A machine learning approach for prioritising lung carcinoma biomarkers

Bachelor Thesis

Author:	Youssef Fahim
Supervisors:	Dr.Mohamed Hamed Fahmy

Submission Date: 19 May, 2024

A machine learning approach for prioritising lung carcinoma biomarkers

Bachelor Thesis

Author:	Youssef Fahim
Supervisors:	Dr.Mohamed Hamed Fahmy

Submission Date: 19 May, 2024

This is to certify that:

(i) the thesis comprises only my original work toward the Bachelor Degree

(ii) due acknowlegement has been made in the text to all other material used

---

Youssef Fahim
19 May, 2024

Acknowledgments

I extend my gratitude to my family for their great support and encouragement throughout my thesis.

To my colleagues, thank you for the helpful discussions and collaborative spirit that improved my work.

I am grateful to my supervisor, Dr. Mohamed Hamed, whose guidance and patience were great in the completion of this thesis.

Abstract

Lung cancer, a leading cause of cancer-related mortality, necessitates innovative approaches for early detection and personalized treatment. The chance that a man will develop lung cancer in his lifetime is about 1 in 16; for a woman, the risk is about 1 in 17[8], furthermore, there are still improvements for prioritizing lung cancer biomarkers. This thesis leverages machine learning and bioinformatics to identify potential genetic biomarkers for lung cancer, utilizing RNA sequencing data from the Cancer Genome Atlas (TCGA). A differential expression analysis of 598 samples (539 cancer samples and 59 control normal tissues) revealed 515 differentially expressed genes (DEGs). Gene set enrichment analysis of these DEGs highlighted significant enrichment in pathways like neuroactive ligand-receptor interaction and diseases like thrombophilia, suggesting their different roles in lung cancer biology. Machine learning models, including Random Forest and Support Vector Machines, were able to effectively predict cancer samples, achieving 94 percent accuracy for RF and 100 percent accuracy for SVM; furthermore, variable importance analysis, along with STRING Network database, pinpointed genes such as ALPP (downregulated), Bex1 (downregulated), HSD17B6 (upregulated), and TCEAL2 (upregulated) as potential biomarkers that need further investigation. This study combines machine learning with bioinformatics analysis to spot potential lung cancer biomarkers, potentially paving the way for improved diagnostic, prognostic, and therapeutic strategies.

Contents

Chapter 1

Introduction

1.1 Cancer

1.1.1 Biology of Cancer

Cancer is characterized by the uncontrolled division of cells that can invade and destroy normal body tissue, forming tumors where the cells divide uncontrollably. With the potential to affect any part of the body, cancer is a leading cause of morbidity and mortality worldwide. The complexity of cancer is influenced by genetic, environmental, and lifestyle factors.

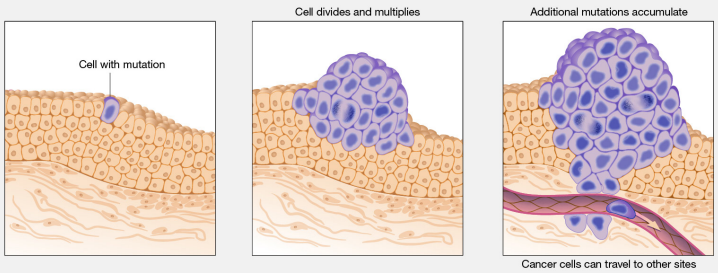


Figure 1.1: Cancer [9]

1.1.2 Lung Cancer

Lung cancer is the leading cause of cancer deaths worldwide. It can be triggered by environmental factors, such as smoking. Lung cancer cells can invade surrounding tissues

and metastasize to other organs through the bloodstream. The disease is divided into two main types: Non-Small Cell Lung Cancer (NSCLC), which includes subtypes like adenocarcinoma and squamous cell carcinoma, and Small Cell Lung Cancer (SCLC), known for its rapid growth and spread.

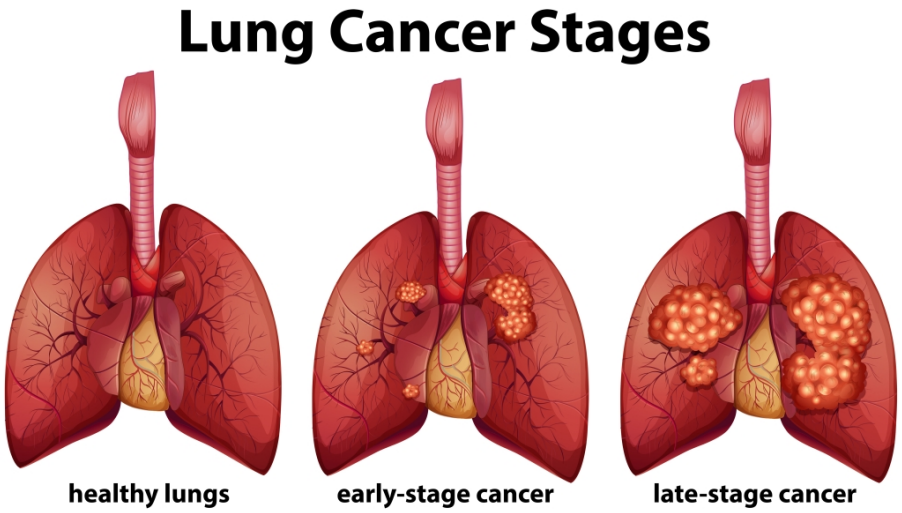


Figure 1.2: SERO, What are the stages of Lung Cancer [15]

1.2 Molecular Biomarkers

Molecular biomarkers (biological markers) are biological molecules that can be a sign of a normal or abnormal process or of a condition or disease. They can indicate various physiological states, including the presence or risk of disease, and can guide decisions on the best treatment approaches. Molecular biomarkers include DNA, RNA, proteins, and metabolites.

1.3 Omics

Omics refers to measuring biological molecules in order to analyze the roles, relationships, and actions of the various types of molecules that make up the cells of an organism.

1.3.1 Omics' layers

Genomics

the study of the whole genome, the complete sequence of DNA in a cell or organism. Genomics are used to identify genetic predispositions to diseases in order to understand and find cures for them.

Transcriptomics

The transcriptome is the complete set of RNA transcripts from DNA in a cell or tissue. Transcriptomics are used to show gene expression levels under certain conditions, assisting in understanding disease mechanisms.

Proteomics

The proteome is the complete set of proteins expressed by a cell, tissue, or organism. Proteomics are used to analyze the expression level of proteins to identify disease diagnosis, prognosis, and the development of personalized medicine approaches.

Epigenomics

The epigenome consists of reversible chemical modifications to the DNA, or to the histones that bind DNA. Epigenomics are used to identify epigenetic markers for diseases applied in cancer research and investigating environmental effects on gene expression.

metabolomics

The metabolome consists of reversible chemical modifications to the DNA or to the histones that bind DNA. In agriculture, metabolomics allows us to enhance genetically modified plants.

In personalized medicine, biochemical tests are used to measure individual metabolite concentrations to identify disease states.[1]

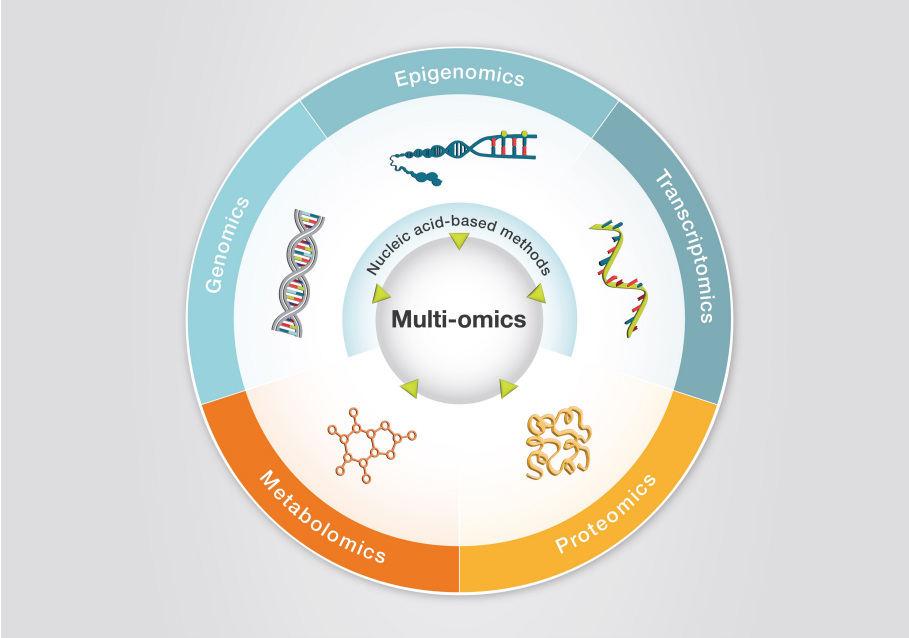


Figure 1.3: A Commonality between Various Omes in Multi-omics Approaches. [3]

1.4 Machine Learning

Machine learning is a branch of artificial intelligence that allows computers to learn from data, identify patterns, and make decisions. It involves algorithms that improve through experience.

1.4.1 Machine Learning Types

Supervised Learning

It involves training computers using labeled datasets, inputs, and outputs in order to predict outcomes and recognize patterns.

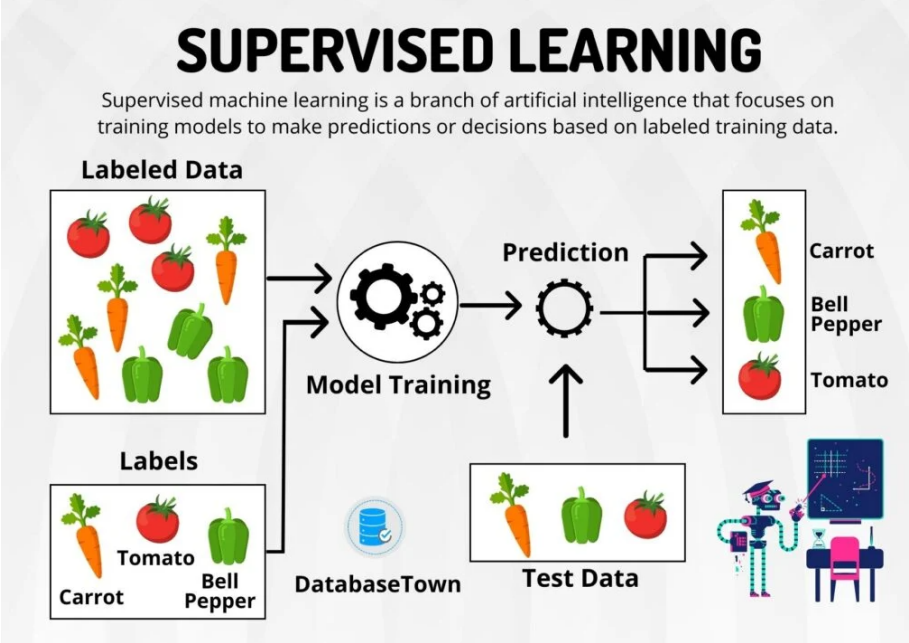


Figure 1.4: Supervised Learning. [13]

Unsupervised Learning

It involves learning from data without human intervention, it is given unlabeled data to discover patterns without any guidance.

