A Hybrid Statistical–Machine Learning Approach Uncovers Distinct Immune and Proliferative Pathways in Lung Cancer Transcriptomes.

Submitted By
Tasnia Rahman
221-35-982
Department of Software Engineering
Daffodil International University

Supervised By
Musabbir Hasan Sammak
Lecturer (Senior Scale)
Department of Software Engineering
Daffodil International University

Thesis submitted in fulfilment of the requirements for the award of the degree of Bachelor of Science
Department of Software Engineering (Major in Data Science)

Fall 2025

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	Tasnia Rahman
Date of Birth	:	01/01/2001
Title	:	A Hybrid Statistical–Machine Learning Approach Uncovers Distinct Immune and Proliferative Pathways in Lung Cancer Transcriptomes.
Academic Session	:	Fall 2025

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☐	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

Tasnia

_____
(Student's Signature)

221-35-982
_____
Student ID
Date: 13 December 2025

_____
(Supervisor's Signature)

Musabbir Hasan Sammak
_____
Name of Supervisor
Date: 13 December 2025

APPROVAL

This thesis titled on "A Hybrid Statistical–Machine Learning Approach Uncovers Distinct Immune and Proliferative Pathways in Lung Cancer Transcriptomes", submitted by **Tasnia Rahman (ID: 221-35-982)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. S. M. Hasan Mahmud
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

A.H.M Shahariar Parvez
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Khalid Been md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Dr. Md Sazzadur Rahman
Professor
Institute of Information technology
Jahangirnagar University, Bangladesh

SUPERVISOR's DECLARATION

I hereby declare that I have checked this thesis "A Hybrid Statistical–Machine Learning Approach Uncovers Distinct Immune and Proliferative Pathways in Lung Cancer Transcriptomes" and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.

(Supervisor's Signature)

Full Name : Musabbir Hasan Sammak

Position : Lecturer (Senior Scale)

Date : 13 December 2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

Tasnia

(Student's Signature)

Full Name : Tasnia Rahman

ID Number : 221-35-982

Date : 13 December 2025

A Hybrid Statistical–Machine Learning Approach Uncovers Distinct Immune and Proliferative Pathways in Lung Cancer Transcriptomes

TASNIA RAHMAN

THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE OF

BACHELOR OF SCIENCE

DEPARTMENT OF SOFTWARE ENGINEERING (MAJOR IN DATA SCIENCE)

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor for their continuous support, guidance, and valuable insights throughout this research. I also thank the faculty members and the Department of Software Engineering at Daffodil International University for providing the academic environment and resources necessary to complete this work. My appreciation extends to my friends and colleagues for their encouragement, and to my family for their patience, strength, and unwavering support during this journey. Without their contributions, this thesis would not have been possible.

Thank you all.

DEDICATION

This thesis, titled "A Hybrid Statistical–Machine Learning Approach Uncovers Distinct Immune and Proliferative Pathways in Lung Cancer Transcriptomes" is dedicated to the people whose unwavering love, support, and inspiration have shaped my journey. To my beloved parents, whose sacrifices, prayers, and constant encouragement have been my greatest strength. Everything I achieve is because of your endless support and belief in me. To my family, who stood beside me in every challenge with patience and motivation. To my teachers and mentors, whose guidance and dedication inspired me to pursue knowledge and excellence. And finally, to all cancer patients and researchers working tirelessly toward better diagnosis and treatment may this work contribute, even in a small way, to a future of improved healthcare and hope.

ABSTRACT

Lung cancer is the greatest cause of death in the world among all types of cancer. It is brought about by complicated genetic variations, variation in transcription and immune system interactions. RNA-seq data (N= 739 samples and 59,429 genes) and machine-learning algorithms (Random Forest, LightGBM and Elastic Net) and statistical methods of differential expression (limmavoom and DESeq2) assist us to identify strong molecular signatures. Important pathways within tumors which regulate the cell cycle, copy DNA, and preserves good shape of chromosomes were found through statistical research. Meanwhile, the machine learning methods identified non-linear immune responses such as like T-cell stimulation, cytokine signaling, and antigen presentation which are frequently missed in the fold-change based tests. The four complementary sets of genes have been obtained through the attempts to merge the two strategies (Stat-only, ML-only, Common, Union), and each shows specific enhancements in the pathway but adds architectural power to identifying biomarkers, analyzing the pathways, and the prospective tool in the field of precision medicine.

KEYWORDS: Hybrid Computational Analysis; RNA-seq; Differential Gene Expression; limma-voom; DESeq2; Machine Learning; Random Forest; LightGBM; Elastic Net; Tumor Microenvironment; Immune Pathways; Cell-Cycle Regulation; High-Dimensional Transcriptomics; Biomarker Discovery; Lung Cancer; NSCLC; Pathway Enrichment; GO/KEGG/Reactome; Statistical–ML Integration; Precision Medicine.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LST OF SYMBOLS

LIST OF ABBREVIATIONS

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Lung cancer is one of the major forms of cancer that occur among people around the world and it continues to claim the lives of a significant number of cancer patients. The current cancer statistics in the world indicate that lung cancer leads to over 1.8 million cancer deaths annually not to mention it causes over 2.2 million novel instances of the disease. That is, over one out of five cancer deaths worldwide is as a result of lung cancer [1]. The 5-year premature mortality has not increased much despite improvements in the immunotherapies, targeted drugs and even surgical procedures. This is most especially in the case of people who are in the advanced stages of the disease. Due to the aggressiveness of lung cancers and the tendency to be diagnosed too late as well as the high level of molecular variability we should have improved genome analysis and biomarker discovery to aid in their early discovery and the correct type of treatment [3].

The occurrence of changes in genes, transcription problems, interactions of the tumor with the environment, and mechanisms through which the immune system is evaded to attack the cancer cause lung cancer. The oncogene pathways are EGFR, KRAS, ALK, PI3K/AKT, and MAPK and they assist the cancer to grow, survive and then metastasize [4]. The tumor progression and resistance to treatment are also caused by dysregulation of DNA-damage response genes, chromatin remodeling factors, and metabolic reprogramming [5]. The recent multi-omics and single-cell technologies have demonstrated that lung tumors contain many different cell types, including malignant epithelial cells, immune cells that have infiltrated the tumor microenvironment (TME) [6]. The common types of lung cancer are of different biologies, clinical behaviour and management. The most prevalent and most common is in the form of the non-small cell lung cancer (NSCLC) that constitutes about 80-85% of the cases. It has also discussed several histological types, which are adenocarcinoma, squamous-cell carcinoma, as well as large-cell carcinoma. They differ in terms of their molecular patterns, patterns of spread and vulnerability to treatability, which is the matter of the molecular tests and choice of the targeted therapy that best suits [2]. The most widespread type is the small-cell lung cancer (SCLC). It is rapidly spreading, rapidly growing and the prognosis is normally poor.

They studies on clinical trials tend to suggest that SCLC is very likely to respond to the initial line of treatment based on chemotherapy and radiotherapy but that it often returns and is resistant to therapy hence necessitating the implementation of high-resolution genomic and single-cell methods in order to enhance patient stratification and offer personalized treatment approaches [2].

The inter-patient heterogeneity of non-small cell lung cancer (NSCLC) is so high that it has been the main focus of most recent studies, and is caused by a number of factors, including genetic abnormalities, exposure to the environment (including smoking), and the constituents of the tumor microenvironment (TME) [8]. The recent single-cell sequencing studies [9] suggest that conventional bulk approaches do not have the power to cover the multicellular states and immunological states of NSCLC tumours. It is this heterogeneity which is the impetus to the use of computational frameworks capable of describing both linear and non-linear molecular signatures.

The recent method gaining much popularity in studying gene expression patterns is the RNA sequencing (RNA-seq). It is so sensitive and has a large dynamic range that allows scientists to measure the transcripts with high precision [10]. RNA-seq has become critical in lung cancer studies in identifying and describing the expression differences in genes, describing molecular subtypes, and exploring pathways that cause drug resistance [10]. Recent studies have used the RNA-seq to identify conservators of lung cancer progression, including MUC1, LGR5, SYK and some immune-related feature, which can affect tumor behavior [11]. RNA-seq in combination with single-cell sequencing has also enabled the further enhancement of the capacity to investigate intratumoral heterogeneity and detect rare malignant or immune cell populations that cannot be detected by bulk tissue analysis. The advantages of this type of research make RNA-seq a fundamental instrument of the current molecular profiling and computational biomaker discovery in lung cancer studies [12].

Two statistical packages most commonly used in an RNA-seq-based differential expression analysis include limma-voom and DESeq2, which is most reliable in identifying differences in the expression between two biological groups and has a solid mathematical basis. The expression of Limma-voom models is based on the line models with precision weighting needed to consider mean-variance dependencies, and DESeq2 is based on the negative binomial models with dispersion shrinkage to explain the variability in count data. Both methods provide results that are easily interpretable, such as p-values, log2 fold-changes and FDR-adjusted significant scores, which can be used to identify genes with significant or self-enriching differences in expression between tumor and normal samples [16].

These statistical approaches are fine at identifying large changes of transcription, but not necessarily at diseases that differ, such as lung cancer .When there is a high level of immunocellular infiltrate in tumours or a high level of microenvironmental heterogeneity, the group-level changes may be diluted and mask interesting biological processes. Due to this reason, traditional analyses of differential expression can fail to capture immune-based, metabolic, or microenvironmental signals, which can by either nonlinear or context-specific mechanisms [17].

This consideration is important in terms of the machine-learning (ML) models, which can analyse the high-dimensional gene-expression profiles without being dependent on the assumption of linearity. Likely Random Forest, LightGBM, and Elastic Net are among the algorithms that can find complex patterns of interaction with features, determine weak but helpful features, deal with the correlation of sets of genes, and can use class weighting to deal with imbalanced sample distributions. Moreover, the models are interpretable through feature-importance values, coefficient values, or SHAP. Very many research works implemented machine learning with lung cancer RNA-seq data have found biomarkers related to epithelial-mesenchymal transition, immune infiltration, metabolic variations, and responses to treatment, which indicates the usefulness of machine learning in unearthing regulatory processes that traditional statistical methods may neglect [19].

Machine-learning methods are better than traditional differential expression methods at finding nonlinear correlations and higher-order interactions between genes.These measure subtle variation that leads to sample classification, and are resistant in the face of noisy or heterogeneous expression patterns, and can effectively combine many sources of biological variation including immune signals, changes in epithelial state, and microenvironmental effects. This enables ML models to indicate immune-driven pathways such as T-cell activation, cytokine signaling and antigen presentation, when such pathways fail to achieve statistical significance when tested on a group-level due to the high patient to patient variance [23].

The various but practical viewpoints are presented by statistical and machine-learning means. A combination of the two can help you to circumvent the issues with both of them. Statistical pipelines are highly effective in identifying large magnitude and consistent variation in expression, whereas the ML models can identify tiny nonlinear and contextual patterns. A combination of the two methodologies enhances easier sample of new genes, more stable and biologically relevant sets of features, easier determination of the difference between tumor-intrinsic and immune-derived signals, and easier identification of biomarkers with cross-validated interpretability. The result of this integrated framework was four gene groups Stat-only, ML-only,

Common and Union-all, which each explained distinct aspects of the biology of lung cancer. It is upon these combined results that the biological interpretations and other analysis will be discussed in the following chapters.

CHAPTER 2

LITERATURE REVIEW

2.1 Literature Review

The application of next-generation genome sequencing, single-cell technologies, multi-omics, and advanced computational modeling have contributed to the improvement of lung cancer research. This chapter overviews the principal findings of the most recent high-impact articles (2022-2025) with particular attention to literature on molecular profiling, pathways analysis, biomarker discovery, and machine-learning-based transcriptomics as the basis of this study.

Lung cancer is the number one cause of death all over the world with million of new cases being reported each year. The focus has been on the significance of early detection and molecular stratification as well as individualized therapy to enhance outcomes in recent studies. Worldwide cancers studies underscore the growing significant of transcriptome profiling and also the immunological characterization and precision medicines therapy [3].

The ease of seeing the functioning of the lung cancer has all been made much easier due to next-generation sequencing, large-scale transcriptome profiling and advanced computational frameworks. Bulk RNA sequencing (RNA-seq) is one of the technologies that are commonly used in this field because this technique gives a broad overview of the changes in gene expression in relation to malignancies. The investigation of transcriptomics with large scale RNA-seq has shown the downregulation of numerous oncogenic pathways, including probable PI3K/AKT and EGFR, in lung adenocarcinoma (LUAD), and their role in the development and progression of NSCLC [5].The uses of RNA fingerprinting in the circulation have been shown. Indicatively, deep generative models of blood-based RNA features could accurately predict early-stage lung cancer [17].Other studies on therapy resistance have presented stress-response and DNA-damage genes related to radiotherapy failure in radioresistant NSCLC models [11].All this evidence highlights the highly crucial role of bulk RNA-seq in the identification of genes including MUC1, LGR5, and SYK features associated with phenotypic transformation, microenvironmental remodelling as well as tumor growth [9].

