**AI and Explainable Machine Learning in Predicting Biomarkers for Drug Discovery**

***A***

***Project Report***

***submitted in partial fulfillment of the***

***requirements for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**in**

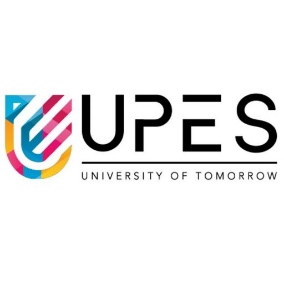
**COMPUTER SCIENCE & ENGINEERING**

**by**

|  |  |
| --- | --- |
| **Name** | **Roll No.** |
| **Saurav Singh** | **R2142221504** |
| **Rishabh Verma** | **R2142221020** |
| **Mohit Kumar** | **R21422211109** |

***under the guidance of***

**Dr. Pooja Sarin**

****

**School of Computer Science**

**University of Petroleum & Energy Studies**

**Bidholi, Via Prem Nagar, Dehradun, Uttarakhand**

**January – 2025**

**CANDIDATE’S DECLARATION**

We hereby certify that the project work entitled **“AI and Explainable Machine Learning in Predicting Biomarkers for Drug Discovery”** in partial fulfilment of the requirements for the award of the Degree of BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING with specialization in Artificial Intelligence And Machine Learning and submitted to the Department of Systemics, School of Computer Science, University of Petroleum & Energy Studies, Dehradun, is an authentic record of our work carried out during a period from **January**, **2025** to **April**, **2025** under the supervision of **Dr. Pooja Sarin, Designation and Affiliation>**.

The matter presented in this project has not been submitted by us for the award of any other degree of this or any other University.

Rishabh Verma-R2142221020

Saurav Singh-R2142221504

Mohit Kumar-R2142221109

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: 30-04-2025 Dr. Pooja Sarin

Project Guide

**ACKNOWLEDGEMENT**

We wish to express our deep gratitude to our guide **Name**, for all advice, encouragement and constant support he/she has given us throughout our project work. This work would not have been possible without his support and valuable suggestions.

We sincerely thanks to our respected **Name of HoD,** **Head Department of SOCS,** for his great support in doing our project in **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.**

We are also grateful to Dean SoCS UPES for giving us the necessary facilities to carry out our project work successfully. We also thanks to our Course Coordinator, Sandeep Chand Kumain and our Activity Coordinator (NAME) for providing timely support and information during the completion of this project.

We would like to thank all our **friends** for their help and constructive criticism during our project work. Finally, we have no words to express our sincere gratitude to our **parents** who have shown us this world and for every support they have given us.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Rishabh Verma** | **Saurav Singh** | **Mohit Kumar** |  |
| **Roll No.** | **R2142221020** | **R2142221504** | **R2142221109** |  |

**ABSTRACT**

Early diagnosis and molecular stratification of **Non-Small Cell Lung Cancer (NSCLC)** is of utmost importance in order to enhance patient outcomes and facilitate precision therapy. Utilizing multi-omics information (**RNA-Seq**, **methylation**, **CNA**, **protein arrays**, and **clinical data**) from **TCGA-LUAD (Lung Adenocariconama)** and **TCGA-LUSC (Lung Squamous Cell Cariconama)**, in this research, we used **Explainable Artificial Intelligence (XAI)** to detect biomarkers that can predict **NSCLC** subtypes and survival prediction. **Multi-Layer Perceptron** and **Logistic Regression** based deep learning models were trained to classify subtypes and were interpreted in terms of feature importance through **SHAP (SHapley Additive exPlanations)** and **LIME (Local Interpretable Model-agnostic Explanations)** to reveal genes modulating the tumor microenvironment, cell cycle, and the immune response. **Protein-protein interaction** and **KEGG** pathway analyses identified biologically relevant hub proteins that were combined with a **Cox regression model** to model the survival prediction. **Kaplan-Meier** and KDE plots showed good risk stratification with subtype-specific distinct survival probabilities.