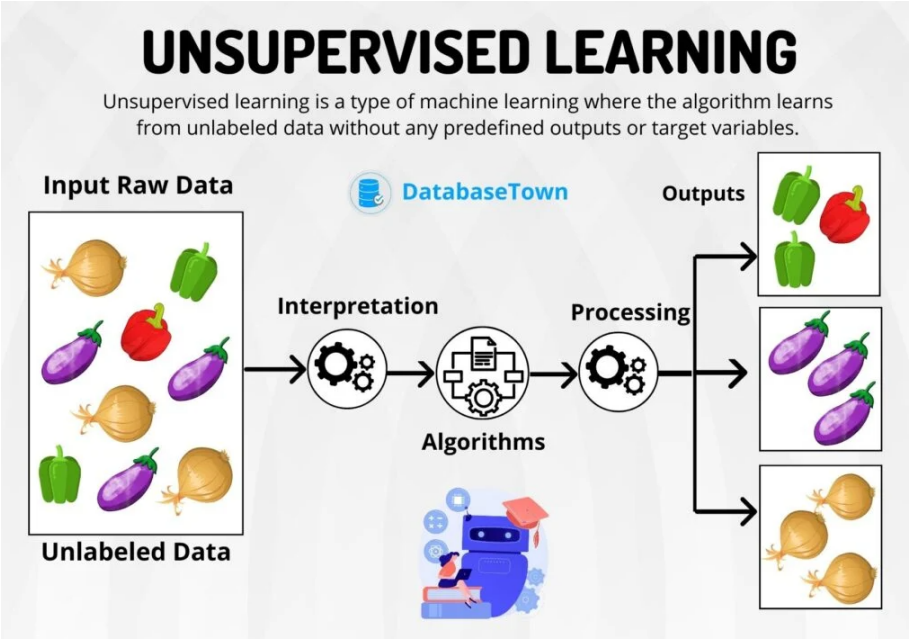


Figure 1.5: Unsupervised Learning. [14]

1.5 Aim and motivation of the thesis

1 in 16 people will be diagnosed with lung cancer in their lifetime—1 in 16 men and 1 in 17 women[8]. Traditional diagnostic techniques often detect lung cancer at later stages, reducing the effectiveness of treatment.

Machine learning is a great approach for discovering biomarkers that could enable earlier detection and personalized treatment plans, significantly improving patient outcomes.

This thesis aims to leverage the power of machine learning to identify potential genetic biomarkers for lung cancer, contributing to the advancement of lung cancer diagnosis.

1.6 Structure of this Thesis

This thesis is structured in 4 chapters

Background

In the background section, we review previous studies related to our scope regarding potential lung cancer biomarkers

Methodology

In the methodology section, We introduce the data analysis pipeline, the tools used to implement our research into lung cancer potential biomarkers, the visualization techniques used to evaluate the data analysis results, and finally the machine learning models used to predict new cases whether they are normal or cancer patients.

Results

In the results section, we review the exploratory analysis done on the data, differential expression analysis results, visualizations that were made to review the potential biomarkers for lung cancer, STRING network analysis results, gene set enrichment analysis results, machine learning results, and finally the feature importance extracted from the machine learning models.

Discussion

In the discussion section, we give off a brief review of the whole thesis, conclude the work done, introduce the potential biomarkers detected during the thesis and introduce our future works suggestions for future studies regarding lung cancer potential biomarkers.

Chapter 2

Background

2.1 Literature Review

A study done by FangweiWang et al. (2022)[32] aimed to provide a gene expression analysis in order to find novel genetic biomarkers that were most associated with non-small lung cancer (NSCLC). In the study, a total of 371 DEGs (165 up-regulated genes and 206 down-regulated genes) were identified, and enrichment analysis revealed that these DEGs might be linked to the development and progression of NSCLC. This is a study that highlights the potential genes as they provide results. ABCA8, ADAMTS8, ASPA, CEP55, FHL1, PYCR1, RAMP3, and TPX2 genes were identified as novel diagnostic biomarkers for NSCLC

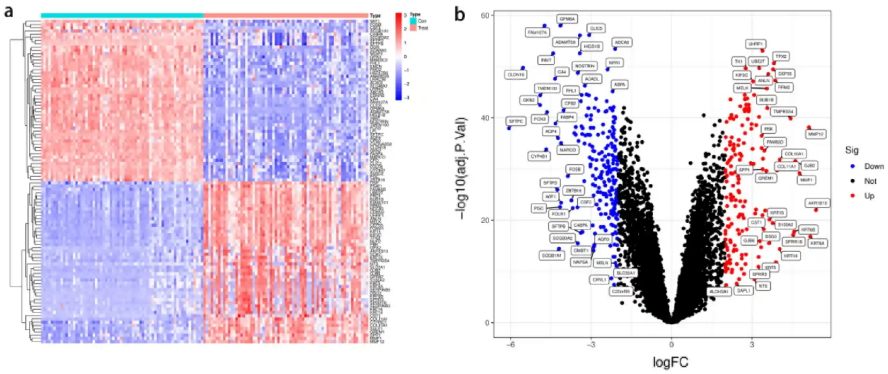


Figure 2.1: (a) Heatmap of DEGs. (b) Volcano diagram of DEGs, red indicates up-regulated, blue indicates down-regulated.[32]

another study done by Abel Sanchez-Palencia et al. (2010)[29] aimed to determine any correlation between the phenotypic heterogeneity and genetic diversity of lung cancer. The expression level of 92 selected genes was validated. High upregulation was observed for KRT15 and PKP1, which may be good markers to distinguish squamous-cell carcinoma samples. High downregulation was observed for DSG3 in stage IA adenocarcinomas.

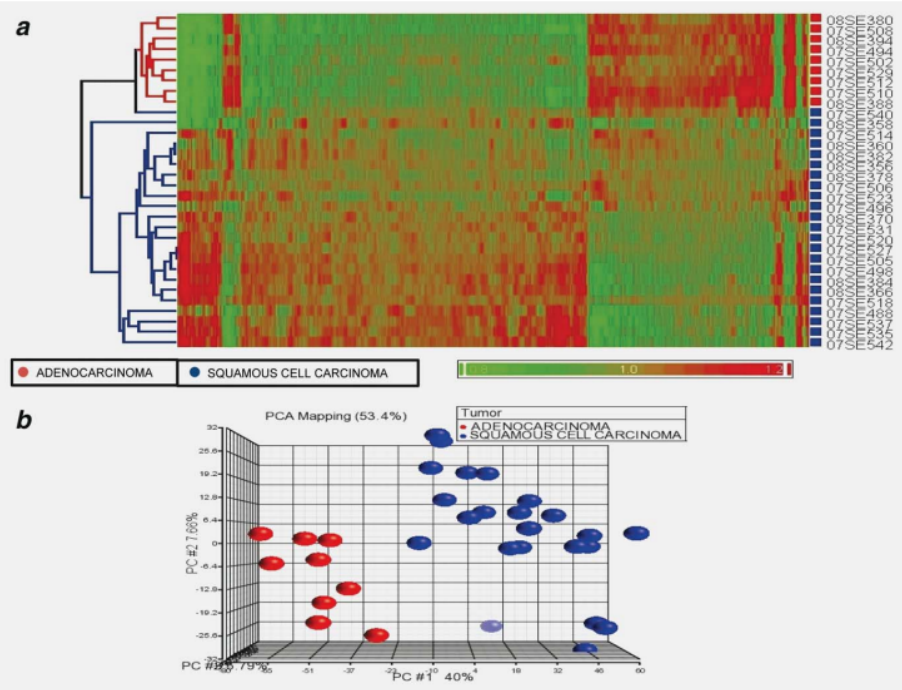


Figure 2.2: (a) Unsupervised hierarchical clustering and (b) PCA of the differentially expressed sequences for the comparison between Stage IB adenocarcinoma and Stage IB squamous-cell carcinoma samples[29]

Another study done by R. Rosell et al. (2013)[28] aimed to personalize lung cancer treatment through the identification of genetic and biomarker characteristics. Mutations in the epidermal growth factor receptor (EGFR), a protein on the surface of cells, have been identified as a primary oncogenic event, a change or occurrence in a cell that leads to the development of cancer, in lung adenocarcinomas. EGFR tyrosine kinase inhibitors showed remission in 60 percent of patients, though responses were temporary due to resistance mechanisms like the pre-existing EGFR Thr790Met mutation, a mutation that makes the cancer cells less sensitive to the inhibitors.

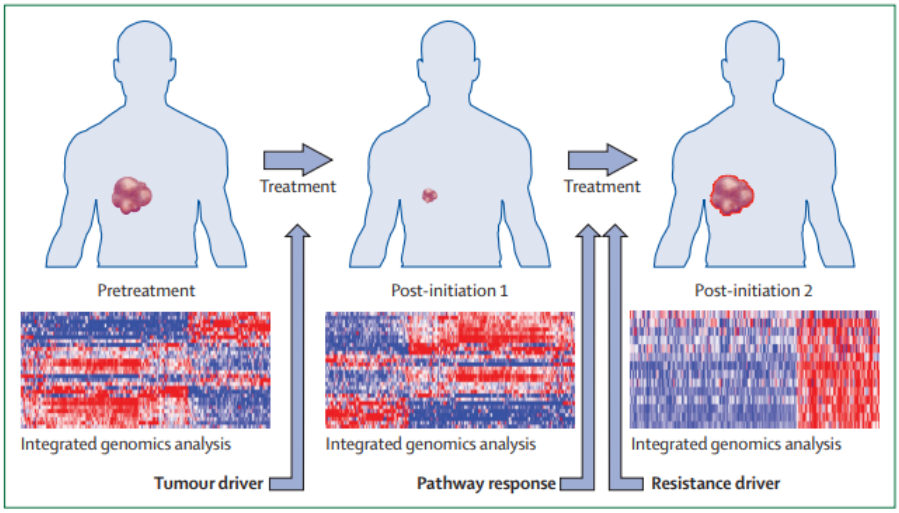
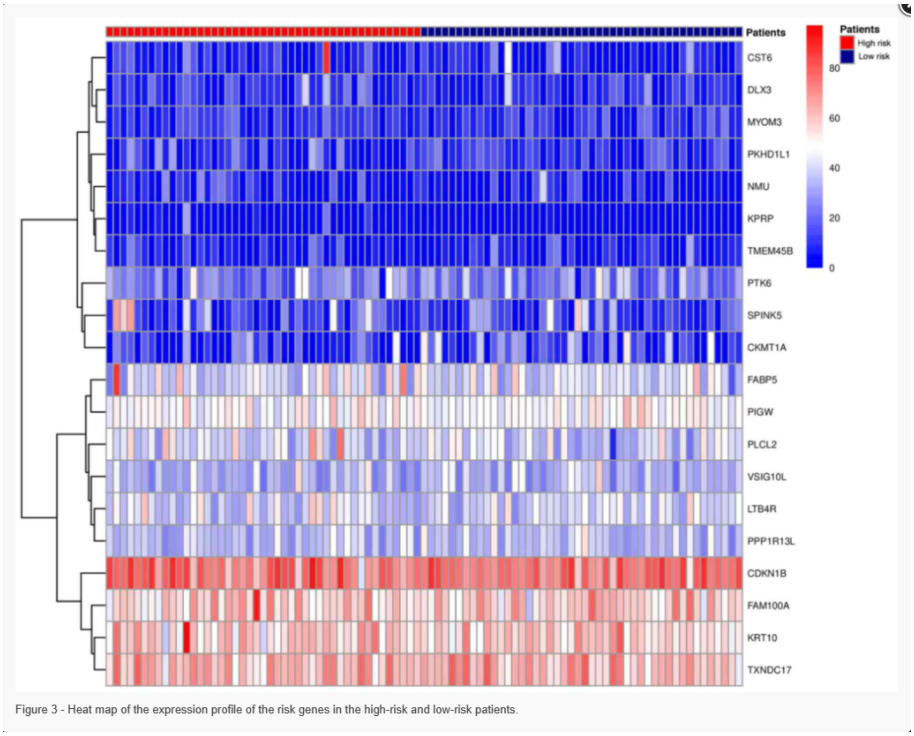


Figure 2.3: Integrated genomic analysis to identify biomarkers of drug response and personalise treatment of patients with non-small-cell lung cancer[28]

Another study done by Liu et al. (2020)[27] aimed to identify prognostic gene biomarkers for metastatic skin cancer using data mining. The least absolute shrinkage and selection operator regression analysis method was used with 10-fold cross verification to reduce the dimensions of gene data in patients with skin cancer, and subsequently, 20 gene biomarkers were screened. Among the 20 prognostic genes identified in this study, several are closely associated with skin cancer and metastatic skin cancer, including DLX3, PTK6 and CST6 genes.