The bulk RNA-seq provides us with some beneficial information, yet on thousands of cells, which performs averaging of gene expression thus obscuring the various cell variations. Single-cell RNA sequencing (scRNA-seq) has become a groundbreaking instrument as it can be used to examine single malignant, immunological, and stromal cells in tumors on a high-resolution scale. Recent single cell research discovered immunosuppressive myeloid niches that support NSCLC evading immune system and avoiding treatment [18]. Subsequent research has characterized malignant epithelial status and stromal relationships and provided detailed information regarding the functional initiatives that are being executed within tumor ecosystems. scRNA-seq has also shown that the immune system changes over time after therapy with a PD-1 inhibitor. This is because T-cell populations become more active in patients who respond to the medication [14]. Millions of transcriptomes of various studies are now available as large integrated single-cell atlases, making it easier to harmonize data of disparate genetic sources, in particular, within locales enriched in immune infiltration or at tumor-microenvironment interfaces [29].Collectively, this evidence shows that single-cell transcriptomics uncovers tumor heterogeneity that bulk-level methods often fail to capture, especially in regions enriched for immune infiltration or at tumor–microenvironment interfaces [29].

The route level of transcriptomic studies has uncovered the key biological pathways that are involved in lung cancer. EGFR signaling is still one of the routes that is most often changed in LUAD. This makes it a main target for numerous particular inhibitors [4]. More investigations at the route level have demonstrated how important lipid metabolism and immune phenotypes that are controlled by fatty acids are.These affect the degree of aggressiveness of tumors and immune system reaction [15]. It has been demonstrated that long non-coding RNAs such as NEAT1 and MALAT1 can modulate the immune cell activity and tumor microenvironment to contribute to tumor growth and immune evasion [24]. Meanwhile, The spatial transcriptomics has illuminated how tumour-tissue resolution influences the metastatic dissemination, particularly in the brain metastases, in which the framework of neighbouring tissue influences for tumour cell plasticity [31]. Bulk RNA-seq coupled with single cell RNA-seq and pathways enrichments have enhanced our concept of the lung cancer biology by distinguishing between alterations that occurred in the tumor and alterations that occurred in the environment. Computer-analysis of gene expression data typically involves the use of such statistical frameworks as limma-voom and DESeq2. The tools are accurate at modeling variability of expression and are quite sensitive to identifying large and consistent fold-changes in proliferative or tumor-intrinsic pathways [16]. These statistical methods, however, have weaknesses in the problems of nonlinear gene interactions, complex immune signals, heterogeneous microenvironments, high-

dimensional measurements, or imbalance issues of samples that frequently allow the oversight of high-effect but small-effect genes of biologic interest. Machine learning tools have evolved to be more helpful in analysis to manage these issues. The machine learning methods have been widely utilized in lung cancer transcriptomics to find biomarkers and to explain complex molecular patterns. The difference between the tumor samples and normal tissue can be detected using classification systems with bulk RNA-seq profiles. These are able to distinguish trends that could not be detected by normal fold-change analysis that can assist them in coming up with their predictions [10]. The lipid-immune properties of machine learning analyses have offered the opportunity to classify NSCLC subtypes and select the most effective treatment [15]. In addition, such algorithms as Random Forest, LightGBM, and Elastic Net have been successfully used to identify molecular features related to recurrence, immunological infiltration, and clinical response and yield more sophisticated studies of transcriptome changes than univariate statistical techniques [22]. Regularly featuring immune-associated pathways that statistical endeavors might reject, feature-importance methods, e.g. SHAP scoring, gain metrics, and coefficient-oriented rankings, never cease to highlight feature-importance metrics that are nonlinear and typically have moderate effect sizes [23]. New advances in pathway enrichment schemes have been able to make transcriptome data considerably less complicated to interpret. New techniques evaluate behavior of routes on a disease-network basis, which enhances the comprehension of orchestrated biological action on numerous pathways [21]. Pathway-centric modelling is becoming widespread in the multiomics and transcriptome investigations to the determine shared with biological signatures associated with the subtypes of lung cancer. Enrichment methods based on cross-talk and an example is CT pathway have provided deeper functional observations by providing explanations on how regulatory processes and signaling networks relate [19]. Such this advancements reflecting the shift from traditional single-gene analysis to the overall view of the pathway of interactions. There is strong support within modern literature for hybrid methods of analysis that combine statistical differential expression with feature selection based upon machine-learning. Biomarkers are more credible with the Network-based approaches and a difference between various datasets with the help of co-expression modules, various expressions, and machine learning features [30]. The integration of bulk and single-cell RNA sequencing in studies assists us to discover more about the important activities on the population and cell scale. This makes it slow to discover the paths [9].Consensus frameworks alleviate the limitations associated with reliance on a specific methodology, hence improving the detection of biologically relevant signals across various datasets [19]. The combined results corroborate the approach used in this thesis that combines both statistical and machine-learning frameworks to determine long-term tumor-intrinsic and complex

immunological nonlinear patterns to improve the molecular understanding of lung cancer.

CHAPTER 3

METHODOLOGY

3.1 Study Overview

This chapter are covering the data acquisitions, preprocessing, statistical modeling and machine-learning analysis, feature integration, and pathway enrichment. Each step was carefully planned to deal with the large RNA-seq collection and the clear imbalance between the normal and tumor samples. The methodological design takes the best parts of both old-fashioned statistics methods and newer machine-learning methods to get a fuller picture of how lung cancer works. Here, figure 3.1 is a diagram that shows how the whole process works.

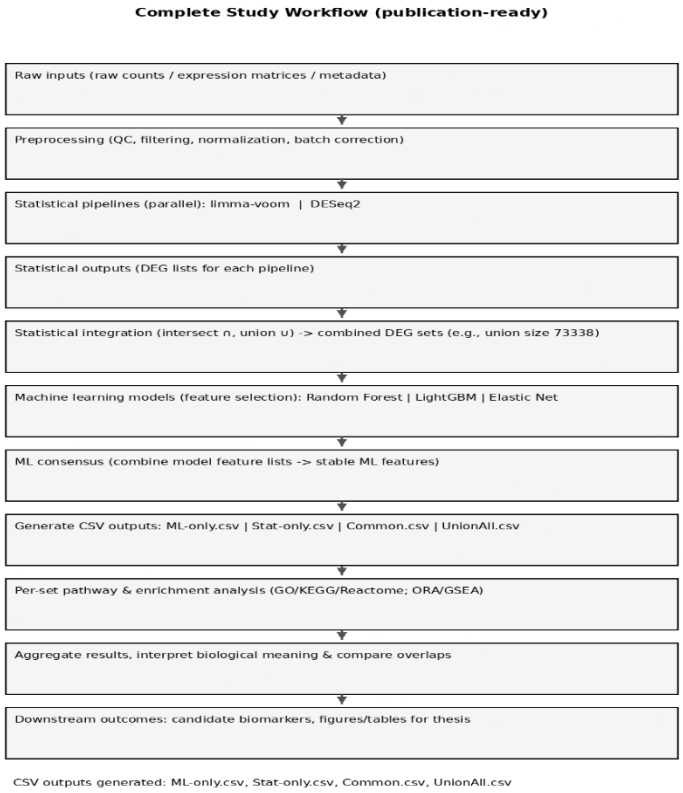


Figure 3.1.1 Overall Workflow of the Analytical Pipeline

3.2 Data Acquisition and Summary

The RNA-seq dataset used in this study consisted of 739 samples, including 702 tumor and 37 normal samples. Raw gene-expression matrices contained 59,429 genes. The accompanying phenotype information was used to distinguish tumor and normal tissues.

Table 3.2.1 Dataset Summary

Component	Description
Total samples	739
Tumor samples	702
Normal samples	37
Total genes (raw count)	59,429
Genes retained after filtering	~30,000+
HVGs selected for ML models	10,000

The highly imbalanced classes distribution required the additional preprocessing of the strategies during model development.

3.3 Preprocessing Dataset

Quality Filtering and the Gene Removal. To reducing the noise and improving the statistical models and low-count genes were removed. Genes with essentially negligible expression across samples were eliminated, reducing the feature space by nearly half while leaving more than 30,000 genes for statistical tests.

Ways to the Approach Normalization Depending on the study that came after, different normalization methods were used:

For the statistical modeling, two normalization approaches were used to keep the expression values consistent across samples. The limma-voom method applied log normalization with precision weighting and while the DESeq2 corrected library-size differences through size-factor adjustment and stabilized variance with its transformation.

For the machine-learning workflows, this log2(x+1) normalization was applied first, followed by selecting genes that showed higher variability across samples. Additional scaling and distribution adjustments helped keep all the features on comparable ranges for the models. To manage the imbalance between tumor and normal samples, several correction steps were taken. Class weights were adjusted within the models, decision thresholds were tuned to improve detection accuracy, and suitable evaluation metrics were used to maintain fairness in performance assessment. These steps were prevented the models from being biased toward the majority class.

3.4 Statistical Differential Expression Analysis

Two widely accepted RNA-seq statistical frameworks were employed to detect differentially expressed genes (DEGs): limma-voom and DESeq2.

limma-voom:
This pipeline uses linear modeling combined with precision weights derived from the voom transformation. DEGs were identified using this criteria:

i. adj.P.Val < 0.05.
ii. |log2FC| ≥ 1.

DESeq2:
DESeq2 models count data using a negative binomial distribution and applies shrinkage to dispersion and fold-change estimates. DEGs were selected based on:

i. padj < 0.05.
ii. |log2FC| ≥ 1.

DEG Summary Across Statistical Methods:
The two statistical pipelines produced overlapping as well as distinct gene sets:

Table 3.4.1 Statistical DEG Summary

Category	Count
DEGs significant in both limma & DESeq2	≈ 2,155
limma-only DEGs	≈ 1,612
DESeq2-only DEGs	≈ 3,570
Union of all statistical DEGs	7,337

The union of these statistical DEGs was used for integration with machine-learning features.

Machine-Learning Analysis:
Three machine-learning (ML) models were applied independently on 10,000 highly variable genes to identify predictive and biologically informative features: Random Forest, LightGBM, and Elastic Net. All models used weighted training to address class imbalance.

Random Forest:

The Random Forest (RF) model was applied to classify the samples and to estimate gene importance through its ensemble of decision trees. Two complementary importance measures were extracted: the native Gini-based score (Mean Decrease Gini) and SHAP-derived interpretability values. To improve stability, the results from both metrics were merged into a combined ranking. Model performance was then assessed using standard evaluation indicators, including ROC-AUC, PR-AUC, accuracy, sensitivity, and specificity.

The LightGBM:

LightGBM (LGBM) was used as a fastest and efficient gradient-boosting method suitable for the high-dimensional datasets. Feature importance was evaluated using both of the gain-based scores and the SHAP values to the capture complementary aspects of model interpretation. To address the imbalances between the sample groups and also the threshold adjustments and weighted training were applied. The most informative LGBM features were then selected for downstream integration.

Elastic Net:
Elastic Net (EN) performs penalized the logistic regression with both L1 and L2 penalties, suitable for correlated gene sets. Genes with the non-zero coefficients were considered informative. SHAP values were also calculated to the support interpretability. The final Elastic Net genes lists represented to the features consistently contributing to the prediction across the model folds.

Implementation of the machine learning and statistical properties :
One of the key features of the given study was to combine variables produced in both statistical analysis and machine learning pipelines. Four categories of genes were built:

i. Stat-only genes (1,714).
ii. ML-only genes (4,377).
iii. Common genes between Statistics ∩ ML (5,625).
iv. Union of all genes from both approaches (11,714).

This multi-level integration allowed a clearer understanding of which biological processes are captured by linear vs. non-linear methods.

Pathway Enrichment Analysis:

The analysis was carried out individually on each of the sets of genes to determine its biological relevancy. The databases that were enriched included the multiple databases such as the Gene ontology which includes Biological Process, Cellular Component, and Molecular functional, the KEGG pathways and Reactome pathways. These auxiliary resources were assisted to the kindling of the functional roles and the molecular.examples of such mechanisms are the genes.

Distinct pathway patterns emerged:

The patterns of pathways were diverse in the various groups of genes. The genes discovered due to the statistical approaches were predominantly enriched in the pathways of the cell-cycle regulations and proliferation, whereas those found due to the machine learning were enriched in the immune activation and cytokine-mediated signaling pathways. The genes found by the two methods exhibited the connections into the developmental and the metabolic processes whilst the union set represented a wider spectrum of metabolic and signaling networks. These differences highlight the complementary nature of strengths of the methods of analysis.

3.5 Final Synthesis and Analytical Rigor

The hybrid framework integrates:

Several complimentary features improved the molecular signal detection in the procedure. While robust statistical approaches detect large expression shifts, machine-learning algorithms reveal subtle non-linear interactions. These phases are supported by thorough preprocessing, balanced evaluation, physiologically informed feature selection, and pathway-level interpretability. They provide a comprehensive analytical pipeline that can detect lung cancer progression-associated molecular markers.

3.6 Ethics, computation, and reproducibility considerations

Each analysis used the open-source frameworks likely R/Bioconductor, Python, scikit-learn, and LightGBM. Cross-validation and statistical controls were used throughout modeling to reduce overfitting. Deterministic seeds and well-defined protocols ensured reproducibility of all results.

CHAPTER 4

RESULTS

4.1 Results Summary

Preprocessing, statistical differential expression, machine-learning model performance, multi-level gene integration, and pathway enrichment results are presented in this parson. Lung cancer molecular changes are better understood thanks to the combined strengths of conventional and data-driven techniques.

4.2 Normalization and Preprocessing The Results

The original collection included the 739 RNA-seq samples and the 702 are tumors and 37 are normalized with 59,429 raw gene characteristics. After removing genes with extremely low expression and applying normalization procedures, the dataset was refined to a more analytically meaningful gene set.