Parallel with that, a colorectal cancer gene expression study based on the **E-MEXP-3756** dataset detected **differentially expressed genes (DEGs)** that served as early-detection **biomarker** candidates. Translational value of **XAI** was supported in oncology by normalization, **PCA**, enrichment analysis and **SHAPE-based** feature selection with **R/Bioconductor** tools.

These observations reinforce the strength of **XAI** in discovering interpretable biomarkers and risk-stratifying patients with targeted treatment. Subsequent research will include experimental validation of candidate genes (for example, **NRAS and SMAD4**) and prospective clinical evaluation to translate computational insights to actionable precision medicine. This two-pronged approach—across **NSCLC** and colorectal cancer—demonstrates the generalizability of **XAI-driven** biomarker discovery in cancer diagnostics and treatment enhancement.

**TABLE OF CONTENTS**

**S.No. Contents Page No**

1. **Introduction 1**
   1. History 1
   2. Requirement Analysis 2
   3. Objective 2
2. **System Analysis 3**
   1. Motivation 3
   2. Modules 3-7
      1. Data Loading and Collection
      2. Data Preprocessing and correlation analysis of proteins
      3. Differentially Expressed Genes
      4. Protein-Protein Interaction Analysis
      5. Machine Learning Implementation Applying SHAP and LIME
3. **Implementation 8**
   1. Data Preprocessing 8-14
      1. Data Loading for Colorectal Cancer and Data Collection for Lung Cancer
      2. Normalization and Transformation
      3. Probe annotation and feature selection
      4. Differentially Expressed Genes
   2. Methodology 14-28
      1. Key Statistics
      2. Significant Gene Attributes
      3. Results and Visualization of DEGs and Protein Lung Cancer Data
      4. Volcano Plot
      5. Correlation Heatmap
      6. Observations
      7. Top 10 DEGs
      8. Gene Enrichment Analysis
      9. Protein-Protein Network
      10. Centrality Analysis
   3. **Gene Ontology BP enrichment and KEGG Pathway Analysis 29**
   4. **Machine Learning Model 33**
   5. **XAI Techniques 40**
   6. **Survival Analysis 47**
4. **Limitations and Future Enhancements 54-56**
   1. **Limitations**
   2. **Future Enhancements**
5. **Conclusion 59**
6. **References 60**