Chapter 3

Methodology

3.1 Biological Repositories

Biological repositories are collections of different types of biological data. These biological repositories were used in order to gather the open-access data needed for this thesis.

3.1.1 The Cancer Genome Atlas (TCGA)

A cancer genomics program offered by the National Cancer Institute (NCI) molecularly characterized over 20,000 primary cancers and matched normal samples spanning 33 cancer types.[2]

For this thesis, The Cancer Genome Atlas (TCGA) was used to download open-access RNASeq genomic profiling data for lung tissue samples, both for normal healthy tissues and Lung Adenocarcinoma (LUAD), which contains gene expression level data for 19963 genes per sample.

3.2 Description of Dataset

A total of 598 samples were downloaded for this study, 59 of which were control normal tissue samples, and the remaining 539 were LUAD tumor samples. Each sample contains gene expression level read counts for 19963 different genes.

LUAD Pheno Table, A descriptive file of the data, was also downloaded to gain further insights.

3.3 Data pre-processing

The data from TCGA was already ready to use, where FastQ files were transformed into raw count expression data for every gene, furthermore, no extra pre-processing steps were needed. Consequentially, the data was loaded directly into the development environment to be directly used.

3.3.1 Filtration

Genes that were detected as non-significant were filtered out of the dataset.

Genes that were found to have near-zero or zero read count values across all samples were removed from the study, leaving us with 19133 genes to work with.

3.3.2 Normalization

Data normalization is the organization of data to appear similar across all records and fields[33], as normally not all features span the same range of counts, since some are long and some are short.

For genes, normalization of the data was done by the variance stabilizing transform (VST) using the DESeq2 R-project package.

3.4 Development Environment

3.4.1 R Project

R is a programming language for statistical computing and data visualization. It has been adopted in the fields of data mining, bioinformatics, and data analysis.[17]

R Project was the chosen programming language for this thesis to manipulate the downloaded data from TCGA.

3.4.2 Bioconductor

Bioconductor is an open-source project that provides tools for the analysis and comprehension of genomic data.

Bioconductor offers helpful packages for the scope of this thesis. The package that will be extensively used in this thesis is the DESeq2 package[16] package.

3.5 Analysis pipeline

3.5.1 Exploratory Analysis

Exploratory Data Analysis (EDA) is a fundamental step in bioinformatics analysis, providing a basis for the initial examination of datasets for many reasons, like identifying patterns, spotting anomalies, confirming hypotheses, and checking assumptions.

In this thesis, visualization analysis will be the main objective, as it will clearly show the patterns exhibited by the samples and genes.

3.5.2 Differential Expression Analysis

Differential Expression Analysis (DEA) is a statistical method used to compare gene expression levels across different conditions. It is a fundamental analysis in the field of bioinformatics, especially in studies concerning diseases like cancer, where understanding which genes are upregulated (increased expression) or downregulated (decreased expression) will provide insights for potential biomarkers.

Differential expression analysis will be performed using the DESeq2 package in R. It normalizes the data and calculates significant numerical values for each gene based on the raw read counts of the samples. These numerical values will then be applied to the computed numerical values in order to filter the genes according to specific constraints. As a result, the most statistically significant genes will be listed as a list of differentially expressed genes. Below is an explanation of the numerical values that will be used in the differential expression analysis.

Fold Change

Fold change is calculated simply as the ratio between the expression level in the second condition (B) and the expression level in the first condition (A). This is written as B/A.

For a given comparison, a positive fold change value indicates an increase in expression, while a negative fold change indicates a decrease in expression.

P value

It tells you how likely it is that the observed difference is due to chance, whether the gene analyzed is likely to be differentially expressed or not.

Low p-value (typically less than 0.05): This suggests the observed change is likely due to a biological effect. High p-value (typically greater than 0.05): The observed change might not be biologically relevant.

Adjusted p value

The p-value obtained for each gene above is recalculated to correct for running many statistical tests (as many as the number of genes). As a result, we can say that all genes with an adjusted p-value less than 0.05 are significantly differentially expressed.[4]

3.6 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA)[6] It is a statistical technique used to identify groups of genes that are differentially expressed in a biological experiment. It shifts the

focus from analyzing individual genes to analyzing groups of genes with known biological functions.

In this stage, every gene is run through Enrichr[5], a comprehensive gene set enrichment analysis web server, in order to identify the functions and diseases that are most associated with the recorded DEGs.

3.7 String Network Analysis

STRING[12] (Search Tool for the Retrieval of Interacting Genes and Proteins) provides a network analysis of proteins and their interactions. It integrates information from various sources to predict functional associations between proteins across organisms.

The top 100 differentially expressed genes (DEGs) were used to generate a STRING network in order to investigate the associations between them.

3.8 Machine Learning Models

3.8.1 Machine Learning Models for Predictive Analysis

Machine Learning (ML) has been increasingly used in biomedical research to enhance the prediction and classification of disease outcomes based on any data. In this thesis, ML models will be applied to classify lung cancer samples, whether they are cancer or tumor, using DEGs identified in the previous sections as features.

3.8.2 Model Selection

Several machine learning models will be evaluated for their accuracy in classifying whether the patient has lung cancer or not. The models to be tested include:

Random Forest

Random Forest (RF)[11] is An ensemble, refers to combining multiple models in order to improve performance, learning method for classification and regression that operates by constructing multiple of decision trees at training time. the output of the random forest is the class selected by most trees.

Support Vector Machines (SVM)

RF[11] is An ensemble refers to combining multiple models in order to improve performance. It is a learning method for classification and regression that operates by constructing multiple decision trees at training time. The output of the random forest is the class selected by most trees.

Logistic Regression

Logistic Regression (LR)[7] estimates the probability of a binary event (for example, cancer or normal) occurring and models the decision boundary of this binary event using a logistic function.

Model performance will be assessed using standard metrics such as accuracy, precision, F1 score, and recall.

3.8.3 Model Training and Validation

The selected models will be trained using a split of the dataset into training (70 percent) and testing (30 percent) sets to ensure that the models generalize well to new data. Confusion Matrix and receiver operating characteristic curve (ROC) will be used to provide insights into the model's performance across different subsets of the dataset.

3.9 Variable Importance

Variable importance shows us the most important features of the model that affect it the most. It is a crucial step after completing the machine learning step, which would allow us to dig deeper into lung cancer potential biomarkers, precisely identifying them.

Chapter 4

Results

4.1 Computational pipeline

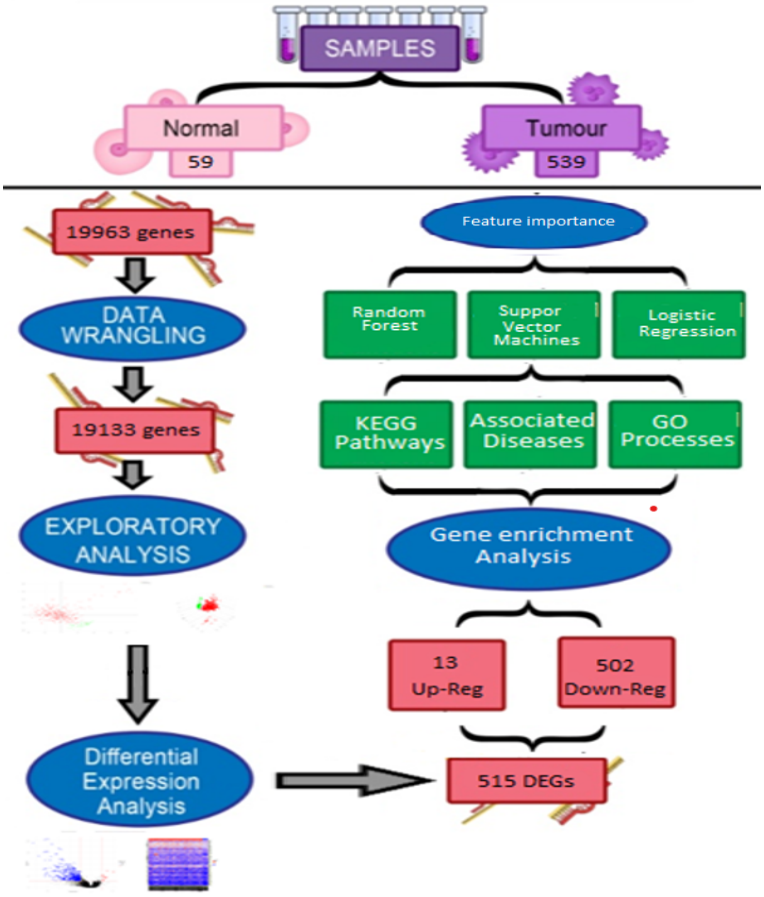


Figure 4.1: Computational Pipeline

In figure 4.1, The computational pipeline is showed, we started our analysis with 19963 genes, which is basically the human genome. after performing some data wrangling like removing duplicates, zero, and constant columns, we ended up with 19133 genes which were up for exploratory analysis. After performing PCA on the data, differential expression analysis was performed on the data. We ended up with 515 DEGs, 15 upregulated and 502 downregulated. Gene set enrichment analysis was performed, moreover, we trained 3 machine learning models on the DEGs to predict new cases. Finally, most important features were extracted from the models.

4.2 Exploratory Analysis and Sample visusalization

After performing data preprocessing, the data downloaded from TCGA was explored by means of visualization using various plots and analyses, such as the Principle Component Analysis, Volcano Plot, and Heat Map. The results and findings of each visualization step are discussed in the following sections.

Principal Component Analysis

The first step was performing Principal Component Analysis (PCA) for all 598 samples involved in the study. PCA is a statistical technique used to simplify the complexity of high-dimensional data while retaining trends.