Table 4.2.1 Summary of Preprocessing Outputs

Step	Outcome
Raw gene count	59,429 genes
Genes removed due to low expression	~29,000+
Genes retained for statistical testing	~30,000+
HVGs (for ML models)	10,000 highly variable genes
Normalization applied	DESeq2 size-factors, limma-voom, log2(x+1), scaling
Imbalance handling	Class weights + threshold tuning

The preprocessing stabilized input distributions and corrected the extreme class imbalance.

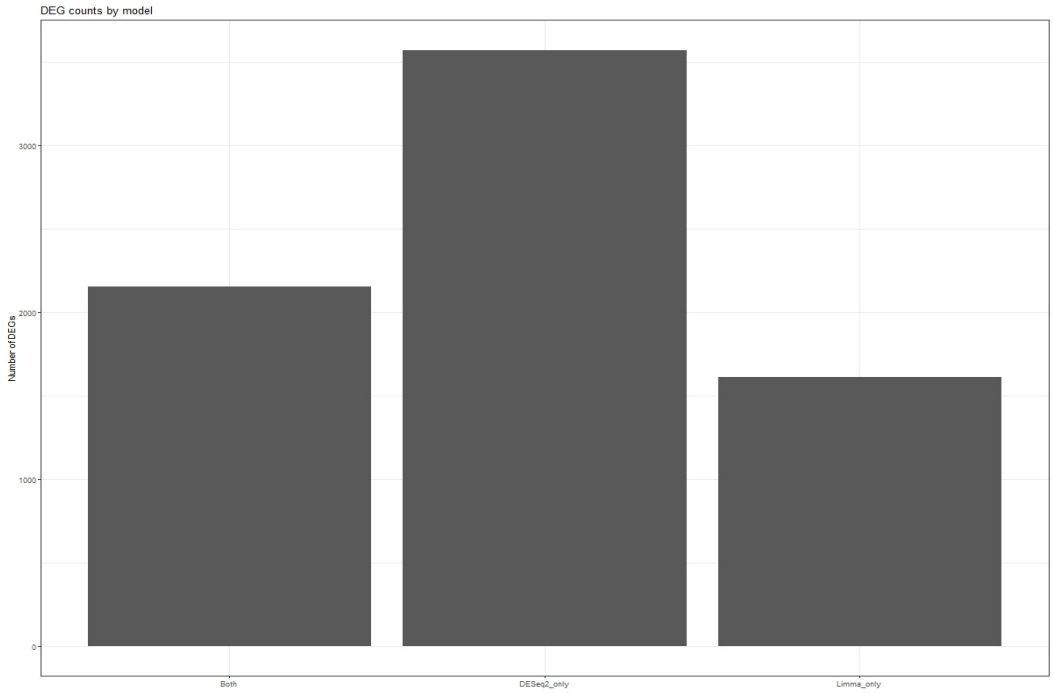
4.3 Statistical Differential Expression Analysis

Both limma-voom and DESeq2 were applied independently to detect significantly altered genes between tumor and normal tissues. Each method provided a unique yet complementary set of results.

Table 4.3.1 Differential Expression Summary

Category	Count
DEGs significant in BOTH methods	**≈ 2,155**
limma-only DEGs	**≈ 1,612**
DESeq2-only DEGs	**≈ 3,570**
Union of statistical DEGs	**7,337 genes**

These findings indicating that the statistical pipelines primarily detect the strong and consistent expression for shifts.



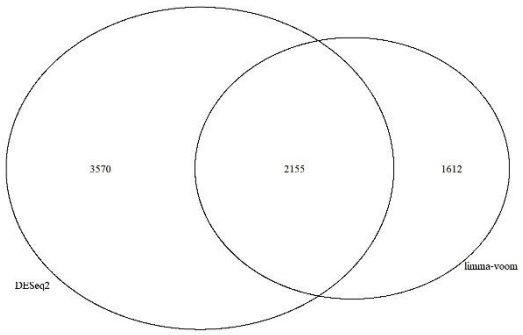


Figure 4.3.1 Bar Plots and Venn Diagram Visualization

4.4 Machine-Learning Model Performance

The three machine-learning models, which are Random Forest, LightGBM, and Elastic Net, were modelled using 10,000 high-variance genes. Weighted training and threshold adjustment compensated against the grievous imbalance in the class. Each model generated the predictive performance indicators and a list of ranked influential genes.

Random Forest Results

Random Forest produced good classification behaviors and cross-validation folds were consistent. Its feature-important conclusions were obtained by integrating Gini importance and SHAP-rankings.

Table 4.4.1 Random Forest Model Evaluation

Metric	Value
ROC-AUC	0.971
PR-AUC	0.998
Accuracy	0.922
Sensitivity (Recall)	0.920
Specificity	0.998
F1 score	0.957
LogLoss	0.207

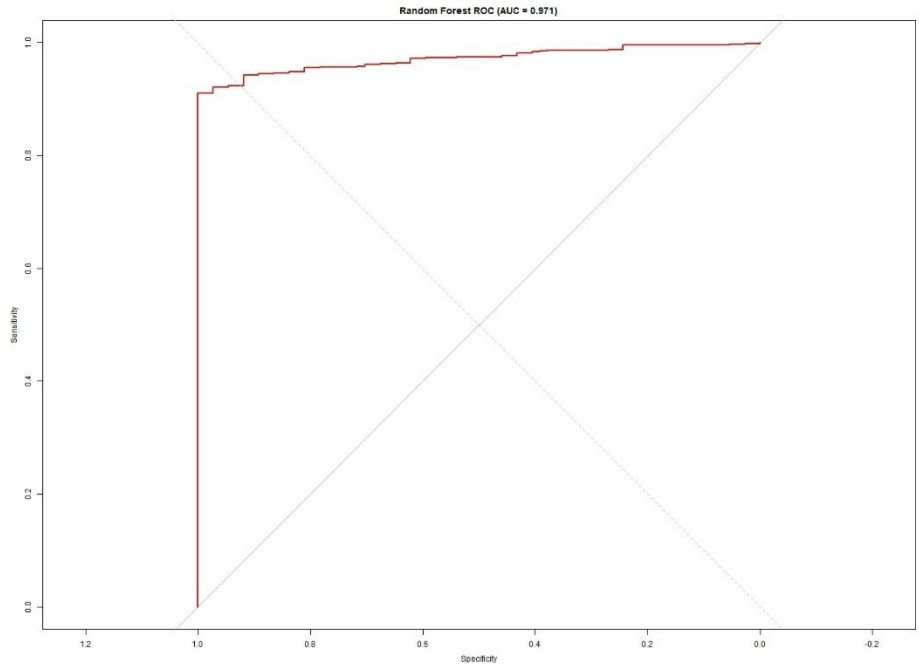


Figure 4.4.1 Random Forest ROC Curve

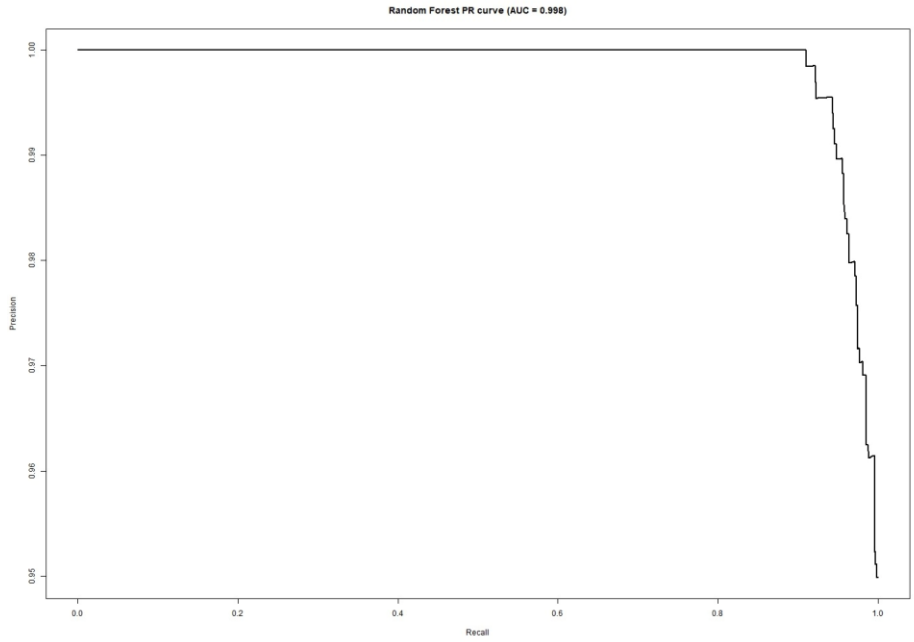


Figure 4.4.2 Random Forest Precision–Recall Curve

Two major measures that one should consider to select features in the Random Forest model were used. This provided us a more stable and a physically appropriate ranking of genes. The first one, Mean Decrease Gini, informs us the degree to which every gene assists in cleaning decision trees, which is a method of characterizing the

structure of the model. The second SHAP significance measure provides an explanation, independent of the model, of how individual genes manipulate all sample predictions. This research made use of the two benchmarks to determine the degree to which the Random Forest method performed independently and the manner in which each gene modified the group, overall, outcomes of the categorization. The list of the most important gene features is easier to understand and apply with the help of this two-ranking method. It also prevented noises-based forecasts being too big making the final list of attributes more consistent.

Table 4.4.2 Top 10 RF Gene Features (Using Combined Gini + SHAP Ranking)

Rank	Gene	Combined Importance Rank	Notes
1	**CLDN18**	1.5	Highest combined RF feature score
2	**AGER**	2.0	Strong SHAP contributor
3	**SFTPC**	5.0	Consistent high importance
4	**LINC01996**	6.0	High Gini + SHAP rank
5	**AL606469.1**	8.0	Strong native importance
6	**FAM107A**	10.0	Key lung-tissue marker
7	**MUC1**	12.0	Known cancer-associated gene
8	**SLC34A2**	13.0	Lung epithelium-associated
9	**SFTPA1**	15.0	Surfactant protein gene
10	**SFTPA2**	16.0	Surfactant protein gene

LightGBM Results

LightGBM also showed good predictive performance and modeling a number of non-linear gene interactions, especially between immune-related features.

Table 4.4.3 LightGBM Model Evaluation

Metric	Value
ROC-AUC	0.967
PR-AUC	0.998
Accuracy	0.941
Sensitivity (Recall)	0.945
Specificity	0.945
Threshold	0.85
F1 score	0.968
LogLoss	0.108

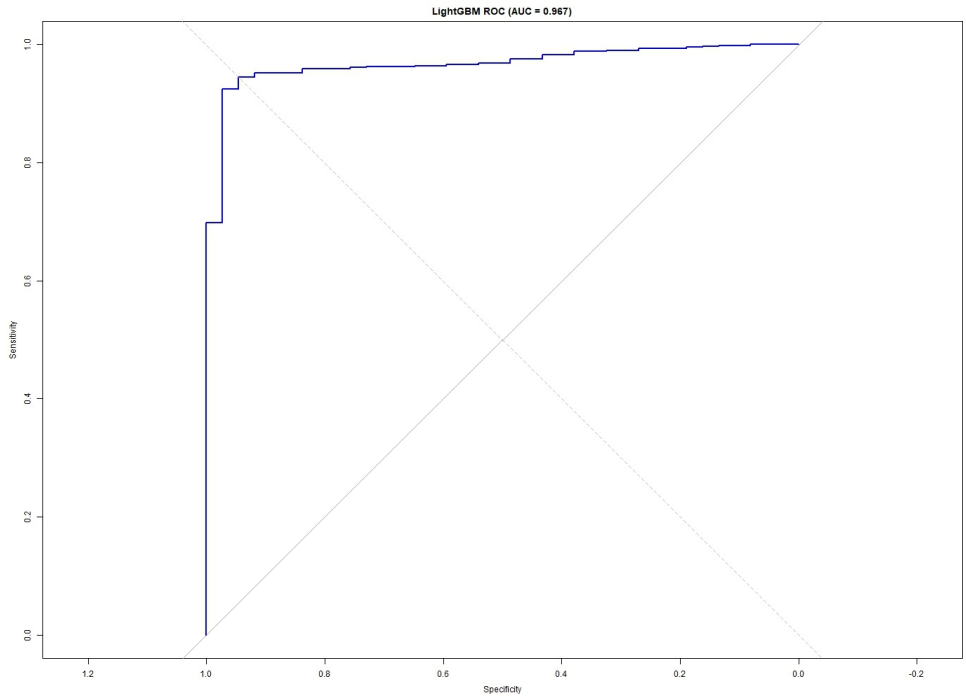


Figure 4.4.3 LightGBM ROC- AUC Curve

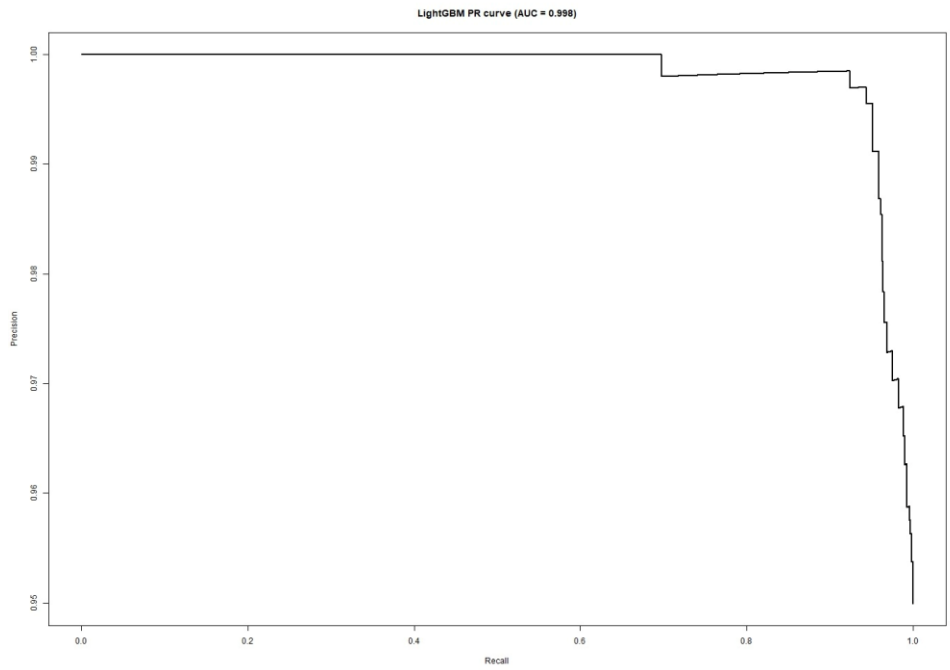


Figure 4.4.4 LightGBM PR- AUC Curve

A number of importance factors were employed by us to ensure that the LightGBM feature ranking has been stable and understandable. We firstly applied ratings of importance using gains to identify how individual genes contributed to make decisions regarding splitting of the new trees. Then SHAP values were plotted to demonstrate the extent to which each gene influenced the model predictions of all of the samples. This allowed discovering interactions and effects not on straight lines, and based on the circumstances. Lastly, a combination of these two standards was done to come up with a list that was concurred upon by all. This aided in ensuring that any changes that occurred when using a single approach were less prone to interfere with the list of the prognostic genes. This two part scoring method ensured that LightGBM was able to give up on both large world aspects and minor aspects that actually impacted real biologically.