**LIST OF FIGURES**

**S.No. Figure**

Fig 1: Methodology for colorectal cancer

Fig 2: Methodology for Lung cancer

Fig 3: steps for data collection for colorectal cancer

Fig 4: Collected data of colorectal cancer

Fig 5: Collected data of TCGA-LUAD

Fig 6: Data collected for TCGA-LUSC

Fig 7: box plot before normalization for colorectal cancer

Fig 8: box plot after normalization for colorectal cancer

Fig 9: PCA plot before normalization for colorectal cancer

Fig 10: PCA plot after normalization for colorectal cancer

Fig 11: Heatmap for colorectal cancer

Fig 12: Correlation matrix

Fig 13: Probe annotation for colorectal cancer

Fig 14: Gene symbols and GeneIDs for Lung cancer

Fig 15: Gene Symbols

Fig 16: Differential expressed gene (DEGs) for colorectal cancer

Fig 17: RNA Sequential Gene Expression data

Fig 18: DNA Methylation HM27 and HM450 gene expression data

Fig 19: CNA(Copy Number Variations) gene expression data

Fig 20: Reverse Phase Protein Array(RPPA) data

Fig 21: Clinical Sample IDs

Fig 22: Colorectal Cancer Data

Fig 23: TCGA Sample Patient ids

Fig 24: Violin Plot for Healthy v/s Colorectal Cancer

Fig 25: Correlation matrix for Protein Pairs of Lung Cancer Subtypes

Fig 26: Top 10 unregulated gene symbols

Fig 27: Top 10 Downregulated genes

Fig 28: Top 10 DEGs

Fig 29: GO BP Enrichment Analysis of Genes

Fig 30: KEGG Pathway Enrichment Analysis

Fig 31: PPI Network

Fig 32: Top Matched Genes from StringDB

Fig 33: Network Graph Visualization of LUAD and LUSC

Fig 34: Top Hub Proteins by degree

Fig 35: Top Highly Correlated Protein Pairs

Fig 36: Centrality Analysis Of genes

Fig 37: Network graph of top hub proteins

Fig 38: Symbol and GeneId of Proteins

Fig 39: GO BP Enrichment-2

Fig 40: KEGG Enrichment-2

Fig 41: KEGG PPI Enriched data LUSC

Fig 42: KEGG PPI Enriched data LUSC

Fig 43: Evaluation Metric-1

Fig 44: Evaluation Metric-2

Fig 45: Evaluation Metric-3

Fig 46: Model Training and Loss

Fig 47: Sigmoid Function

Fig 48: Log loss function

Fig 49: Evaluation Metric

Fig 50: SHAP values

Fig 51: Beeswarm Plot

Fig 52: Per Gene SHAP Importance by Condition

Fig 53: LIME Plot

Fig 54: KM Plot

Fig 55: Risk Score Plot

Fig 56: KDE Plot

**LIST OF TABLES**

**S.No. Table**

1. Neural Network Model
2. Model Comparison
3. **INTRODUCTION**
   1. **History**

Biomarkers have traditionally been critical in reshaping the landscape of therapeutic monitoring and diagnosis and prognosis in medicine, they are the measurable biological process of physiological or pathological responses or processes. Biomarkers can be traced from genomics (DNA/RNA/proteins) expressions patterns or from complex metabolomic (metabolites) process and imaging data. Molecular biological breakthroughs had resulted in the biomarker expansion in drug development adding advancement in Personalized medicine as a diagnostic, prognostic and predictive method and advancement in genomics have facilitated the use of microarray and RNA-seq technologies to study gene expression patterns, offering insights into underlying molecular mechanisms.

On the other hand, Institutions and Regulators such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) had promoted the use of validated biomarkers in clinical studies to expedite drug approvals and lower the cost of development. Historically, biomarker discovery has been based on hypothesis-driven and manual methods. Though efficient, they were laborious, costly, and subject to bias from humans. High-throughput technologies which includes next-generation sequencing (NGS), mass spectrometry, and metabolomics have substantially changed this field and thereby put a demand on the use of artificial intelligence (AI) to deliver useful insights with efficiency.

* 1. **Requirement Analysis**

This project was conducted in Rstudio and python using a comprehensive suite of packages and framework, ensuring a robust and reproducible workflow.

Summary of data used in this work:

* TCGA-LAUD: RNASeq data which consisted of 510 samples of patients across more than 20000~ gene samples. Similarly, DNA HM27 and HM450 methylation data consisted of 562 samples of patients and Copy Number Alterations (CNA) data consisted of 511 samples and Protein Array data consisted of 360 samples.
* TCGA-LUSC: RNASeq data which has 484 patient samples, DNA HM27 and HM450 methylation data consisted of 486 samples of patients, Copy Number Values (CNV) were consisted of 487 samples and Protein array data were consisted of 317 samples.
* E-MEXP-3756: Consists of High-quality microarray gene expression data.

comprehensive computational tools for annotation, visualization and enrichment analysis:

* To accurately annotate the gene expression they must map probe identifiers to gene symbols and biological functions. Using annotation packages such as hugu133plus2.db and org.Hs.eg.db to ensure that each probe is correctly linked to its corresponding gene.
* For visualization packages like ggplot, pheatmap, matplotlib, seaborn and enhancedVolcano are utilized to generate PCA plots, heatmaps and volcano plots.
* Enrichment analysis is performed using clusterProfiler which examines Gene Ontology (GO) and KEGG pathway enrichments.
* Machine learning techniques for feature selection and interpretability using random forest and frameworks like fastshap, shap and lime.
  1. **Objective**

The objective of this study is to identify significant biological roles of genomics in disease identification and reveal potential biomarkers for early detection and therapeutic targets using novel explainable AI techniques for the feature importance extraction of input genomics features.