The PCA (Fig 4.2) shows that there is a clear distinction in the count values between tumor samples and normal tissue samples, as the tumor cluster and the normal tissue clusters are clearly and mostly separated from each other.

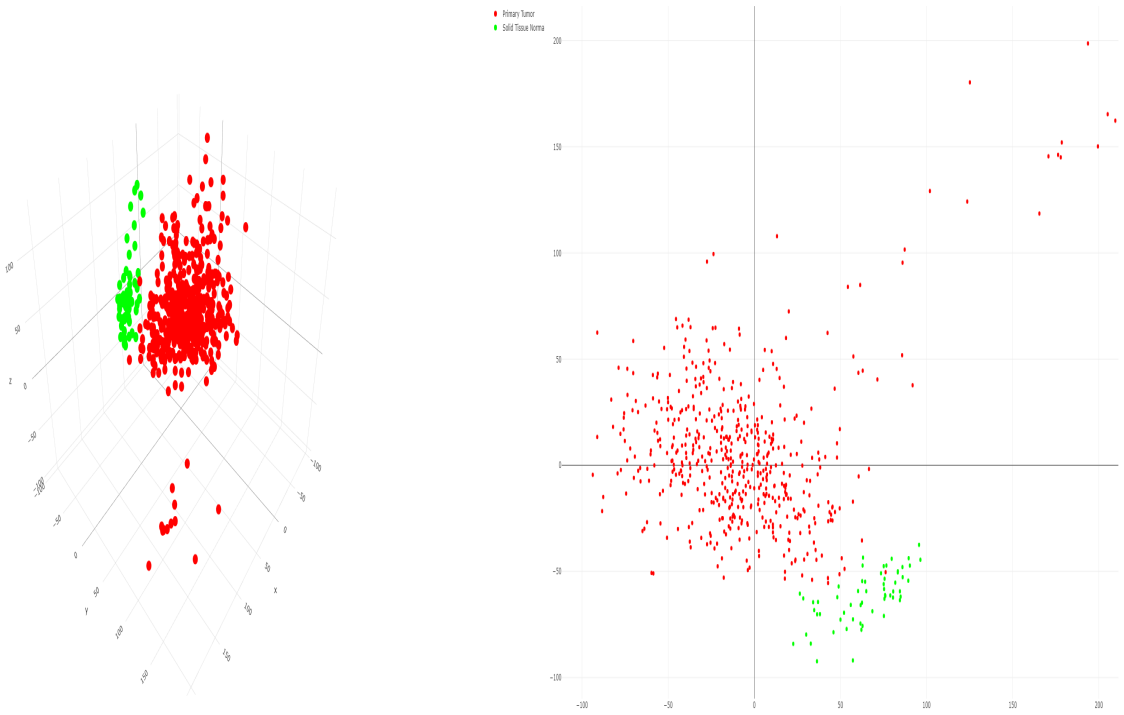


Figure 4.2: Data Pipeline

4.3 Downstream Analysis

After the initial review of plots during the exploratory analysis stage and the confirmation of the study's significance, the next phase involves conducting downstream analysis. This process aimed to identify the most statistically significant features among the 19,133 genes examined. Additionally, create plots for a clearer visualization of the data under analysis. The stage began with the differential expression analysis of all genes, utilizing the computational methods provided by the DESeq2 package. Next, the visualization of the data was conducted, focusing on the Differentially Expressed Genes (DEGs) that were identified.

4.3.1 Differential Expression Analysis

The raw count data available to us was run through the DESeq2 package in order to perform the differential expression analysis discussed in the methodology section. The calculated p-value, FDR, and LFC values for each miRNA were then used as constraints in order to filter out the most statistically significant miRNAs. In this study, the accepted range of constraints was as follows: The adjusted p-value must be 0.05 or less, and the log2 Fold Change (LFC) must be between 1 and -1.

Running the genes through these constraints resulted in 515 differentially expressed genes, Those DEGs are classified into two types: up-regulated, and down-regulated, which means that the expression level of the genes in tumor samples relative to the normal tissue samples are expressed in either higher or lower quantities, respectively. Table 4.3 shows the relevant statistical data acquired, as well as the classification of each identified gene.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
OLFM4	144.5530756	-4.223486	0.4475247	-9.437436	3.820113e-21	3.654511e-17
POU3F2	96.6433931	-3.968068	0.4179893	-9.493228	2.239878e-21	3.654511e-17
ZIC1	37.6872426	-5.308957	0.5847617	-9.078838	1.097406e-19	6.998887e-16
PCSK1	786.3013306	-3.585360	0.4012701	-8.935030	4.070665e-19	1.947101e-15
ABCC2	722.3426109	-2.952174	0.3486467	-8.467522	2.506689e-17	9.592096e-14
NPC1L1	158.2675258	-2.746604	0.3288623	-8.351837	6.720949e-17	2.143199e-13
ASCL1	869.2901194	-3.973827	0.4921286	-8.074773	6.760254e-16	1.760014e-12
ONECUT3	23.3385141	-4.135171	0.5127680	-8.064409	7.359070e-16	1.760014e-12
KRT20	99.2309432	-4.299977	0.5362281	-8.018933	1.066677e-15	2.267637e-12
LBP	97.6545509	-3.182281	0.4044588	-7.867998	3.603621e-15	6.894807e-12
CHRNA9	140.1404450	-3.194658	0.4112946	-7.767323	8.016250e-15	1.394317e-11
REG1A	51.5984865	-4.798231	0.6381317	-7.519186	5.511842e-14	8.788173e-11
ZFP42	75.1863223	-3.677587	0.4929573	-7.460255	8.635520e-14	1.270949e-10
ERVH48-1	106.1737048	-2.637710	0.3666396	-7.194285	6.278872e-13	8.580975e-10
H1-5	23.3611388	-2.909731	0.4088065	-7.117623	1.098040e-12	1.400586e-09
H4C3	14.5491642	-2.688649	0.3805578	-7.065023	1.605899e-12	1.920354e-09
PCDH8	16.7796541	-3.768289	0.5389268	-6.992208	2.705930e-12	3.045445e-09
CGA	119.0430588	-4.210082	0.6033278	-6.978101	2.991976e-12	3.180304e-09
GAL	108.7437287	-2.481063	0.3568025	-6.953601	3.560768e-12	3.442365e-09
INSM1	44.9200926	-2.662305	0.3829486	-6.952121	3.598353e-12	3.442365e-09

Figure 4.3: Top 20 differential expressed genes

After the Differential Expression Analysis (DEA) step, the data can be visualized in various ways for ease of observation.

4.3.2 Visualization

Volcano Plot

As explained in the methodology, the methods of computation used in the downstream analysis stage were: p-value, adjusted P-value, and log fold change. Each gene carries one of each of those values, and so, for all genes, two of those values were used in order to plot a volcano plot: LFG and adjusted p-value (see Fig 4.4). The purpose of this plot is that it serves as one way to visualize the DEGs acquired; all 19133 genes are plotted, but only the ones exhibiting an absolute LFC between 1 and -1 and an adjusted p-value of less than 0.05 were highlighted. The highlighted genes are the 515 identified DEGS.

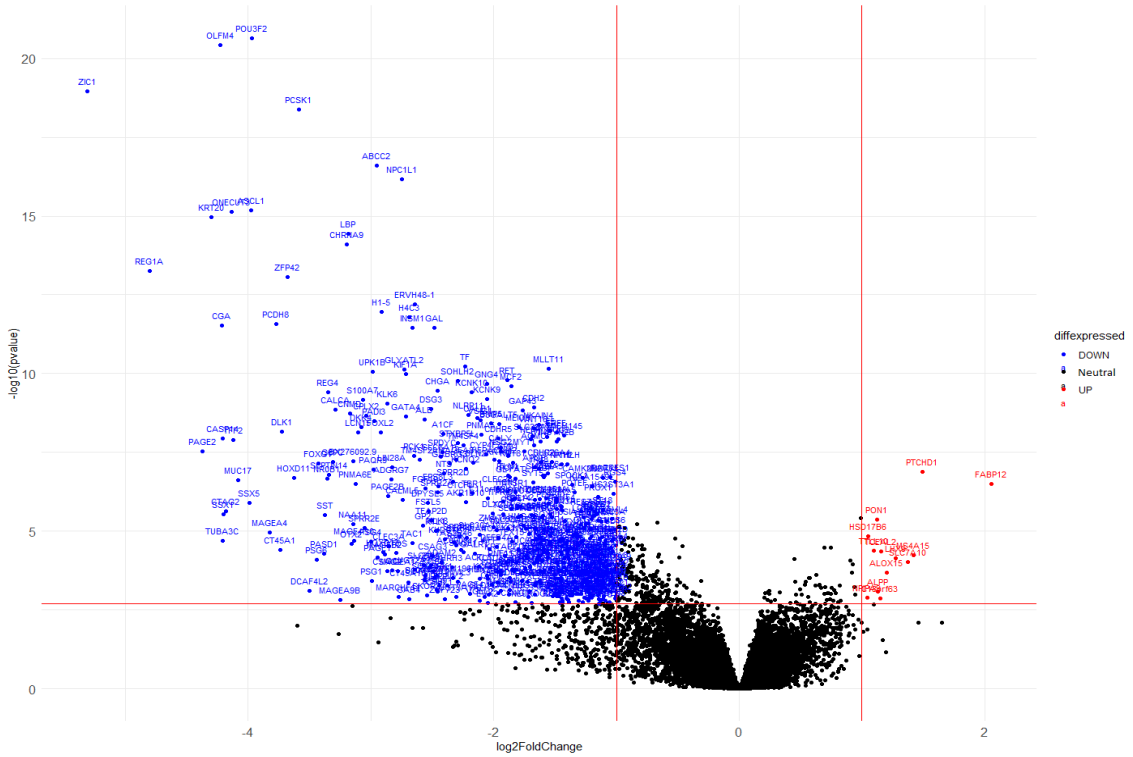


Figure 4.4: Volcano plot visualizing DEGs

Heatmap

Heatmaps are used to visualize and interpret the relative levels of expression across a dataset, enabling the identification of patterns and clusters in the data. They are useful for presenting large datasets in an easily interpretable form.

A heatmap was used to show the gene expression level of the top 100 significant genes identified by our Analysis. The result is an overall pattern that shows a significant change in the expression level of each gene between both types of samples (Fig 4.5), confirming the conclusions derived from the already displayed visualizations and the results of the DEA.