Table 4.4.4 Top10 LGBM Gene Features (Gain + SHAP Rank Aggregated)

Rank	Gene	Final Score (avg_rank)	Notes
1	SFTPC	1.0	Strongest LGBM predictor
2	LINC01996	2.0	High SHAP + Gain weight
3	CLDN18	3.5	Consistently important across models
4	FAM107A	5.0	Distinguishes tumor vs normal
5	LANCL1-AS1	6.0	High SHAP-driven rank
6	AGER	7.0	Multiple ML model support
7	SCGB1A1	8.0	Lung secretoglobin family
8	SFTA3	9.5	Surfactant-related
9	SLC34A2	10.0	Tissue-specific marker
10	MUC1	11.0	Oncogenic pathway involvement

Elastic Net Results

Elastic Net identified genes with significant predictive value based on their non-zero coefficients and SHAP contributions.

Table 4.4.5 Elastic Net Model Evaluation

Metric	Value
ROC-AUC	0.967
PR-AUC	0.998
Accuracy	0.941
Sensitivity (Recall)	0.945
Specificity	0.945
Threshold	0.85
F1 score	0.968
LogLoss	0.108

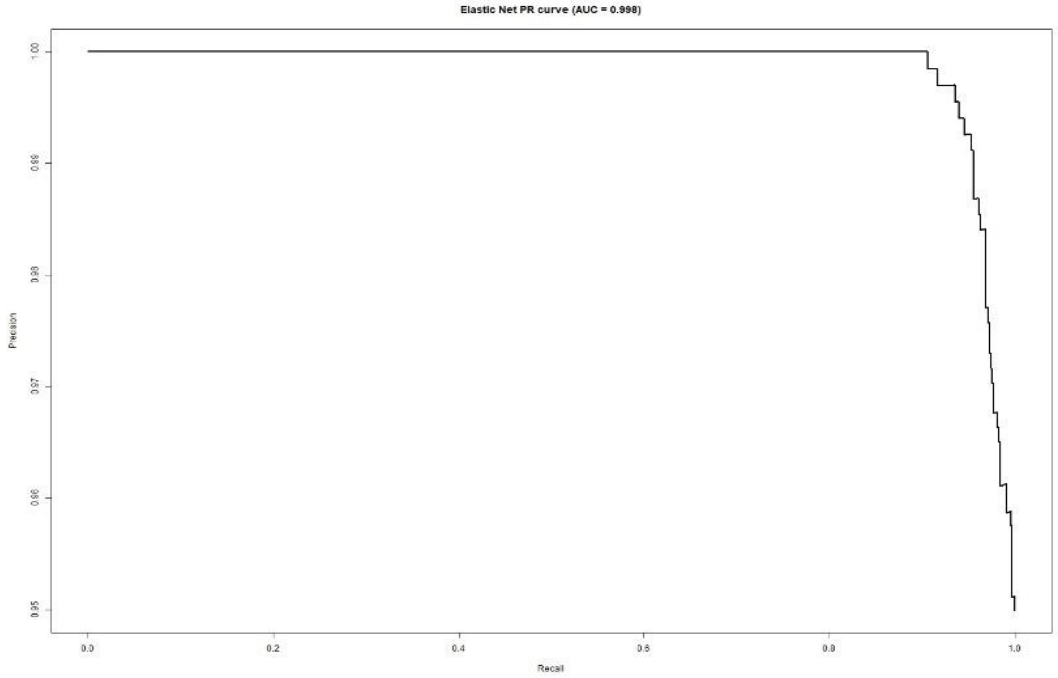


Figure 4.4.5 Elastic Net PR- AUC Curve

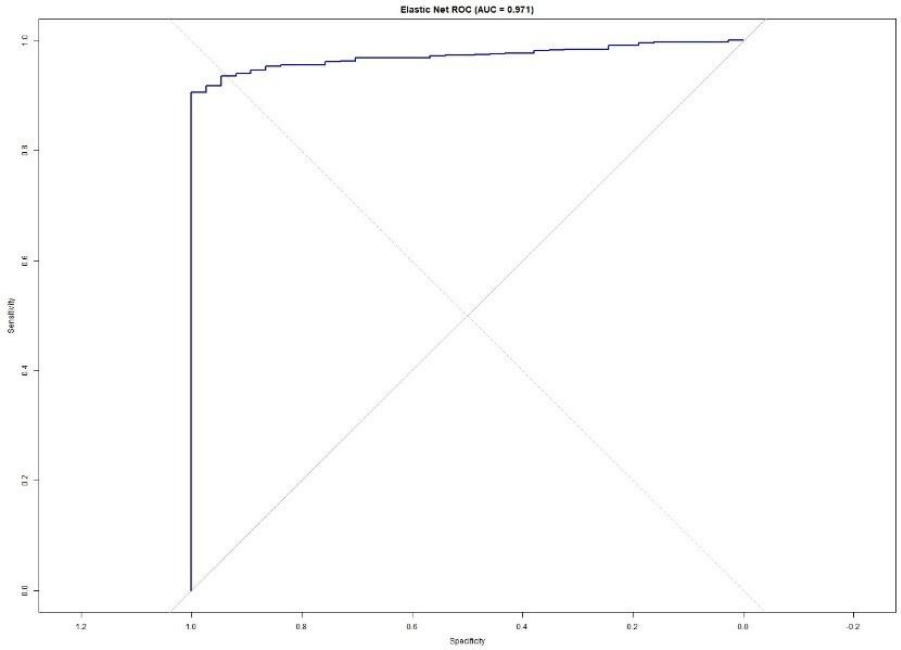


Figure 4.4.6 Elastic Net ROC - AUC Curve

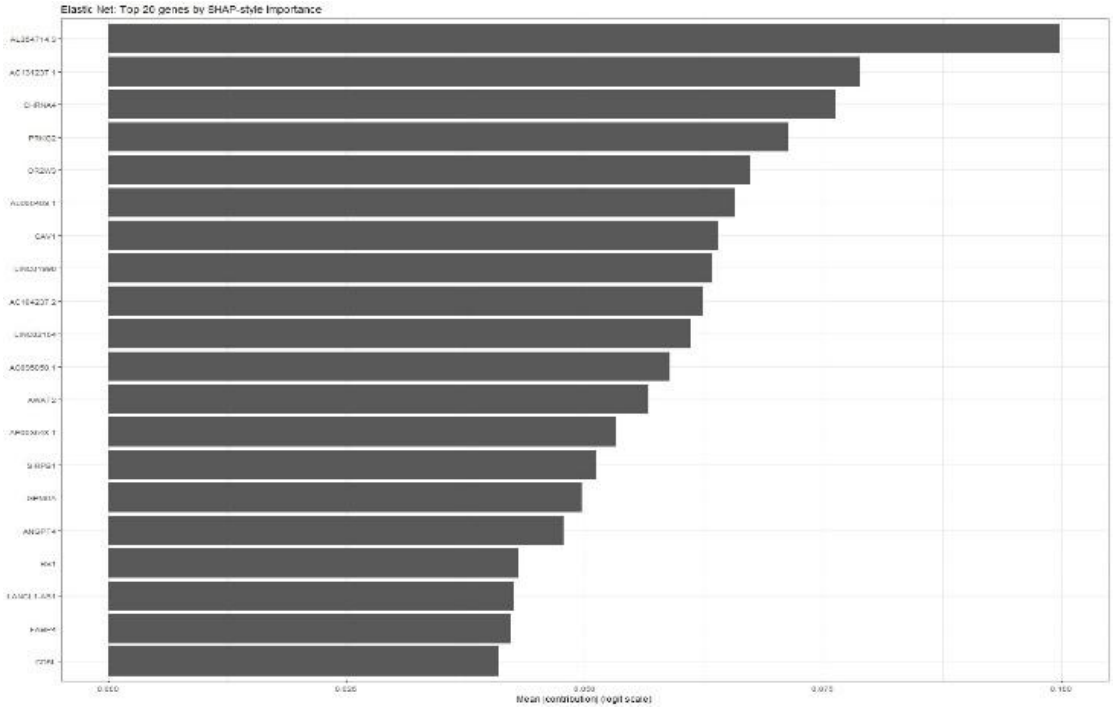


Figure 4.4.7 Elastic Net top 20 Genes by SHAP Importencr

Table 4.4.6 Top 10 Elastic Net Predictive Genes

Rank	Gene	Coefficient	SHAP Rank	Notes
1	**AL354714.3**	-0.073619	1	Highest EN impact, strongest SHAP signal
2	**AC131237.1**	-0.067579	2	Strong predictive contribution
3	**CHRNA4**	-0.054603	3	Highly stable, strong negative coefficient
4	**PRKG2**	-0.055291	4	Robust regulatory gene
5	**OR2W3**	-0.054005	5	Stable, consistent SHAP importance
6	**AL606469.1**	-0.043835	6	Strong mid-level predictive weight
7	**CAV1**	-0.048182	7	Known signaling + tumor microenvironment gene
8	**LINC01996**	-0.033692	8	Cross-model important lncRNA
9	**AC104237.2**	-0.051864	9	Strong model contribution
10	**LINC02154**	-0.032440	10	Stable predictive lncRNA

4.5 Comparison of Feature Sets (Statistical vs ML)

To understand the complementary strengths of both analytical approaches, four distinct gene categories were generated:

Table 4.5.1 Gene Integration Summary

Category	Count	Description
Stat-only genes	1,714	Present only in limma/DESeq2 union
ML-only genes	4,377	Selected by at least one ML model
Common Stat ∩ ML	5,625	Overlap of statistical + ML features
Union Stat ∪ ML	11,714	All unique genes recovered by any method

4.6 Pathway Enrichment Results

Each of the four gene groups underwent independent enrichment using:

i)GO Biological Process (GO:BP).

ii)GO Cellular Component (GO:CC).

iii)GO Molecular Function (GO:MF).

iv)KEGG pathways.

v)Reactome pathways.

The following subsections summarize the results.