**2. SYSTEM ANALYSIS**

* 1. **Motivations**

The main goal is to identify early detection biomarkers in colorectal cancer and to classify the TCGA(The Cancer Genome Atlas) Lung cancer data into its cancer subtypes LUAD(Lung Adenocarcinoma) and LUSC(Lung Squamous Cell Carcinoma) and to explore the explainable AI techniques like SHAP, LIME that helps to interpret the black box models.We will also derive the biological significance from the Protein array data through Protein-Protein interaction analysis that can help in drug target identification and identifying top differential hub proteins along with the clinical data for survival prediction.

* 1. **Modules**

The analysis of Colorectal cancer workflow consists of the following modules:

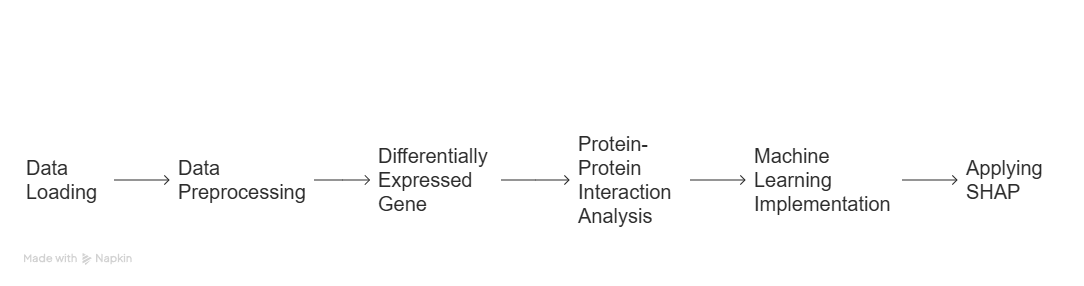


Fig 1: Methodology for colorectal cancer

The analysis of Lung cancer workflow consists of the following modules:

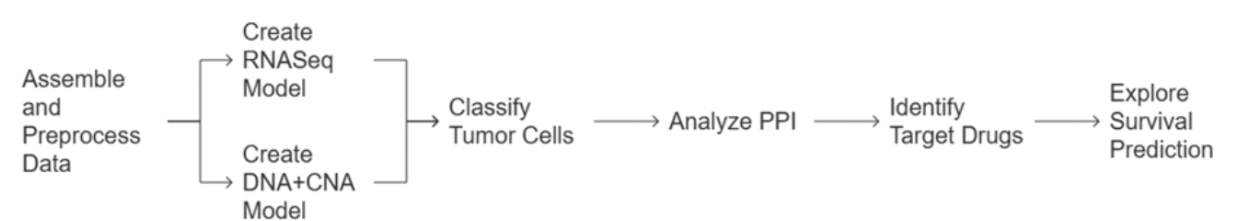


Fig 2: Methodology for Lung cancer

* + 1. **Data loading and collection**

Here is the pipeline for data collection of Colorectal Cancer data below.

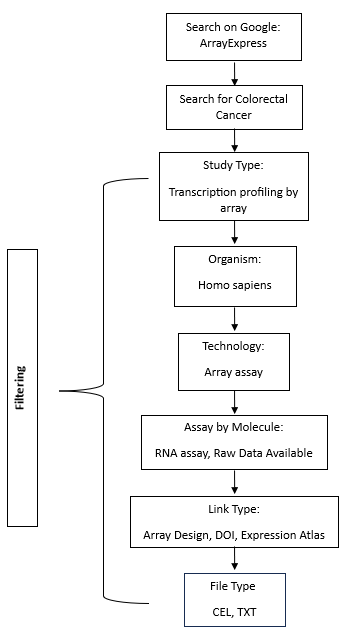


Fig 3: steps for data collection for colorectal cancer

RStudio utilizes functions from packages oligo and Biobase to read the raw CEL files and their associate SDRF metadata file "E-MEXP-3756.sdrf.txt." The SDRF file becomes an AnnotatedDataFrame through read.delim() function then conversion. By using oligo::read.celfiles() the program imports CEL files while linking each file to its respective metadata. This step creates a foundation since it verifies that all further procedures will be conducted on data that has been properly annotated.