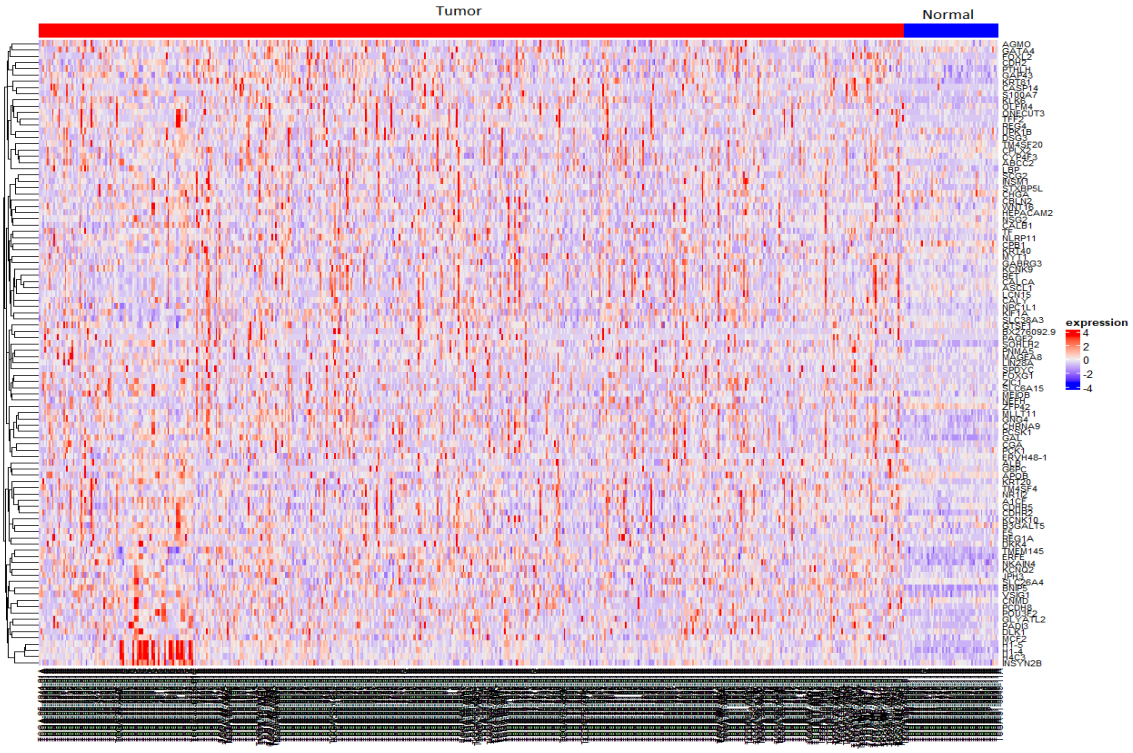


Figure 4.5: Heatmap visualizing the top 100 DEGs

The last step of the downstream analysis pipeline is performing gene set enrichment analysis (GSEA) in order to identify associated diseases and cell functions.

4.4 Gene Set Enrichment Analysis

DEGs were run through Enrichr with the purpose of identifying the functions, diseases, and processes that are potentially associated with the DEGs. The top 10 ranked kegg pathways, GO processes, and OMIIM diseases based on p-values were selected and visualized.

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways

We examined the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways bar plot (Figure 4.7), which provided insights into the involvement of our gene set in various biological pathways (biological pathways refer to a series of actions among molecules in a cell that leads to a certain product or a change in the cell) to identify specific pathways that were significantly enriched with our genes.

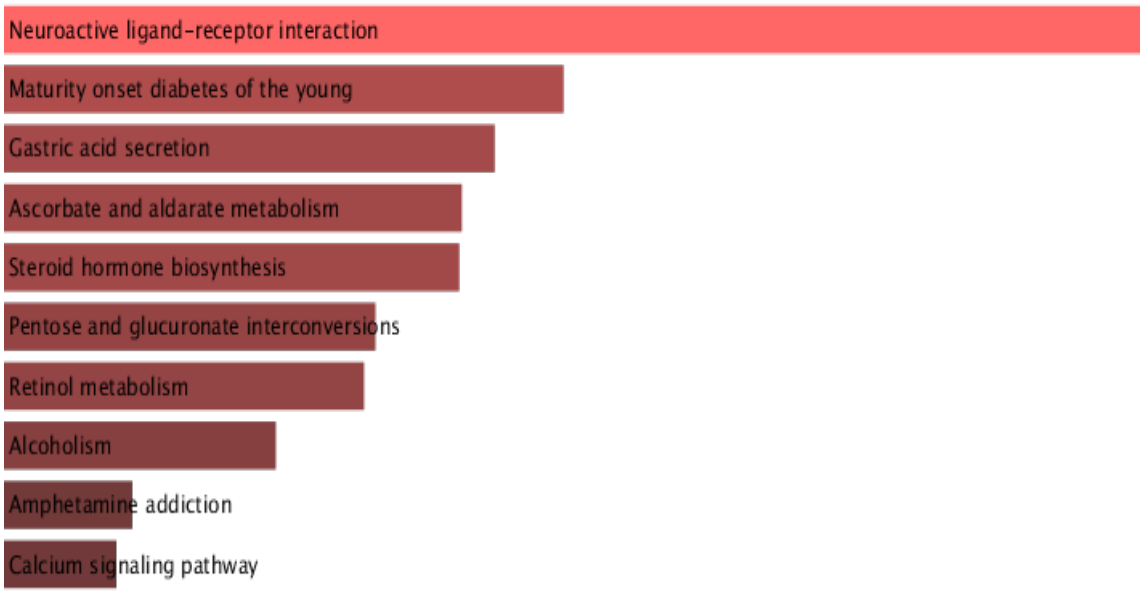


Figure 4.6: Enriched KEGG pathways analysis of the differentially expressed genes (DEGs)

Upon examining the KEGG pathway bar plot, the neuroactive ligand-receptor interaction pathway exhibited the highest level of enrichment among our DEGs. This finding suggests that our genes may be fundamental for transmitting signals across the nervous system, affecting everything from muscle movement to mood and cognition. The neuroactive ligand-receptor interaction signaling pathway was found to be associated with the progression of bladder cancer and renal cell carcinoma in a previous study[24].

Additionally, maturity-onset diabetes of the young (MODY) was found to be the second highest enriched pathway among our DEGs. MODY is a type of monogenic (controlled by a single gene) diabetes first described as a mild and asymptomatic (showing no symptoms) form of diabetes that was observed in non-obese children, adolescents, and young adults. MODY was found to be potentially related to pancreatic cancer in a previous study[20].

The third highest enriched pathway among our DEGs is gastric acid secretion. This pathway details the biological processes involved in the secretion of gastric acid in the stomach. This finding suggests that lung cancer and gastric acid secretion might share common risk factors such as smoking. A previous study suggests that smoking might be associated with increased gastric acid secretion[26].

In addition to these findings, several other pathways showed enrichment: ascorbate and aldarate metabolism, steroid hormone biosynthesis, pentose and glucuronate interconversions, retinol metabolism, and alcoholism. Although not extensively discussed in this context, their presence among the enriched pathways highlights their relevance to the understanding of our gene set's functionality. Further research into these pathways may provide additional insights into the roles of our DEGs.

Online Mendelian Inheritance in Man (OMIM)

We explored the Online Mendelian Inheritance in Man (OMIM) Disease bar plot(Figure 4.8), which highlights potential associations between our genes and disease conditions.

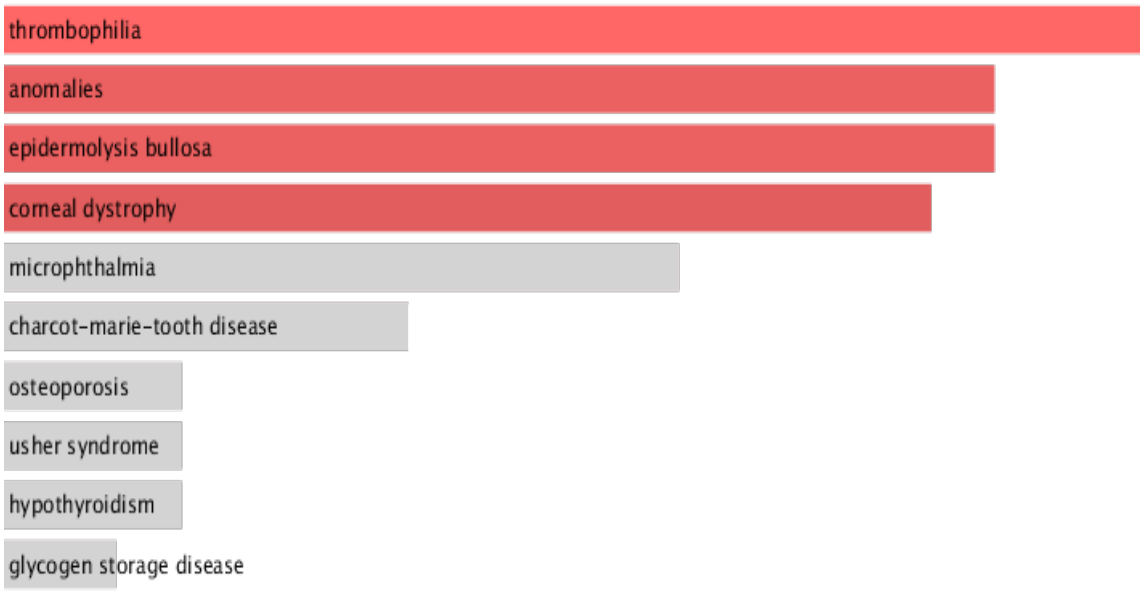


Figure 4.7: Enriched KEGG pathways analysis of the differentially expressed genes (DEGs)

The most enriched disease is thrombophilia, which implies that your blood can form clots easily. This finding suggests that thrombosis might be related to patients with lung cancer. A previous study found a relationship between thrombophilia and patients with cancer[22] suggesting that thrombosis might be a complication of cancer in the body.

Furthermore, anomalies refer to mutations, rearrangements, or alterations in the DNA that can contribute to the initiation and progression of cancer. This finding suggests that patients with lung cancer might have their DNA altered in comparison with normal people.

Additionally, the third most enriched disease, epidermolysis bullosa (EB), is a rare skin condition that causes fragile, blistering skin. This finding suggests that EB might be related to lung cancer in any way. A study previously found a relationship between EB patients and those who also have squamous cell cancer[19], where squamous cell cancer arises more aggressively in EB patients.

The fourth most enriched disease, corneal dystrophy, It refers to the buildup of material in one or more layers of the cornea, which may cause the cornea to lose its transparency, thus causing a loss of vision or blurred vision. This finding suggests that there might be a relationship between lung cancer and the cornea. A previous study[34] found

that patients with Fuchs Endothelial Corneal Dystrophy (FECD) aged 65 years or older may be at increased risk for cancer.

Even though the remaining diseases in the OMIM disease bar plot were not enriched enough, their presence among the enriched diseases should not be ignored. These diseases may still have potential connections to our DEGs. Further investigation and analysis are required to find the roles of our genes in these disease conditions.

Gene Ontology Processes (GO)

We examined the gene ontology of our DEGs. GO describes our knowledge of the biological world with respect to three aspects: molecular function, cellular location, and biological process. Biological process will be the focus of this analysis.