Table 4.6.1 ML-Only Top 5 Pathways

Category	Highlighted Findings
GO:BP	Adaptive immune response, immune system process
GO:CC	T-cell receptor complex, immunoglobulin complex
GO:MF	Antigen binding
KEGG	Herpes simplex virus 1 infection, cytokine–cytokine receptor interaction
Reactome	B-cell receptor (BCR) signaling, CD22 mediated BCR regulation

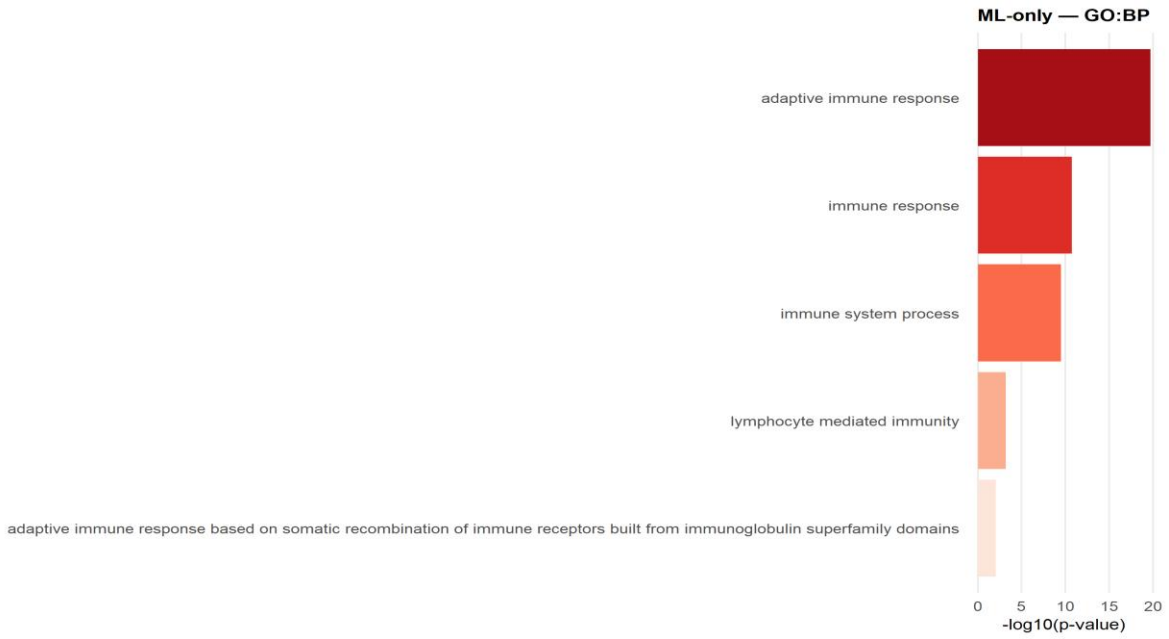


Figure 4.6.1 Top GO Biological Processes enriched in ML-only genes

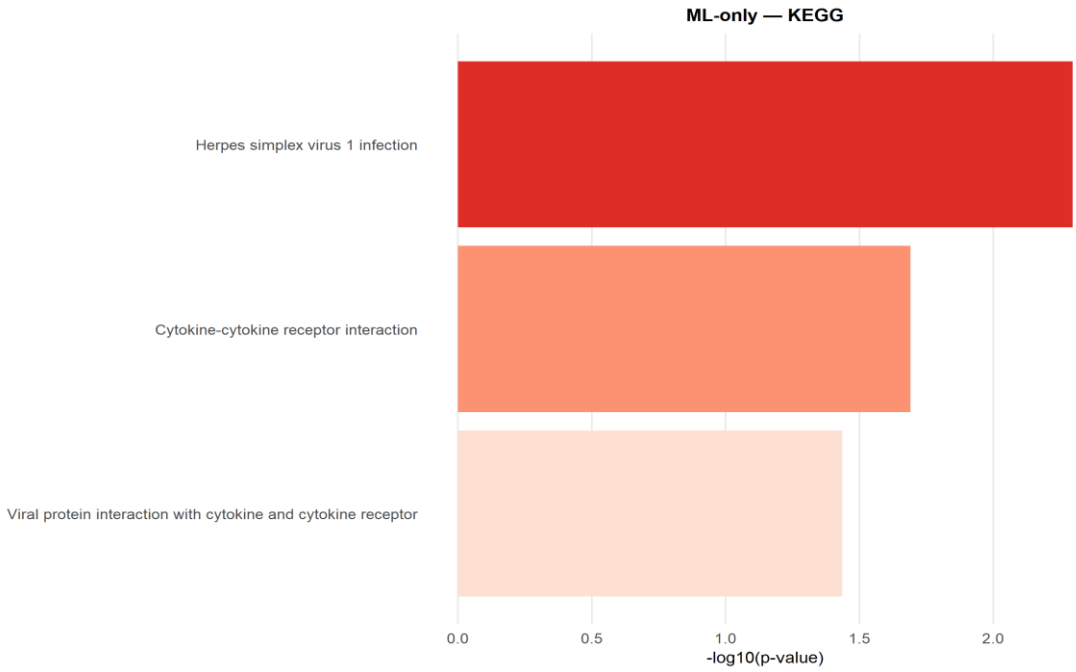


Figure 4.6.2 Top KEGG pathways enriched in ML-only genes (ranked by p-value)

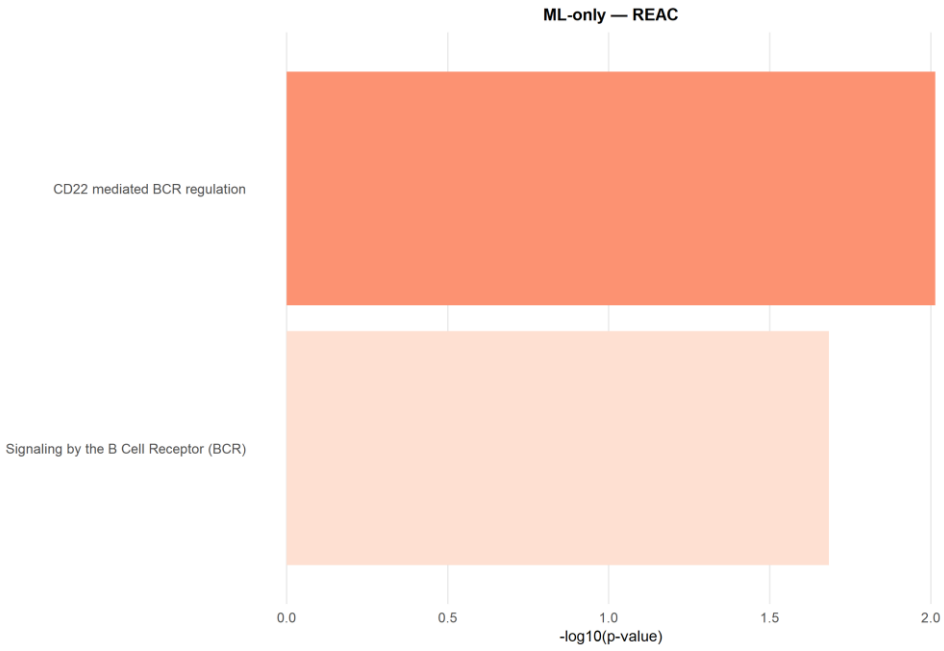


Figure 4.6.3 Top Reactome pathways genes only for Machine Learning.

ML-only pathways were predominantly immune-related and especially the B-cell and T-cell activation.

Table 4.6.2 Stat-Only Top 5 Pathways

Category	Key Findings
GO:BP	Cell cycle, DNA replication, DNA repair
GO:CC	Nucleoplasm, cytosol
GO:MF	Nucleotide binding, enzyme binding
KEGG	Adherens junction, Osteoclast differentiation
Reactome	RHO GTPase cycle, S-phase regulation

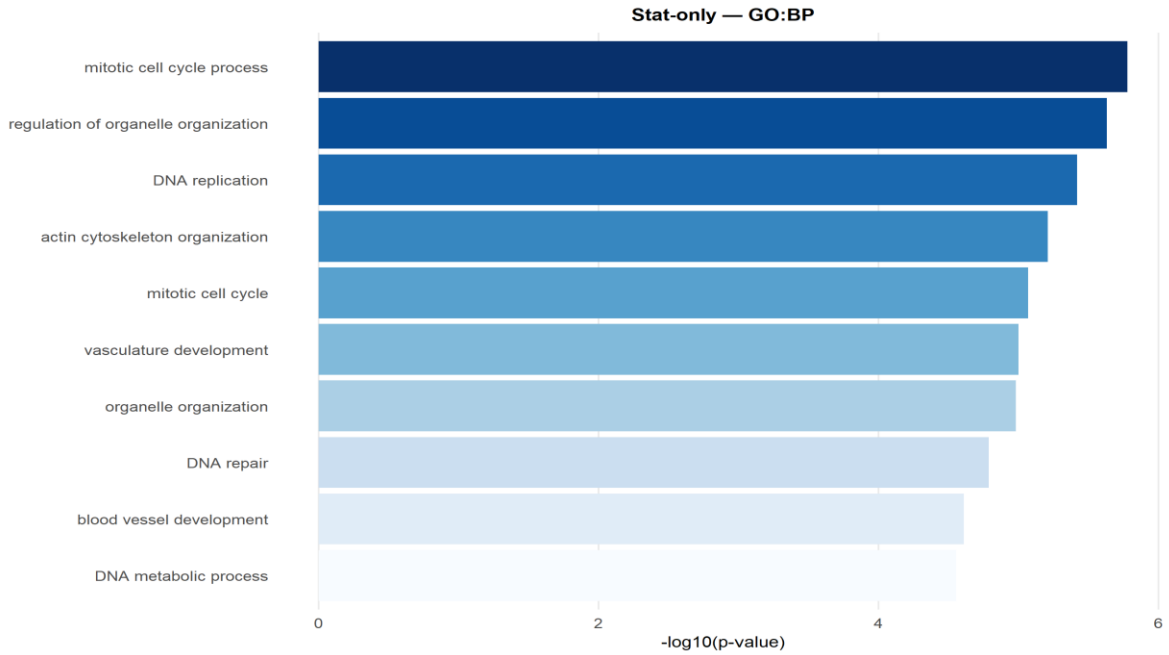


Figure 4.6.4 Top GO Biological Processes enriched in stat-only genes (top terms by significance)

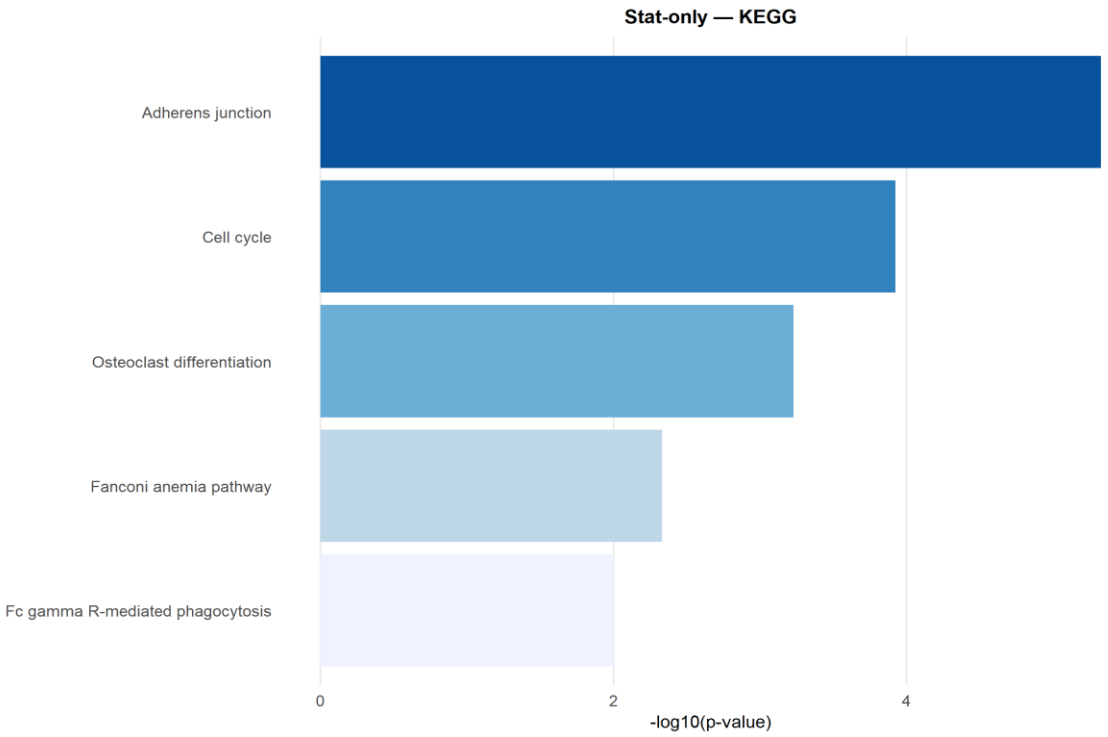


Figure 4.6.5 Top KEGG pathways enriched in stat-only genes (ranked by p-value)

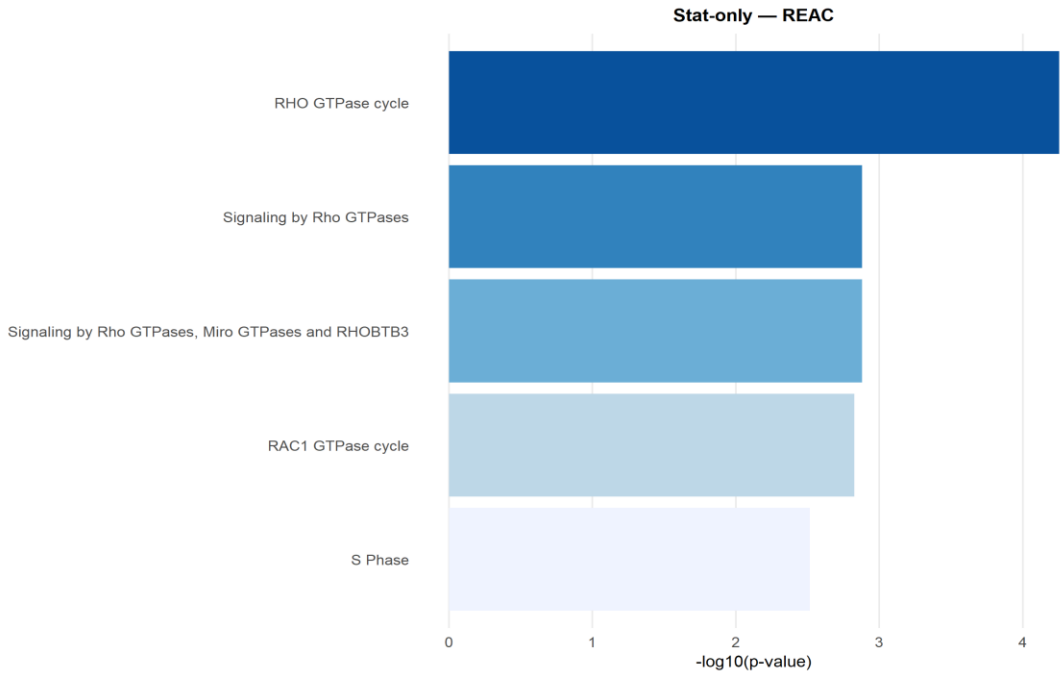


Figure 4.6.6 Top Reactome pathways enriched in stat-only genes (ranked by significance)

These pathways showing us poroliferation-driven tumorigenic signatures.

