and positioning LAMP3 as a potential target for cancer immunotherapy. (Hui et al., 2022).

The LIME local explainer used in the test dataset for tumor samples emphasizes the potential roles of long non-coding RNAs (lncRNAs) and membrane proteins in NSCLC. LncRNAs are highlighted for their potential involvement in tumor progression and cellular processes critical to malignancy (Ratti et al., 2020). The positive associations of these genes with tumor samples suggest they could serve as biomarkers or therapeutic targets, warranting further investigation into their roles in NSCLC. The FIG4 gene, identified in both Random Forest feature selection and LIME local explainer tests, seems to be more associated with non-cancer-related processes. FIG4 is known for its role as a polyphosphoinositide phosphatase, crucial in regulating lipid signaling pathways (Mironova et al., 2018), but its direct link to non-small cell lung cancer (NSCLC) remains unclear. While FIG4 was flagged as relevant in these analyses, its involvement likely pertains more to general cellular functions rather than specific cancer mechanisms, suggesting that its role in NSCLC may be limited and requires further investigation in broader contexts.

5 Conclusion

The study demonstrated the effectiveness of the tree-based models in classifying imbalanced tabular data, with emphasis in the Random Forest model in classifying NSCLC with high accuracy and specificity, surpassing other models. The identification of genes, such as TCF21, CDO1, PTPRB, and SRPK1, underscores their significant roles in NSCLC pathogenesis, providing valuable insights into potential biomarkers and therapeutic targets. While genes like FIG4 may play more general cellular roles, their relevance in cancer remains uncertain and warrants further investigation. The integration of advanced machine learning methods, such as LIME, further elucidates the contributions of specific genes and pathways in NSCLC, offering a promising avenue for enhancing cancer diagnosis and treatment. Our findings emphasize the importance of continued research into these genetic and molecular mechanisms, which could lead to more precise and effective therapeutic strategies in the fight against lung cancer.

6 Reference

Bergstra, J., Yamins, D. e Cox, D. D. (2013) “Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures”, em TProc. of the 30th International Conference on Machine Learning (ICML 2013.

Breiman, L. (2001) Machine learning, 45(1), p. 5–32. doi: 10.1023/a:1010933404324.

Cai, Y. et al. (2024) “CD52 knockdown inhibits aerobic glycolysis and malignant behavior of NSCLC cells through AKT signaling pathway”, Journal of cancer, 15(11), p. 3394–3405. doi: 10.7150/jca.86511.

Chen, B. et al. (2018) “Promoter methylation of TCF21 may repress autophagy in the progression of lung cancer”, Journal of cell communication and signaling, 12(2), p. 423–432. doi: 10.1007/s12079-017-0418-2.

Christgen, S., Place, D. E. e Kanneganti, T.-D. (2020) “Toward targeting inflammasomes: insights into their regulation and activation”, Cell research, 30(4), p. 315–327. doi: 10.1038/s41422-020-0295-8.

GEO Accession viewer Nih.gov. Available on: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81089

Géron, A. (2022) Hands-on machine learning with scikit-learn, keras, and tensorflow. 3o ed. O’Reilly Media.

Hajihosseinlou, M., Maghsoudi, A. e Ghezelbash, R. (2023) "A novel scheme for mapping of MVT-type Pb–Zn prospectivity: LightGBM, a highly efficient gradient boosting decision tree machine learning algorithm", Natural resources research, 32(6), p. 2417–2438. doi: 10.1007/s11053-023-10249-6.

Huang, J.-Q. et al. (2023) "Serine-arginine protein kinase 1 (SRPK1) promotes EGFR-TKI resistance by enhancing GSK3β Ser9 autophosphorylation independent of its kinase activity in non-small-cell lung cancer", Oncogene, 42(15), p. 1233–1246. doi: 10.1038/s41388-023-02645-2.

Hui, Z. et al. (2022) "Single-cell profiling of immune cells after neoadjuvant pembrolizumab and chemotherapy in IIIA non-small cell lung cancer (NSCLC)", Cell death & disease, 13(7), p. 607. doi: 10.1038/s41419-022-05057-4.

Kang, Y. P. et al. (2019) "Cysteine dioxygenase 1 is a metabolic liability for non-small cell lung cancer", eLife, 8. doi: 10.7554/eLife.45572.

Mironova, Y. A. et al. (2018) "Protective role of the lipid phosphatase Fig4 in the adult nervous system", Human molecular genetics, 27(14), p. 2443–2453. doi: 10.1093/hmg/ddy145.

Pugliese, R., Regondi, S. e Marini, R. (2021) "Machine learning-based approach: global trends, research directions, and regulatory standpoints", Data Science and Management, 4, p. 19–29. doi: 10.1016/j.dsm.2021.12.002.

Qi, Y., Dai, Y. e Gui, S. (2016) "Protein tyrosine phosphatase PTPRB regulates Src phosphorylation and tumour progression in NSCLC", Clinical and experimental pharmacology & physiology, 43(10), p. 1004–1012. doi: 10.1111/1440-1681.12610.

Ratti, M. et al. (2020) "MicroRNAs (miRNAs) and long non-coding RNAs (lncRNAs) as new tools for cancer therapy: First steps from bench to bedside", Targeted oncology, 15(3), p. 261–278. doi: 10.1007/s11523-020-00717-x.

Stańczyk, U. (2015) "Feature evaluation by filter, wrapper, and embedded approaches", em Studies in Computational Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 29–44.

Sun, H. e Hu, X. (2017) "Attribute selection for decision tree learning with class constraint", Chemometrics and intelligent laboratory systems: an international journal sponsored by the Chemometrics Society, 163, p. 16–23. doi: 10.1016/j.chemolab.2017.02.004.

Torres, R. e Judson-Torres, R. L. (2019) "Research techniques made simple: Feature selection for biomarker discovery", The journal of investigative dermatology, 139(10), p. 2068-2074.e1. doi: 10.1016/j.jid.2019.07.682.

UniProt Consortium (2023) "UniProt: The universal protein knowledgebase in 2023", Nucleic acids research, 51(D1), p. D523–D531. doi: 10.1093/nar/gkac1052.

Välk, K. et al. (2010) "Gene expression profiles of non-small cell lung cancer: survival prediction and new biomarkers", Oncology, 79(3–4), p. 283–292. doi: 10.1159/000322116.

Wang, P. et al. (2018) "Distinct prognostic values of alcohol dehydrogenase family members for Non-small cell lung cancer", Medical science monitor: international medical journal of experimental and clinical research, 24, p. 3578–3590. doi: 10.12659/MSM.910026.

Zhao, Y., Shao, J. e Asmann, Y. W. (2022) "Assessment and optimization of explainable machine learning models applied to transcriptomic data", Genomics, proteomics & bioinformatics, 20(5), p. 899–911. doi: 10.1016/j.gpb.2022.07.003.