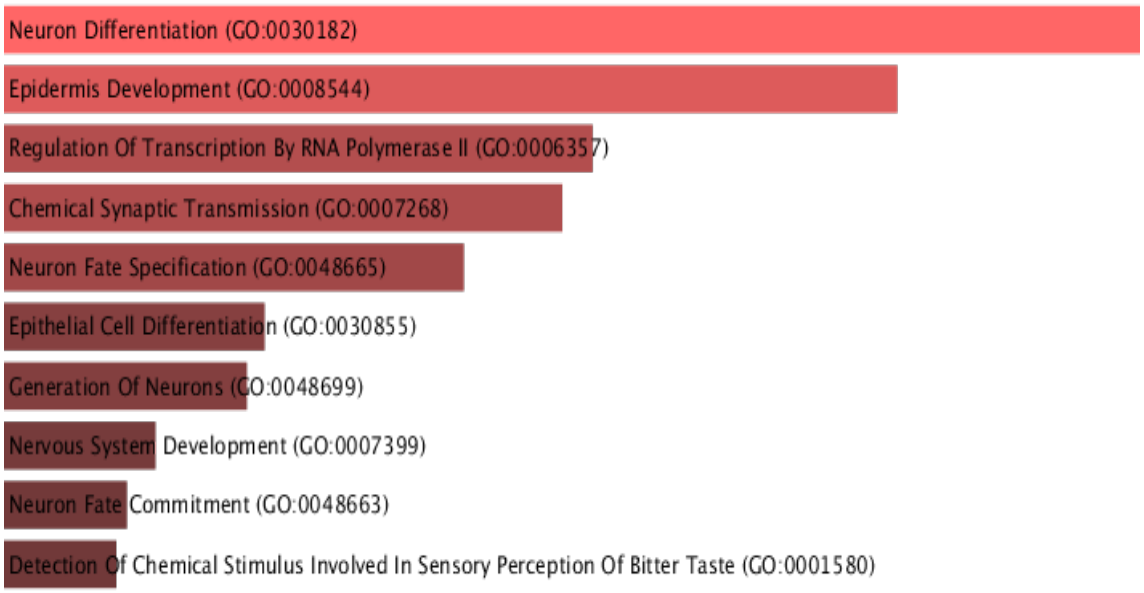


Figure 4.8: Enriched Gene Ontology analysis of the differentially expressed genes (DEGs)

Neuron Differentiation: The process in which a stem cell acquires specialized features of a neuron[10] is the most enriched GO process. This finding suggests that our DEGs might be associated with stem cell specialization.

The second most enriched GO process is epidermis development, which is responsible for the progression of the epidermis (the top layer of skin in your body). This finding suggests that our DEGs might be correlated with skin-related biological processes.

The third most enriched GO process is the regulation of transcription (copying a segment of DNA into RNA) by RNA polymerase II, which is any process that modulates the frequency of transcription made by RNA polymerase II (RNA polymerase is an enzyme that is responsible for copying a DNA sequence into an RNA sequence). This finding suggests that our DEGs might be associated with the transcription process in the body.

Additionally, the remaining enriched GO processes should not be ignored. These processes may still have potential connections to our DEGs. Further investigation and analysis are required to find the roles of our genes in these processes.

4.5 String Network Analysis

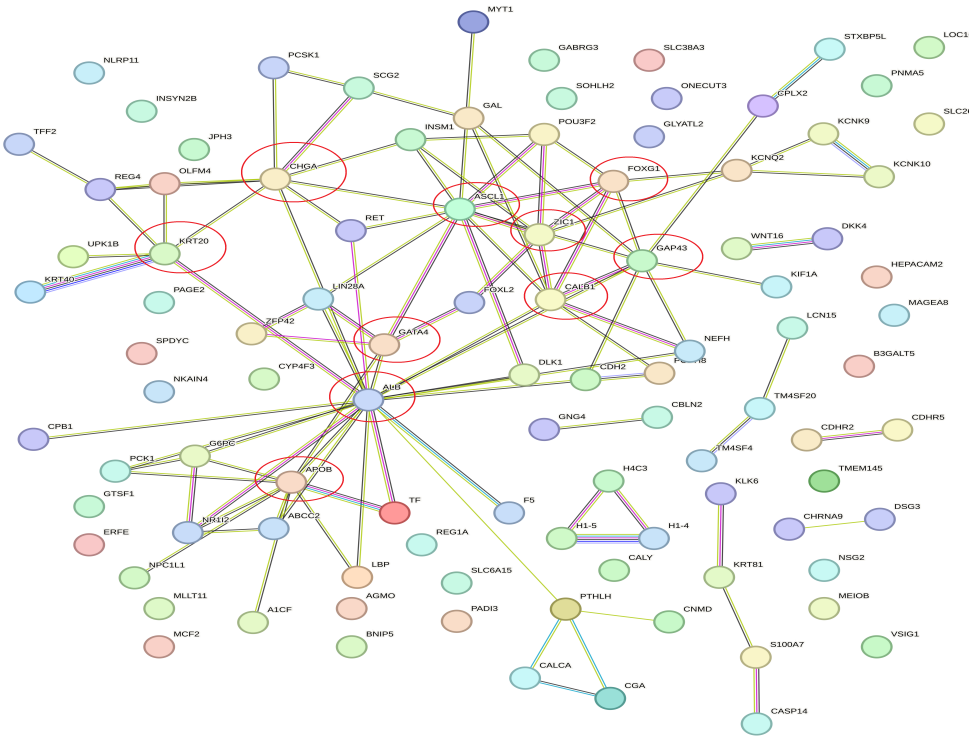


Figure 4.9: String Network of the top 100 DEGs(top 10 highlighted)

Gene (Node)	Degree
ALB	20
ASCL1	12
APOB	10
CHGA	9
GAP43	9
CALB1	8
ZIC1	7
FOXG1	6
GATA4	6
KRT20	6

Table 4.1: Gene Interaction Degree of centrality

Displayed in the table are the top 10 genes out of our top 100 DEGs sorted by degree of centrality. Those genes will be further investigated after training our models, and the Feature importance along with the STRING network will be used to form a final importance table.

4.6 Machine Learning

The final stage of this thesis is analyzing our data, retrieving the DEGs, performing needed exploratory analysis and visualizations, performing string network analysis, and performing gene set enrichment analysis. All we need now is to take a further step to help us in the future with diagnosing new cancer patients. That is where machine learning comes into action. The 3 models discussed before (random forests, support vector machines, and logistic regression) were used to train 3 different models in order to predict whether a patient has lung cancer or not.

4.6.1 Results

The metrics we are going to discuss initially are accuracy, precision, recall, and the F1 score of the models' predictions. These are the most basic and straightforward measurements of a model's performance.

From the first look in figure 4.10, it looks like SVM performed the best while logistic regression performed the worst. However, let's discuss the results further in order to understand what is actually going on.

4.6.2 Model Evaluation Metrics

Accuracy

Accuracy is basically the ratio between the correct predictions and the total number of samples we have. SVM scored the highest accuracy of 100 percent, RF scored around 94 percent, and LR scored around 68 percent. Accuracy alone is not informative enough about the models' performance; therefore, we will dig deeper into the metrics.

Precision

Precision basically shows how often an ML model is correct when predicting the target class (cancer patients in our case), or, in other words, the ratio of true positive predicted cases to the true positive predicted cases added to the negative positive predicted cases. Different from accuracy, SVM scored the highest precision of 100 percent, followed by LR of 97 percent, followed by RF of 94 percent. This result suggests that even though logistic regression resulted in lower accuracy than the other two models, out of the positively predicted cases, 97 percent of them were actually correctly predicted.

Recall

Recall shows whether an ML model can find all objects of the target class, or, in other words, the ratio of predicted true positives to the ratio of predicted true positives added to the predicted false positives. Close to accuracy, SVM and RF scored 100 percent while LR scored 62 percent. This result suggests that the lung cancer predictions by both SVM and RF were all correctly predicted; however, LR correctly predicted only 62 percent of the cases as cancer.

F1 Score

F1 score is the Harmonic mean of both Recall and Precision, It balances both metrics.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

similar to Accuracy, SVM scored the highest followed by RF followed by LR.

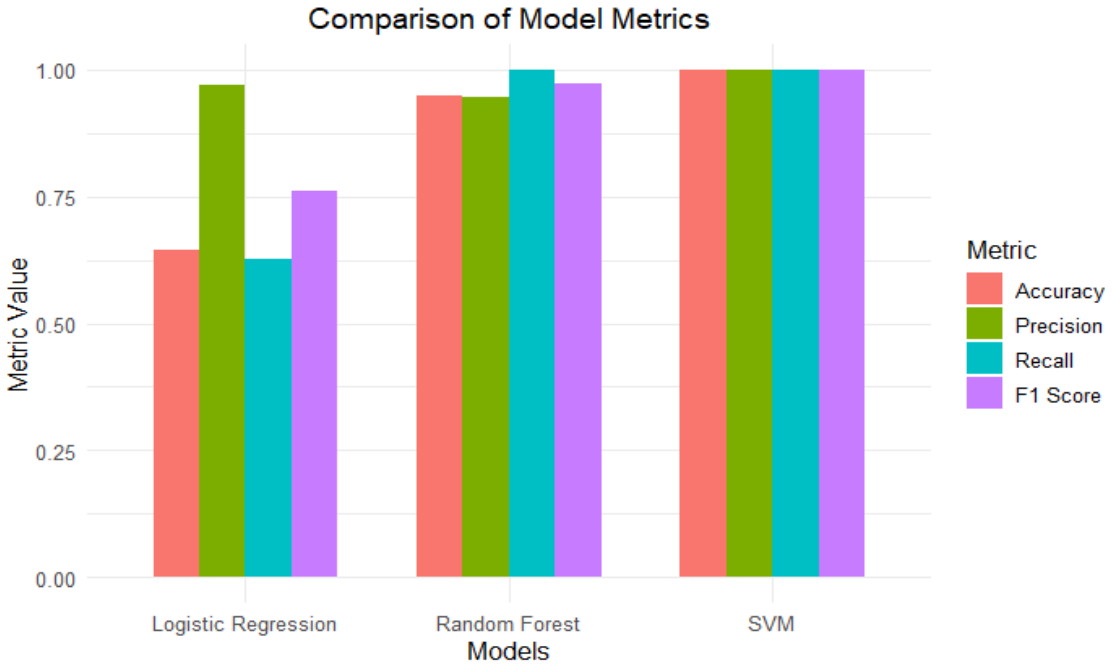


Figure 4.10: Accuracy Comparison of the 3 models

4.6.3 Receiver Operating Characteristic Curve (ROC)

The ROC shows the performance of a classification model at all classification thresholds. This curve plots two parameters: the true positive rate (sensitivity) and the false positive rate (specificity).