Table 4.6.3 Common Top 5 Pathways

Category	Key Patterns
GO:BP	Developmental process, cellular process
GO:CC	Extracellular region, organelle
GO:MF	Binding activity (protein, molecular, ion)
KEGG	Steroid hormone biosynthesis, Retinol metabolism
Reactome	Broad metabolic functions

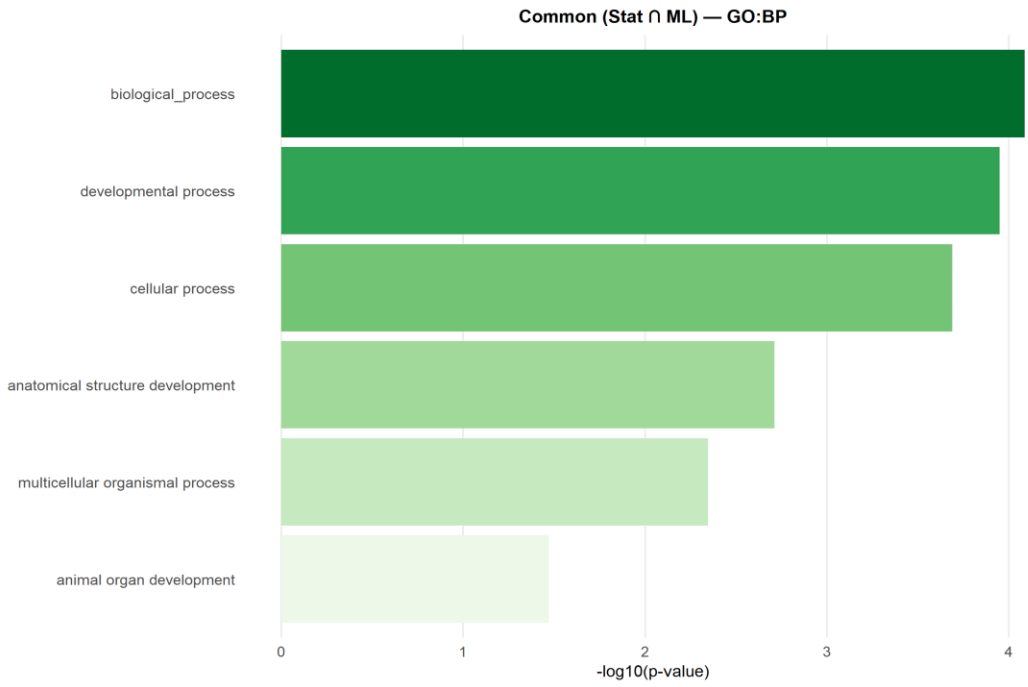


Figure 4.6.7 Top GO Biological Processes enriched in genes common to both statistical and ML approaches.

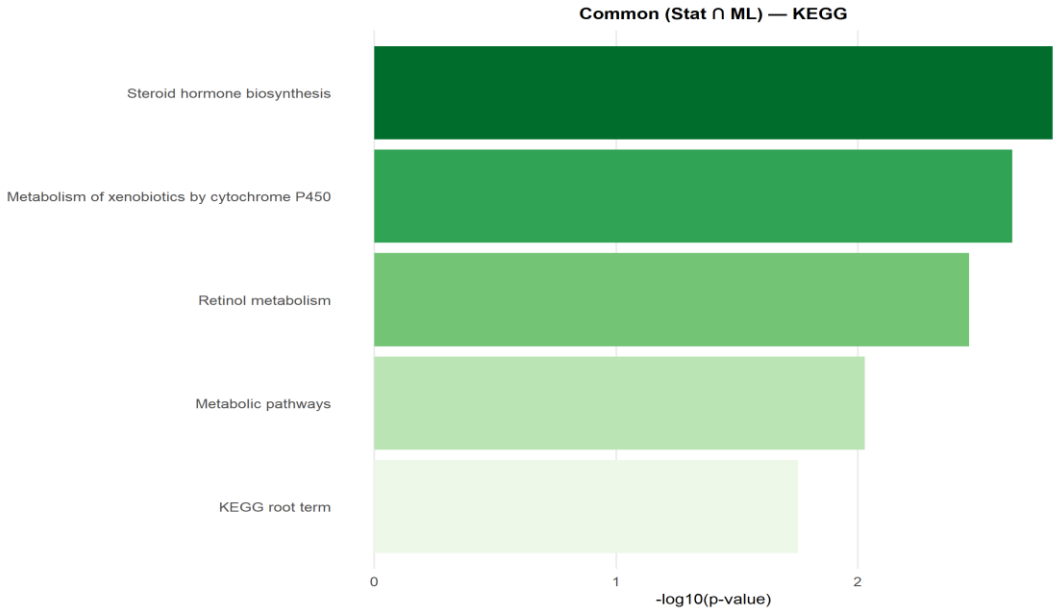


Figure 4.6.8 Top KEGG pathways enriched in the common gene set.

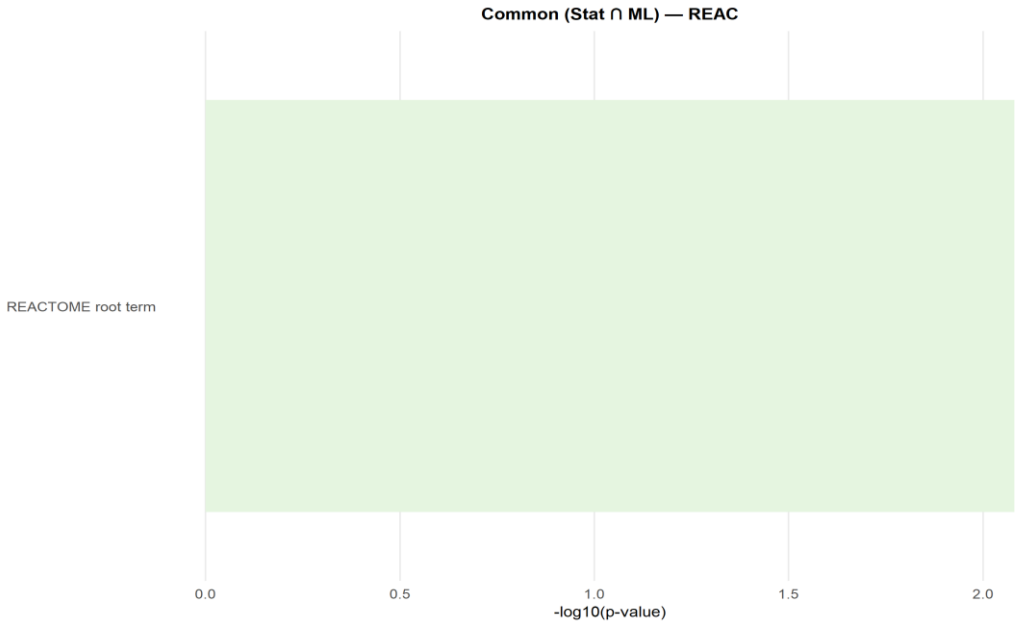


Figure 4.6.9 Top Reactome pathways enriched in the common gene set.

This setup captures biologically stable pathways agree with both the statistics and ML.

Table 4.6.4 Union Pathway Overview

Category	Result
GO:BP	Broadest enrichment — biological process, cellular response, developmental pathways
Reactome	Large-scale metabolic and root-level pathways
KEGG	Multiple metabolic pathways

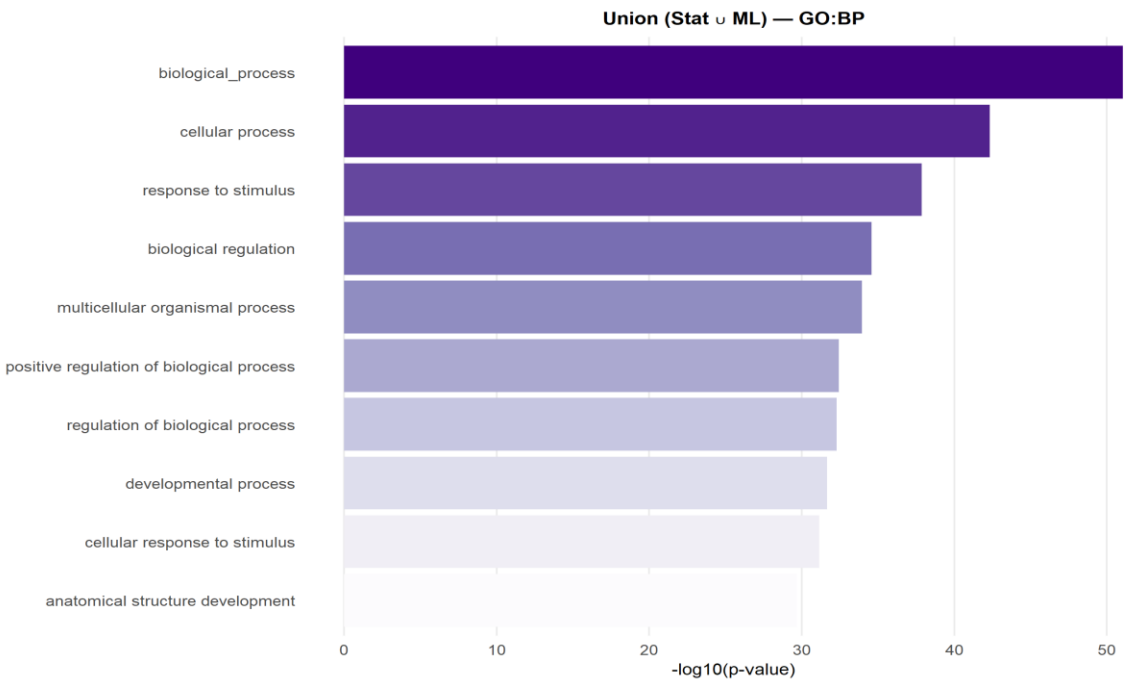


Figure 4.6.10 Top GO Biological Processes across the union of Stat and ML gene sets .

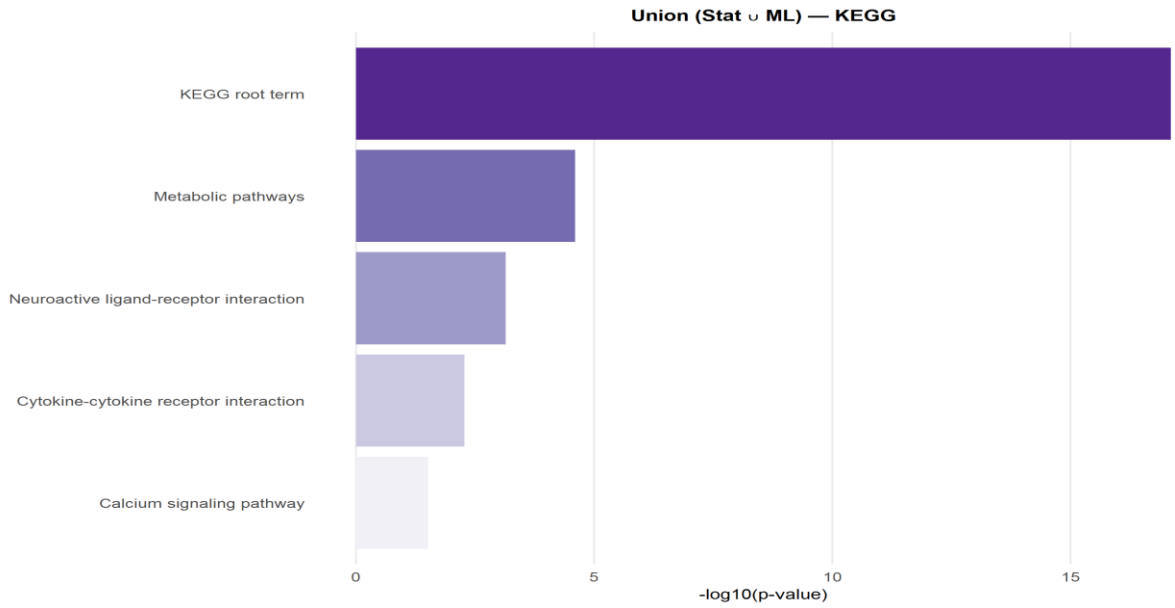


Figure 4.6.11 Top KEGG pathways across the union gene set.