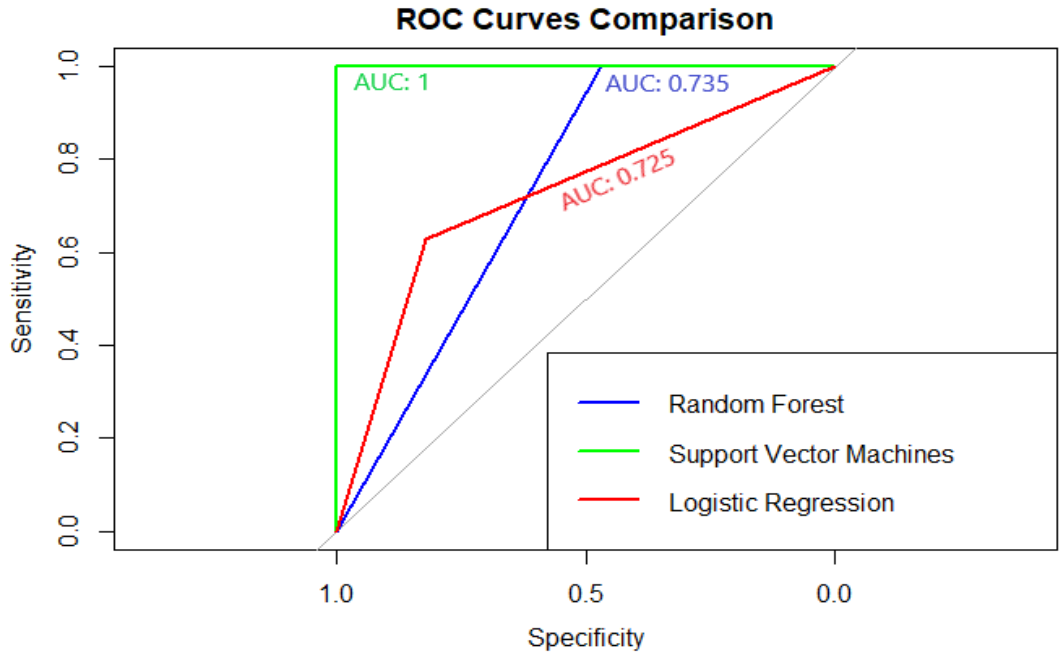


Figure 4.11: ROC Curves for the 3 models

After observing the ROC curve plot, we can clearly see that the SVM model has an area under curve of 1, which indicates an ideal scenario. Furthermore, LR and RF both have nearly equal AUCs of 0.735 for RF and 0.725 for LR, which is also an acceptable AUC.

The gray line in the plot indicates the worst-case scenario (AUC = 0.5), which is basically a random classifier scenario.

4.6.4 Confusion Matrix

A confusion matrix represents the prediction summary calculated previously in matrix form; it represents the true positives (predicted positive and are positive), the true negatives (predicted negative and are negative), the false positives (predicted positive and are negative), and finally the false negatives (predicted negative but are positive).

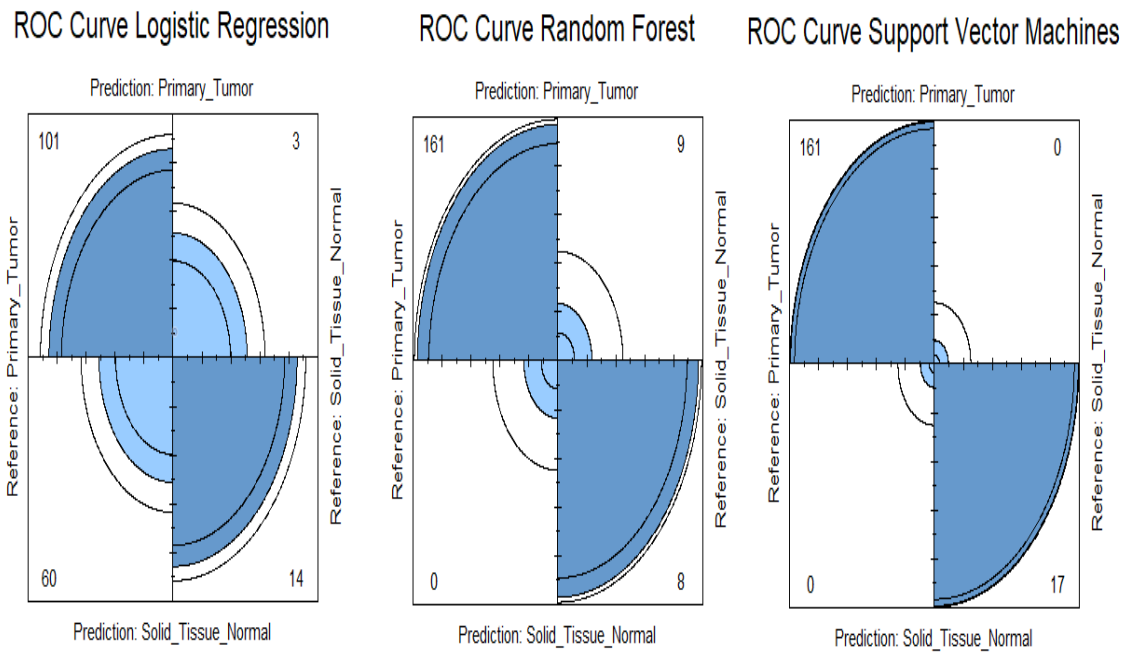


Figure 4.12: Confusion Matrices of the 3 models

For the logistic regression matrix, a total of 63 samples were mistakenly predicted, 60 lung cancer patients were predicted as normal, and 3 normal cases were predicted as lung cancer. This is not favorable, as the error here is critical (predicting cancer patients as normal).

For the Random Forest Matrix, a total of 9 samples were mistakenly predicted; all of them were normal cases predicted as lung cancer patients. This error here is more acceptable because, by predicting a normal patient as a lung cancer patient, we would eventually discover that the diagnosis was wrong after taking the normal procedures.

For the Support Vector Machines Matrix, no wrong predictions were made, which is ideal.

4.7 Feature Importance

After discussing the results of the three machine learning models, we will dive deeper into the most important features of the models—the genes that are most important to the model.

Logistic regression won't be taken into account due to its results; the Feature importance of the logistic regression may not be reliable enough to be taken into consideration.

focusing on support vector machines and random forest models, both their Feature importance scores were calculated.

4.7.1 Results

Gene	Importance
TCEAL2	1.2115
HSD17B6	0.9509
ITLN1	0.8833
PTCHD1	0.8008
GLP2R	0.7687
ALPP	0.7439
BEX1	0.7304
MS4A15	0.6672
SYT4	0.6363
C18orf63	0.6360

Table 4.2: Importance Scores of Genes (Random Forest)

Gene	**Importance**
TCEAL2	0.9869
VTN	0.9710
ITLN1	0.9636
HSD17B6	0.9599
MS4A15	0.9543
ALPP	0.9463
BEX1	0.9383
MYO1A	0.9355
CALB2	0.9115
C18orf63	0.9114

Table 4.3: Importance Scores of Genes (Support Vector Machines)

After observing the previous results, we can see that both models have the same most important genes with differences in importance scores. Moreover, we are going to sum them both up to get a better interpretation of the results.

Gene	Importance
TCEAL2	1.2115
HSD17B6	0.9509
ITLN1	0.8833
PTCHD1	0.8008
GLP2R	0.7686
ALPP	0.7439
BEX1	0.7304
MS4A15	0.6672
SYT4	0.6363
C18orf63	0.6359

Table 4.4: Importance Scores of both models summed and sorted

Now that we have the most important genes according to the 2 models, we can further interpret them better by adding out STRING Network results to the importance scores. This would yield even more precise results as we would have combined more than one analysis technique together.

Before summing them up, we have to scale the importance scores up to match up the string network node degrees. Importance scores were multiplied by a factor (max(network degrees)/max(importance)) to scale them up.

Gene	Importance (Scaled)	String Network Degree	Sum
ALPP	6.11	10	16.11
HSD17B6	7.84	7	14.84
BEX1	6.02	5	11.02
TCEAL2	10.00	1	11.00
GLP2R	6.27	2	8.27
ITLN1	7.26	1	8.26
PTCHD1	6.60	1	7.60
SYT4	5.20	1	6.20
MS4A15	5.45	0	5.45
C18orf63	5.20	0	5.20

Table 4.5: Combined importance scores with String Network Degrees

After scaling the importance scores, network degrees were summed with the importance scores and sorted in descending order. Moreover, we are going to investigate the roles of the top 4 genes in this list in order to see whether any gene is related to the lungs or cancer in any way.

Our study revealed that ALPP as the top important gene. It is an alkaline phosphatase that can hydrolyze various phosphate compounds.

A previous study showed that ALPP mRNA expression was statistically significantly higher in 12 different cancers [18], namely: OV, testicular germ cell tumors (TGCT), uterine corpus endometrial carcinoma (UCEC), pancreatic adenocarcinoma (PAAD), bladder urothelial carcinoma (BLCA), stomach adenocarcinoma (STAD), esophageal carcinoma (ESCA), uterine carcinosarcoma (UCS), rectum adenocarcinoma (READ), head and neck squamous cell carcinoma (HNSC), clone adenocarcinoma (COAD), and acute myeloid leukemia (LAML).

This finding suggests that ALPP might also be related to lung cancer since it is related to other types of cancer; furthermore, ALPP might be a potential lung cancer biomarker.

HSD17B6 was identified as the second most important gene. This gene is responsible for converting the androgen DHT to the estrogen 3 beta-adiol.

A previous study found that HSD17B6 mRNA and protein's expression was frequently lower in LUAD than in normal people[31]; moreover, their findings indicate that HSD17B6 may function as a tumor suppressor in LUAD and could be a promising prognostic indicator for LUAD patients, especially for those receiving radiotherapy.

This finding suggests that HSD17B6 is highly associated with LUAD due to its lower expression in LUAD patients; furthermore, this suggests that HSD17B6 could be a potential biomarker for LUAD prognosis.

Bex1 was found to be the third most important gene in our analysis. It is primarily expressed in the brain and is mostly related to neural development and function.

A previous study found that Bex1 was identified as being involved in cancer invasion, as it was highly expressed in the invasive cell lines[21]. Its downregulation suppressed the invasion and proliferation of the invasive tumor cell lines.

These findings suggest that Bex1 may promote metastasis and could be a potential therapeutic target in lung adenocarcinoma.

TCEAL2 is the fourth important gene in our analysis. The exact function of TCEAL2 is still under investigation. It is likely involved in regulating gene transcription.

A previous study found that TCEAL2's expression was down-regulated in most cancers[30].

This finding suggests that TCEAL2 may be a biomarker for cancer in different cancers and may affect tumors through immune infiltration (the process in which immune cells migrate to areas where they should not be). This has the potential to either suppress or promote tumor growth.

Chapter 5

Discussion

In our study, we conducted a comprehensive differential expression analysis of lung cancer to identify differentially expressed genes (DEGs) that may serve as potential biomarkers. Our analysis showed a total of 515 differentially expressed genes, of which 502 were found to be downregulated and 13 were found to be upregulated in lung cancer patients. The top 5 downregulated differentially expressed genes were OLFM4, POU3F2, ZIC1, PCSK1, and ABCC2, while the top 5 upregulated differentially expressed genes were PTCHD1, FABP12, PON1, HSD17B6, and TCEAL2. After the DEG identification, we used visualizations like heatmaps and volcano plots to further understand our DEGs; furthermore, the enrichR online tool was used to determine enrichments in OMIM diseases, KEGG pathways, and GO biological processes.

Our DEGs show a diverse range of biological functions and pathways, underscoring the complex nature of lung cancer biology. These findings are consistent with previous research demonstrating similar gene expression alterations in various diseases, including lung cancer. For instance, a study highlighted the potential of OLFM4 (our top downregulated DEG) as it might be a potential therapeutic strategy for NSCLC[23]. Additionally, another study identified that POU3F2 (our second top downregulated DEG) has a significant role in neuroendocrine (a system that is made up of nerves and gland cells) differentiation of lung cancers[25].

Regarding our top upregulated DEG, PTCHD1, no studies found a direct relationship between it and any type of cancer; however, TCEAL2 and HSD17B6 were found by previous studies mentioned in the results section [30][31] to be associated with cancer, and HSD17B6 was associated with LUAD specifically. While TCEAL2 is not directly connected to lung cancer, it should not be overlooked because of its connection to cancer and how they can all be relevant to each other.

Additionally, the gene set enrichment analysis for our DEGs was also supported by multiple research papers, as mentioned in the results section. For example, the most enriched KEGG pathway in our enrichment analysis was neuroactive ligand receptor interaction, which is a very popular cause of bladder cancer[24]. Even though this study doesn't support the association with lung cancer, we shouldn't overlook it as it might

still be related. Same thing for both OMIM diseases and GO processes; the third most enriched OMIM disease was EB, which was found to be more aggressive in NSCLC patients by a previous study[19].

Machine learning results are promising for both random forest and support vector machines; however, logistic regression's results might not be promising due to the fact that LR is known to behave and train better in regression scenarios where you have to predict a continuous value instead of classifying either normal or cancer. In our case, LR was used in order to view the differences between classification-focused models and regression-focused models.

Machine learning's most important genes, along with the string network results, were combined, and our top 4 genes accordingly are ALPP, HSD17B6, Bex1, and TCEAL2. Interestingly, TCEAL2 and HSD17B6 were also two of the most upregulated genes in our DEGs, which put my attention into further investigation of those 2 genes. As mentioned previously in the results section, HSD17B6's expression was found to be lower in LUAD patients in a previous study[31], and TCEAL2's expression was found to also be lower in most cancer cases. However, in our case, both TCEAL2 and HSD17B6 were upregulated, which arose as a new topic for further investigation.

5.1 Conclusion

In conclusion, our comprehensive differential expression analysis of lung cancer identified 515 differentially expressed genes (DEGs) between lung cancer and normal tissues. Among these DEGs, 502 were downregulated, while 13 were upregulated. The top 5 upregulated genes, including OLFM4, POU3F2, ZIC1, PCSK1, and ABCC2, along with the top 5 downregulated genes, including PTCHD1, FABP12, PON1, HSD17B6, and TCEAL2, displayed significant differences in expression levels and hold potential as biomarkers for lung cancer. These findings suggest their involvement in the progression and prognosis of lung cancer.

Moreover, gene set enrichment analysis for the 515 identified differentially expressed genes (DEGs) showed connections to various diseases (OMIM diseases), cellular pathways (KEGG pathways), and fundamental biological processes (GO biological processes). This suggests that these genes are important not just for lung diseases but also for many other health problems.

Furthermore, machine learning yielded promising results, with SVM and RF having really positive results, while LR had negative results regarding all the aspects of performance measurement (accuracy, precision, recall, F1 score, ROC curve, and confusion matrix).

Finally, the feature importance of the DEGs showed the most important genes to the machine learning models, which were taken into account (SVM and RF). They were combined with the STRING Network results to output the most important genes. More research is necessary to fully understand the complexities of lung cancer and make new findings that could significantly impact how we diagnose, treat, and predict the course of the disease.

5.2 Future Work

Experimental validation should be done on the identified potential biomarkers, particularly those highlighted by both machine learning and STRING network analysis, to confirm their roles in lung cancer development, progression, and therapy.

Clinical studies with more patients and a balanced ratio between tumor and normal cases are needed to validate the clinical significance of these biomarkers.

Investigating the mechanisms by which the identified genes contribute to lung cancer is crucial. This could involve studying their functions and interactions with the body.

Comparing the performance of other machine learning algorithms could lead to further improvements in prediction accuracy and biomarker identification; moreover, using deep learning approaches such as neural networks.

The unexpected upregulation of TCEAL2 and HSD17B6 in LUAD samples suggests further investigation to understand the underlying potential implications for lung cancer. Analyzing the identified biomarkers in different subtypes of lung cancer (e.g., adenocarcinoma, squamous cell carcinoma) could reveal subtype-specific biomarkers and treatment strategies.

Appendix

Appendix A

Lists

NSCLC	Non-Small Cell Lung Cancer
DEGs	Differentially Expressed Genes
PCA	Principal Component Analysis
DEA	Differential Expression Analysis
SCLC	Small Cell Lung Cancer
LUAD	Lung Adenocarcinoma
TCGA	The Cancer Genome Atlas
NCI	National Cancer Institute
GSEA	Gene Set Enrichment Analysis
OMIM	Online Mendelian Inheritance in Man
KEGG	Kyoto Encyclopedia of Genes and Genomes
ML	Machine Learning
RF	Random Forest
LR	Logistic Regression
ROC	receiver operating characteristic curve

List of Figures

Bibliography

[1] Applications of metabolomics. https://www.ebi.ac.uk/training/online/courses/metabolomics-introduction/the-importance-of-metabolomics/applications-of-metabolomics/. Accessed: 2024-03-18.

[2] The cancer genome atlas. https://www.cancer.gov/ccg/research/genome-sequencing/tcga. Accessed: 2024-04-10.

[3] A commonality between various omes in multi-omics approaches. Omics-BasedClinicalDiscovery:Science,Technology,andApplications. Accessed: 2024-03-18.

[4] Comparing experimental conditions: differential expression analysis. *biocorecrg*.

[5] Enrichr. https://maayanlab.cloud/Enrichr/. Accessed: 2024-04-10.

[6] Gene set enrichment analysis. https://www.gsea-msigdb.org/gsea/index.jsp. Accessed: 2024-04-10.

[7] Logistic regression. https://www.sciencedirect.com/topics/computer-science/logistic-regression#:~:text=Logistic%20regression%20is%20a%20process,%2Fno%2C%20and%20so%20on. Accessed: 2024-04-10.

[8] Lung cancer facts. https://www.lungcancerresearchfoundation.org/lung-cancer-facts/#:~:text=1%20IN%2016%20PEOPLE%20will,and%201%20in%2017%20women.&text=Approximately%20127%2C070%20AMERICAN%20LIVES%20are%20lost%20annually.&text=654%2C620%20PEOPLE%20IN%20THE%20U.S.,some%20point%20in%20their%20lives. Accessed: 2024-03-19.

[9] National human genome research institute, cancer. https://www.genome.gov/genetics-glossary/Cancer. Accessed: 2024-03-18.

[10] Neuron differentiation. https://www.informatics.jax.org/vocab/gene_ontology/GO:0030182#:~:text=neuron%20differentiation%20Gene%20Ontology%20Term%20(GO%3A0030182)&text=Definition%3A,specialized%20features%20of%20a%20neuron.

[11] Random forest classifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Accessed: 2024-04-10.

[12] String. https://string-db.org/. Accessed: 2024-04-10.

[13] Supervised learning. https://databasetown.com/supervised-learning-algorithms/. Accessed: 2024-03-19.

[14] Unsupervised learning. https://databasetown.com/unsupervised-learning-types-applications/. Accessed: 2024-03-19.

[15] What are the stages of lung cancer? https://treatcancer.com/blog/what-are-the-stages-of-lung-cancer/. Accessed: 2024-03-18.

[16] Deseq2. *Bioconductor*, 2024.

[17] R (programming language). *Wikipedia*, 2024.

[18] Mohsen Basiri and Saghar Pahlavanneshan. Evaluation of placental alkaline phosphatase expression as a potential target of solid tumors immunotherapy by using gene and protein expression repositories. *Cell J.*, 23(6):717–721, November 2021.

[19] Domenico Bonamonte, Angela Filoni, Aurora De Marco, Lucia Lospalluti, Eleonora Nacchiero, Valentina Ronghi, Anna Colagrande, Giuseppe Giudice, and Gerardo Cazzato. Squamous cell carcinoma in patients with inherited epidermolysis bullosa: Review of current literature. *Cells*, 11(8):1365, April 2022.

[20] Nuno R Carreira, Catarina Gonçalves, Alexandra Wahnon, Sara Dâmaso, and Joao Martins. Late diagnosis of Maturity-Onset diabetes of the young (MODY) 12 with catastrophic consequences. *Cureus*, 13(2):e13145, February 2021.

[21] Takefumi Doi, Hiroyuki Ogawa, Yugo Tanaka, Yoshitake Hayashi, and Yoshimasa Maniwa. Bex1 significantly contributes to the proliferation and invasiveness of malignant tumor cells. *Oncol. Lett.*, 20(6):1–1, October 2020.

[22] Anna Falanga. Thrombophilia in cancer. *Semin Thromb Hemost*, 31(1):104–110, February 2005.

[23] Xian-Zheng Gao, Guan-Nan Wang, Wu-Gan Zhao, Jing Han, Chang-Ying Diao, Xiao-Hui Wang, Sheng-Lei Li, and Wen-Cai Li. Blocking OLFM4/HIF-1α axis alleviates hypoxia-induced invasion, epithelial–mesenchymal transition, and chemotherapy resistance in non-small-cell lung cancer. *J. Cell. Physiol.*, 234(9):15035–15043, September 2019.

[24] Zhaohui He, Fucai Tang, Zechao Lu, Yucong Huang, Hanqi Lei, Zhibiao Li, and Guohua Zeng. Analysis of differentially expressed genes, clinical value and biological pathways in prostate cancer. *Am J Transl Res*, 10(5):1444–1456, May 2018.

[25] Jun Ishii, Hanako Sato, Takuya Yazawa, Yukiko Shishido-Hara, Chie Hiramatsu, Yukio Nakatani, and Hiroshi Kamma. Class III/IV POU transcription factors expressed in small cell lung cancer cells are involved in proneural/neuroendocrine differentiation. *Pathol. Int.*, 64(9):415–422, September 2014.

[26] Felix W. Leung Kazuo Endoh. Effects of smoking and nicotine on the gastric mucosa: A review of clinical and experimental evidence. 107, September 1994.

[27] Gang Liu, Chen Li, Haiyan Zhen, Zhigang Zhang, and Yongzhong Sha. Identification of prognostic gene biomarkers for metastatic skin cancer using data mining. *Biomed. Rep.*, 13(1):22–30, July 2020.

[28] R. Rosell, T. G Bivona, and N. Karachaliou. Genetics and biomarkers in personalisation of lung cancer treatment. Aug 2013.

[29] Abel Sanchez-Palencia, Mercedes Gomez-Morales, Jose Antonio Gomez-Capilla, Vicente Pedraza, Laura Boyero, Rafael Rosell, and Mª Esther Fárez-Vidal. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *International Journal of Cancer*, 129(2), 2011.

[30] Yu Sun and Jun Zhao. Transcription elongation factor a (SII)-like (TCEAL) gene family member-TCEAL2: A novel prognostic marker in pan-cancer. *Cancer Inform.*, 21:117693512211262, January 2022.

[31] Tian Tian, Fu Hong, Zhiwen Wang, Jiaru Hu, Ni Chen, Lei Lv, and Qiyi Yi. HSD17B6 downregulation predicts poor prognosis and drives tumor progression via activating akt signaling pathway in lung adenocarcinoma. *Cell Death Discov.*, 7(1), November 2021.

[32] Fangwei Wang, Qisheng Su, and Chaoqian Li. Identidication of novel biomarkers in non-small cell lung cancer using machine learning. *Scientific Reports*, 12(1):16693, Oct 2022.

[33] Stephen Watts. What is data normalization? *BMC*, 2020.

[34] Timothy T Xu, Keith H Baratz, Michael P Fautsch, David O Hodge, and Michael A Mahr. Cancer risk in patients with fuchs endothelial corneal dystrophy. *Cornea*, 41(9):1088–1093, September 2022.