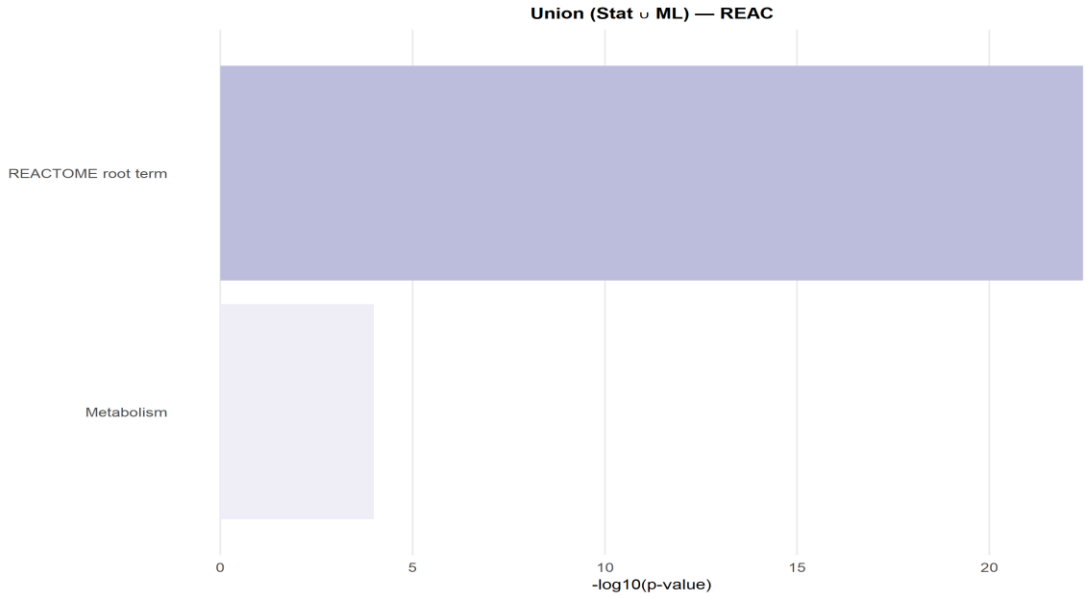


Figure 4.6.12 Union Reactome terms

Union pathways represent the complete functional landscape captured by the any method.

4.7 Comparison of Pathway Volume Across Gene Sets

Table 4.7.1 Pathway Count Summary

Gene Set	Pathway Count
Stat-only	**101**
ML-only	**17**
Common (Stat ∩ ML)	**30**
Union (Stat ∪ ML)	**259**

4.8 Final Key Findings

The study taught us a lot that we can use to learn more about how DNA works in lung cancer cells. Statistical methods missed strong signs of immune response, but machine-learning models did. T-cell receptor complexes, contact between cytokines and receptors, and control of B-cell receptors are some of the ways that the signals got there. This is how ML systems can find small, nonlinear connections between immune system-related genes. The statistical pipelines, on the other hand, mostly found paths that had to do with how cells split and grow, such as mitotic progression, DNA damage repair mechanisms, and replication control. This is an example of how well they can find big, steady changes in fold-change between healthy tissues and tumors. Both approaches led to the same set of developmental and metabolic pathways. These were processes that involved multicellular organism, steroid hormone production and overall metabolic processes. In some molecular ways, this shows that the two methods are the same. Important immune signals were missed by statistical tools, but strong cell-cycle paths were found inside tumors by statistical tools that ML did not show. As this helpful behavior shows, both methods can be useful. We think that a mixed analysis method would help us get a fuller and more biologically useful picture of lung cancer transcriptomics.

CHAPTER 5

DISCUSSION & FUTURE WORK

5.1 Overview

The research used a multi-level pathway enrichment, multi-step machine learning, and statistical use of a differentiable framework of differential expression in order to achieve biologically relevant and complementary molecular signatures associated with lung cancer. The major aim was to assess the value of traditional statistical (limma-voom, DESeq2) and multi-model machine learning (Random Forest, LightGBM, Elastic Net) to understand tumor biology, and to understand whether machine-learning approaches are able to identify non-linear and weak-random patterns that cannot be identified by classical DE tests. The findings reveal that the statistical pipelines and the ML pipelines are distinct but the complementary biological processes necessitate a combined analysis.

5.2 How to Understand Statistics

Statistical analyses performed on DESeq2 and limma-voom have found a total of 7,337 differentially expressed genes, and 2,155 of them had significant and consistently significant differences in expression between tumor and normal tissue. Most of these statistically significant genes were the regulators of such well studied and known proliferative and cell-cycle-regulated pathways, as regulators of mitosis, DNA replication, chromosomal segregation, and genome-maintenance processes. Researchers came to a conclusion that the factor that makes lung cancer progress is that the rate at which cells divide is high, the genome is unstable, and replication is overstressed. Interesting variation of gene tumor markers were the MCM family, TOP2A and CCNB1/2. This proved the reliability of the data and that the statistical models detected tumor gene signals. Statistical methods have some strengths; but they cannot be satisfactorily applied in determining immune-mediated or microenvironment-related patterns of expression. Share with the multiple immune-related pathways that are nonlinearly associated with the genes or shows only mild changes, which although biologically relevant, do not meet the stringent statistical significance criterias. It is also highly heterogeneous in terms of the tumour microenvironment, varying by individual, and it also introduces considerable amount of variation rendering group level statistical differences less powerful. These issues

provided evidence that more complex patterns of interaction than linear fold-change differences require machine-learning techniques.

5.3 Interpretation of Machine Learning Results

The ML models (Random Forest, LightGBM, Elastic Net) were fitted based on 10,000 highly variable genes, where class weights and threshold tuning were performed to take care of the vehemently unbalanced sample (37 normal vs. 702 tumor). The joint ML models were able to find many 4377 ML-only genes that did not have substantial fold changes but still high classification performance results.

Non-linear patterns detected by ML

Conventional statistical techniques overlooked a special category of nonlinear expression configurations that were exhibited by the machine-learning investigation. The genes (ML-only) were highly enriched in immune-related pathways including T-cell activation, antigen presentation, cytokine-receptor signaling and B-cell receptor pivotal procedures. This implies that the tumor immunological microenvironment contributes significantly to lung cancer. These pathways have failed to pass statistical significance in DESeq2 or limma-voom, to a large extent due to the fact that immune-associated signatures are typically very heterogeneous across tumor samples and can usually induce only small fold-changes despite containing biologically important signals. Their useful effect is compared to co-expression networks, not mere group-mean differences, and therefore they are detected with more ease using the models that can acquire hierarchies of interactions. Machine-learning models e.g., Random Forest, LightGBM or Elastic Net can then find these multi-gene relationships and nonlinear relationships, allowing them to identify patterns of immune regulation, which are completely ignored by standard differential expression tests. Subsequently, the ML pipeline identified minor yet significant immunological signals that can give a more in-depth insight into tumor-immune interactions in lung cancer.

5.4 Complementarity Between Statistical and ML Gene Sets

The combination between the statistical and the machine-learning output generated four significant gene groups. They are merely Analysis Driven Signatures based solely on statistics and Immune and Microenvironment Signals based solely on machine-learning and Common, and Union-all which all present a separate layer of lung cancer biology. Comparing of their pathways enrichment patterns revealed that the diverse yet complementary biological processes assemble these groups. This complement is described by the fact that such combination of statistical and machine

learning methodologies provides a more complete and reliable analysis of tumor activity compared with the application of either of these methodologies separately.

5.4.1 Stat-Only Pathways (Statistical Analysis–Driven Signatures)

Stat-only gene set (1,714 genes detected by limma-voom and DESeq2 only) was highly enriched in which the pathways of the core malignant transcriptional program in the lung cancer were involved. These pathways were involved in work with DNA replication, cell-division apparatus, maintenance of chromosomes and several elements of the mitotic checkpoint apparatus. These results underpin the fact that the lung tumors are a result of the uncontrolled proliferation, replication stress, and genomic disorder which is well established in prior studies. These pathways were well represented using statistical models as they depend on high, uniform increases in fold-change that only tumor cells that were growing in very large numbers can generate the right signal. Consequently, the highest tumor-intrinsic molecular changes defining the direction of cancer evolution are mirrored in the only statistical analysis group.

5.4.2 ML-Only Pathways (Machine-Learning Driven Immune and Microenvironment Signals)

The ML-only gene set, however, contained 4,377 genes whose absence in normal DE analysis, nevertheless, was significant in machine-learning prediction. The most probable functions of these genes were in the immune-oncology processes and these include look like interferon signaling, antigen processing and cytokine action. The TME makes messages regarding the inflammation and the disease, which comprise a significant portion of the lung cancer development and propagation. Their genes do not move so widely, they are specific to individuals and therefore, linear statistical models can hardly locate them. That is why they do not get included in DEG statistics lists. Instead, the machine-learning algorithms could accurately identify these types of nonlinear and context-dependent patterns in our prediction of co-expression dynamics and high-order interactions. This is because the ML-only routes identify biologically significant immune signatures that the conventional analyses would not identify.

5.4.3 Common Pathways (Shared Statistical ∩ ML Biological Signals)

Statistical and machine learning methods discovered that 5,625 genes in the Common gene set overlap. These genes exhibited physically stable and reproducible signals across both analytical frameworks. In this group, pathway enrichment exhibited

patterns that were related to how cells grow and divide, how steroid and retinol are used, how the extracellular matrix is formed, and how cells accomplish their most critical activities. These pathways appear to be less sensitive to methodological differences because they involve genes whose expression changes are both statistically strong and predictive for classification. The presence of these shared pathways indicates that some molecular mechanisms such as metabolic reprogramming and structural remodeling are consistently dysregulated in lung cancer and robustly detectable using both linear statistical contrasts and nonlinear ML-based feature selection. This similarity in the biology enhances the trust in the validity of such results.

5.4.4 Union-All Pathways (Broad Combined Landscape of Lung Cancer Biology)

The most comprehensive pathway enrichment profile of the 4664 genes sets is Union-all genes collections and the overall all 11714 different genes identified by either of the two techniques, which showed the 259 significantly enriched pathways. This vast viewpoint had markers that were associated with tumor intrinsic growth and immune based microenvironment regulation. These cell-cycle regulatory, DNA replication, and chromosomal integrity networks were joined by immune-modulatory, metabolic, developmental, and signaling networks. The wide range of the Union-all group attests to how the combination of statistical and machine learning practices will embrace the totality of lung cancer biology, that is, the processive growth of aggressive tumor cells, immunological interactions, and metabolic reconfigurations. This integrative point of view underlies the key observation of the current research: multi-method consensus is a more comprehensive and more detailed model of tumor activity than any single method of analysis.

5.5 Biological Implications of Key Results

There are several biological implications of this study which elucidated the complex tumor-intrinsic and microenvironmental behavior of the lung cancer. Among the most remarkable findings is the fact that is the machine learning models only identify the immune-related mechanisms like B-cell activation, T-cell activation, cytokine signaling, and antigen-presentation. It means that immune heterogeneity is an essential aspect of lung cancer biology, yet it is challenging to observe with the standard fold-change based statistical analysis. New immune pathways tend to show moderate and strongly changing expression in patients that makes them less likely to be reproducible by strict statistical cutoffs. However, such immune signals have been of paramount importance when it comes to predictive responses to immunotherapy, tumor-infiltrating lymphocytes (TILs) and patient identification of those most likely

to respond to checkpoint-inhibitor therapy. The fact that the ML models are able to detect these immune properties indicate that non-linear interactions between genes and context-dependent patterns are a significant factor in determining the interaction between lung tumor and the immune system. Conversely, the statistical DEGs revealed the basic proliferative and cell-cycle processes that were at the center of the tumor development. Recurring calculations showed the genes in the mitotic controls, maintenance of the chromosomes, and the replication of DNA. This supports the fact that these genes could be the cause of irregular division of cells and genetic instability. These tracks indicated the reality regarding the cancers, such as the DNA repair issues and replication stress. These characteristics render the tumors aggressive and this is most of the times the cause of a bad effect in the clinic. The growth genes differ greatly between the malignancies and healthy cells. The regular statistics tools can effortlessly locate the differentials of analysis to demonstrate that they are the right and reliable expressions.

When machine learning immune signatures are combined with statistical proliferation signatures, it is demonstrated that lung cancer is controlled by combined molecular programs. The gene sets, Common and Union have immunity and proliferative mechanisms. Internal processes contribute to the development of cancers and how the immune system treats them depends on the interactions. Numerous indications suggest that it is necessary to use a complex of medicines that will address both the development of tumor cells and the functioning of the immune system. In this study biological mechanisms are combined to explain lung cancer. This makes it easier to be the precise and focused in the medicine.

5.6 Importance of the Integrative Hybrid Modeling Approach

The findings of the current work prove the apparent worth of applying to the hybrid approach of analysis combining conventional statistical tools with machine-learning models. DESEQ2 and limma-voom classical differential expression tools are very effective in identifying large linear and tumor intrinsic transcriptional changes in specific cases i.e. proliferation, mitosis, DNA replication, and chromosome regulation. These notable patterns are apparent in all the samples and play an important role in the functioning of lung cancer. Conversely, the machine-learning models were more effective at detecting nonlinear, context-specific and co-expression signals which are readily missed in the test of statistics. They included Immune-activation markers, cytokine signaling, and antigen-presenting pathways and illustrate the process in which the tumor microenvironment is differentiating and transforming. Combine one of these twos disparate ways of the thinking, and you have a clearer understanding of how this tumors operate. This combination is

biologically significant in the pathways that one or the other method failed to detect, and this leads to reduced false negatives. It also identifies patterns in the rules that would not be apparent in pipelines that utilize a single approach. This increases the list of the biomarker candidates. More to the point, it increases the interpretability of the results by confirming powerful tumor-intrinsic processes using statistics and in subtly identifying immune-derived effects using machine learning. In complicated diseases like lung cancer, where there are several layers of molecules interacting at the same time. This discovery paradigm of multi-resolution provides a more precise and comprehensive picture as compared to any one analytical approach.

5.7 Limitations

Hybrid statistical-machine learning paradigm has proven itself to be quite fruitful in biological discoveries; however, it is important to note that the paradigm has a number of serious limitations that must be taken into account in order to be able to interpret the results properly. To begin with, the sample was not quite balanced. It contained 702 tumor samples and 37 healthy cell samples. This implies that the model must acquire a new rank of classes and threshold. These means may be useful to even the playing field, yet they may alter the way of how the model functions or where the boundary between classes is made. Second, SHAP values give a nice view of how models determine what to do, however, these may not give a full picture of how the ML models are not clear and how the features become more or less important when the model is configured differently. Third, route enrichment findings are determined by the completeness of the databases of annotations. This implies that the enrichment outcomes might not contain new genes or genes which are not well characterized. This can complicate the process of comprehending the working of biology. Lastly, this paper has examined only one of the largest RNA-seq datasets. As no one has checked this dataset, one cannot definitely state whether the gene patterns and pathways identified in this study are similar in other groups of people or other sequencing platforms. These issues do not diminish the importance of the current results; on the contrary, they demonstrate areas in which further research might be conducted to make this model stronger and applicable in different cases.

5.8 Future Work

Such research outcomes allow us to conduct much greater research which will enable us to know biology better and easier to use it in treatment. External confirmation is very essential as the second step. To accomplish this, we have to examine gene sets, ML-derived patterns, and pathway outcomes of various groups of NSCLC cells originating of GEO, ArrayExpress, and TCGA, and new RNA-seq data on a single

cell. It would contribute to determining whether the signatures can be utilized in sequencing a large number of various groups and systems. Subsequent studies ought to involve a great deal of clinical information, such as the effectiveness of treatment, the degree of cancer staging and grading, as well as the extent of immune infiltration. This will assist in establishing the usefulness of combined statistical and machine learning gene signatures in life prediction or in prediction life. The other alternative is to learn numerous omics. Inclusion of the copy-number changes in DNA, DNA modification by methylation, proteomics, and metabolomics will indicate regulation in a higher stage than transcriptomics. This will shed light in tumor activity and response to treatment. Neural networks, transformers, and autoencoders can be trained to detect more nonlinear features due to their ability to uncover the patterns and dynamics of gene-expression in space and time that a normal machine-learning model would overlook, and nonlinear features have been shown to be of use in drug repurposing studies, particularly in immune-related genes that machine-learning models have discovered.Moreover, coupling drug-gene interaction databases with real-world drug-response datasets should be useful in drug repurposing studies by detecting the patterns and nonlinear dynamics of gene-expression in space and time, The mentioned consensus ML features, which are reported in this work, have the potential to be converted to a powerful diagnostic, prognostic, or treatment-selection classifier, potentially conforming to the foundation of a therapeutically significant decision-support tool once tested on large enough patient groups. Later, these guidelines will strengthen and make the hybrid modeling technique that will be proposed in this study useful and more probable to be utilized in other areas.

CHAPTER 6

CONCLUSION

6.1 Discussion & Conclusion

This study uses statistical differential expression, machine learning–based feature selection, and multi-level pathway enrichment to discover lung cancer signals with biological significance. The paper uses a difficult and imbalanced dataset of 739 samples (702 tumor, 37 normal) and 59,429 genes to show how old statistical methods and new machine-learning methods provide complimentary tumor biology insights.

Key Results:

Limma-voom and DESeq2 were jointly statistically analyzed to reveal 7,337 differentially expressed genes, which does suggest that the tumor was producing many signals. Such genes were primarily associated with the cell cycles, DNA repair and replication and integrity of the chromosomes. This can be compared with the standard indicators of malignant progression and complements the processes of the dataset in question.In addition to these findings, machine-learning algorithms, which may be referred to as Random Forest, LightGBM, and Elastic Net, extracted 4,377 predictive genes in the cohort of highly variable variables. A large number of these genes were not found to be of particular importance in standard differential-expression analyses, but patterns of interest were always apparent within the machine learning models, which revealed more intricate biological interactions not known using a simple fold-change metric. Upon coordinated analysis of both statistical analysis and machine learning results, there were four groups of features, with the statistically acquired only group being the proliferative one; the machine learning acquired only group the immune and microenvironmental ones; the ones acquired as shared between the two techniques and the ones with the greatest activity in terms of functional operation represented the successful combination. Along with that, each group was associated with the biological domains including a look like the cell cycle, DNA repair, cytokines signaling, metabolic regulation, and wider functional networks using The pathway enrichment analysis.

6.2 Final Thoughts

Collectively, this evidence confirms that integrated use of traditional statistical modeling, and machine learning methods gives a more multifaceted view of biological phenomena as opposed to using either of the methods individually. Statistical approaches discovering changes in the level of expression which are stable and significant, and machine learning approaches are excellent at discovering little, non-linear interplay among genes. The combination of the two points of view results in improved gene signatures, a expanded pool of pathways that are covered, and future applied research. When applied in the context of lung cancer, the combined approach achieved the accidental emergence of three significant layers of disease biology, which include intrinsic proliferative signals, immune-microenvironment interaction and shared metabolic or developmental processes. These observations have a potential to streamline precocious biomarker identification, inform therapy choices and enhance our knowledge on tumor immune interactions.

REFERENCES

[1] H. Sung *et al.*, "Global cancer statistics 2023: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin.*, vol. 73, no. 1, pp. 1–33, 2023, doi: 10.3322/caac.21763.

[2] N. Suresh *et al.*,"Non-small-cell lung cancer,"*Nat. Rev. Dis. Primers*, vol. 10, pp. 1–28, 2024, doi: 10.1038/s41572-024-00466-6.

[3] J. Kerr *et al.*,"The evolving clinical landscape of non-small-cell lung cancer therapies: advances, challenges and future directions,"*Lancet Oncol.*, vol. 24, no. 3, pp. 345–362, 2023, doi: 10.1016/S1470-2045(22)00859-5.

[4] F. Zhang *et al.*,"Multimodal omics analysis of the EGFR signaling pathway in non-small cell lung cancer and emerging therapeutic strategies,"*Signal Transduct. Target. Ther.*, vol. 8, no. 1, pp. 1–14, 2023, doi: 10.1038/s41392-023-01512-4.

[5] K. Wang *et al.*,"Proteogenomic analysis of lung adenocarcinoma reveals tumor subtypes and therapeutic vulnerabilities,"*Cell Rep. Med.*, vol. 3, no. 2, pp. 100–112, 2022, doi: 10.1016/j.xcrm.2021.100650.

[6] Q. Chen *et al.*,"An integrated single-cell transcriptomic dataset for non-small cell lung cancer,"*Sci. Data*, vol. 10, pp. 1–12, 2023, doi: 10.1038/s41597-023-02068-1.

[7] L. Li *et al.*,"Single-cell profiling after neoadjuvant chemo-immunotherapy reveals immune activation in non-small cell lung cancer,"*Cell Death Dis.*, vol. 13, pp. 1–15, 2022, doi: 10.1038/s41419-022-05021-7.

[8] H. Wu *et al.*,"Molecular profiling of human non-small cell lung cancer by single-cell RNA sequencing,"*Genome Med.*, vol. 14, pp. 1–18, 2022, doi: 10.1186/s13073-022-01040-z.

[9] R. Patel *et al.*,"Single-cell and bulk RNA-seq analysis identifies MUC1 as a key gene for lung adenocarcinoma to neuroendocrine transformation,"*BMC Biol.*, vol. 21, pp. 1–14, 2023, doi: 10.1186/s12915-023-01608-7.

[10] A. Singh *et al.*,"RNA sequencing identifies novel transcriptomic markers and molecular alterations in non-small cell lung cancer,"*J. Transl. Med.*, vol. 20, pp. 1–12, 2022, doi: 10.1186/s12967-022-03440-2.

[11] J. Wang *et al.*,"Investigation of radiation-induced transcriptome profile of radioresistant non-small cell lung cancer A549 cells using RNA sequencing," *Radiat. Oncol.*, vol. 18, pp. 1–11, 2023, doi: 10.1186/s13014-023-02293-9.

[12] M. G. Fisher *et al.*,"Liquid biopsy diagnostics for non-small cell lung cancer via elucidation of tRNA signatures,"*Genome Biol.*, vol. 25, no. 2, pp. 1–19, 2024, doi: 10.1186/s13059-024-03059-x.

[13] S. Kang *et al.*,"LGR5+ stem cell transcriptomic analysis reveals cellular programs linked to lung cancer susceptibility,"*Nat. Commun.*, vol. 13, pp. 1–13, 2022, doi: 10.1038/s41467-022-29997-7.

[14] X. Zhao *et al.*,"Single-cell RNA sequencing reveals enhanced antitumor immunity after PD-1 inhibitor treatment in non-small cell lung cancer,"*Nat. Commun.*, vol. 14, pp. 1–14, 2023, doi: 10.1038/s41467-023-41917-9.

[15] M. Yang *et al.*,"Decoding the role of lipid metabolism in non-small cell lung cancer: from macrophage subtype identification to prognostic model development," *Front. Immunol.*, vol. 14, pp. 1–13, 2023, doi: 10.3389/fimmu.2023.1223456.

[16] A. Petrova *et al.*,"RNA sequencing in non-small cell lung cancer shows gene downregulation of therapeutic targets in tumor tissue compared to non-malignant lung,"*Cancer Genomics Proteomics*, vol. 20, no. 1, pp. 45–55, 2023, doi: 10.21873/cgp.20323.

[17] N. Huang *et al.*,"Deep generative AI models analyzing circulating orphan non-coding RNAs enable detection of early-stage lung cancer,"*Nat. Commun.*, vol. 15, pp. 1–12, 2024, doi: 10.1038/s41467-024-48111-2.

[18] T. Liu *et al.*,"Single-cell and spatial transcriptomics analysis reveals immunosuppressive myeloid niches in non-small cell lung cancer,"*Nat. Commun.*, vol. 15, pp. 1–18, 2024, doi: 10.1038/s41467-024-47657-1.

[19] S. Wu *et al.*,"CTpathway: a CrossTalk-based pathway enrichment analysis method for multi-omics biological interpretation,"*Genome Med.*, vol. 14, pp. 1–15, 2022, doi: 10.1186/s13073-022-01047-6.

[20] E. Kim *et al.*,"Benchmarking enrichment analysis methods with a disease pathway network to improve pathway-based interpretation,"*Brief. Bioinform.*, vol. 25, no. 1, pp. 1–14, 2024, doi: 10.1093/bib/bbad488.

[21] J. Rogers *et al.*,"Interpreting omics data with pathway enrichment analysis: concepts, applications, and future challenges,"*Trends Genet.*, vol. 39, no. 3, pp. 234–249, 2023, doi: 10.1016/j.tig.2022.12.003.

[22] Y. Wang *et al.*,"Decoding the immune microenvironment of non-small cell lung cancer using single-cell RNA sequencing,"*Front. Oncol.*, vol. 13, pp. 1–12, 2023, doi: 10.3389/fonc.2023.1124518.

[23] R. Chen *et al.*,"Identifying key genes associated with recurrence in non-small cell lung cancer through TCGA integration and single-cell analysis,"*Front. Genet.*, vol. 13, pp. 1–14, 2022, doi: 10.3389/fgene.2022.954315.

[24] K. Xu *et al.*,"Human lncRNAs NEAT1 and MALAT1 regulate the tumor microenvironment in lung cancer PDX models,"*Mol. Cancer*, vol. 22, pp. 1–15, 2023, doi: 10.1186/s12943-023-01830-5.

[25] S. Zhao *et al.*,"Ticagrelor inhibits the growth of lung adenocarcinoma by downregulating SYK expression and modulating the PI3K/AKT signaling pathway,"*Cancer Lett.*, vol. 559, pp. 215–228, 2023, doi: 10.1016/j.canlet.2023.01.002.

[26] Y. Liao *et al.*,"Weighted gene co-expression network analysis identifies functional modules related to bovine respiratory disease,"*BMC Genomics*, vol. 23, pp. 1–14, 2022, doi: 10.1186/s12864-022-08777-0.

[27] R. Marjanovic *et al.*,"Single-cell analysis of human non-small cell lung cancer lesions reveals cellular heterogeneity and immune remodeling,"*Cancer Cell*, vol. 39, no. 10, pp. 1295–1314, 2021, doi: 10.1016/j.ccell.2021.08.010.

[28] H. Qian *et al.*,"The spatial transcriptomic landscape of non-small cell lung cancer brain metastasis,"*Nat. Commun.*, vol. 13, pp. 1–12, 2022, doi: 10.1038/s41467-022-29570-9.

[29] J. Kim *et al.*,"Concurrent RNA-based NGS and DNA-based NGS detect more actionable variants in advanced non-small cell lung cancer,"*JAMA Netw. Open*, vol. 7, no. 2, pp. 1–12, 2024, doi: 10.1001/jamanetworkopen.2024.1045.

[30] S. Roy *et al.*,"Comprehensive network analysis of lung cancer biomarkers identifies key genes through integrated RNA-seq data and PPI networks,"*Front. Oncol.*, vol. 13, pp. 1–14, 2023, doi: 10.3389/fonc.2023.1123458.

221-35-982

ORIGINALITY REPORT

15%	12%	7%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	4%
2	Submitted to Midlands State University Student Paper	1%
3	www.frontiersin.org Internet Source	1%
4	core.ac.uk Internet Source	1%
5	Submitted to University of Technology, Sydney Student Paper	<1%
6	Hamid D. Ismail. "Bioinformatics of Autoimmune Diseases", CRC Press, 2026 Publication	<1%