Data collection for TCGA Lung Cancer dataset is done from the available online Genomic data portal cBioportal.The two datasets were downloaded luad\_tcga\_2018 and lusc\_tcga\_2018 dataset.Both the datasets consisted of multi-omics data such as RNA Sequential data, Copy Number Variations Data, DNA methylation HM27 and HM450 data and Reverse Phase Protein Array data and other genomics data.All the data consisted of numerical values of gene expressions like (log2ratio, z\_score normalised, etc).

* + 1. **Data preprocessing and correlation analysis of proteins**

The initial step includes data loading which is then processed by performing log2 transformation before implementing quantile normalization through RMA (Robust Multi-array Average) protocol. This correction method enables technicians to minimize technical sample biases. Boxplots along with PCA plots provide visualization methods to monitor quality before and after normalization by displaying log2-intensity distributions and confirming cluster formation of similar samples. The module implements procedures that select and retain high-quality interchangeable gene expression values essential for subsequent analysis.

The preprocessing in LUAD and LUSC data included RNASeq gene expression log2ratio data, Copy Number Alterations zscore data , DNA HM27 and HM450 numerical data, and RPPA zscore normalized data (Reverse Phase Protein Array).

The duplicate data is removed and then data is standardized.The high dimensional data is dimensionally reduced by applying PCA (Principle Component Analysis).

* + 1. **Differentially Expressed Gene**

The statistical analysis tools lead to gene expression comparison through the limma package. The normalized data receives linear model analysis to identify differentially expressed genes (DEGs) through pre-defined criteria (adjusted p-value < 0.05 and log fold change > specified threshold). The analysis filter process checks for both genes showing increased expression and those showing decreased expression. The EnhancedVolcano package generates volcano plots for displaying DEGs that show biologically important patterns for colorectal cancer.

For the lung cancer Protein array, the correlation analysis of protein array data is carried out to determine the top correlated proteins and their corresponding P-value matrix were taken to determine the protein pairs.

* + 1. **Protein-Protein Interaction Analysis**

The DEGs are converted to proteins while the STRINGdb package seek known and predicted protein–protein interactions. The ipgraph library generates visual network maps to display protein interactions from the retrieved data using STRINGdb. The computational method computes centrality measures including degree and betweenness and closeness to find proteins that serve as important disease mechanism hubs.

In case of Lung Cancer data, the protein pairs determined from the correlation analysis were matched with the known protein pairs from Stringdb and then the matched protein pairs (edges) were visualized in networkx graph and then the degree of each protein is determined to identify the top hub proteins.These top hub proteins were then selected for KEGG pathway analysis to determine the top differential proteins. These proteins were then combined with the clinical survival data for Survival analysis.

* + 1. **Machine Learning implementation Applying SHAP**

The randomForest package constructs a random forest classifier for normal and colorectal cancer sample discrimination by using chosen gene expression features. The dataset undergoes separation into training data and testing data for measuring model performance through computational metrics that include the confusion matrix. Model interpretability receives additional enhancement through the computation of SHAP (SHapley Additive exPlanations) values using the fastshap package. SHAP analysis calculates the exact influence that each feature (such as logFC or average expression or t-statistic) holds toward achieving prediction results. After the visualization methods with beeswarm plot and box plot which provide the knowledge of feature importance to get explainability.

For Lung cancer dataset, Logistic regression, XGBoost, Random Forest Classifier, MLPClassifier is implemented and applied SHAP to determine the feature importance of the model influencing the predictions and LIME to Locally approximate model decisions on feature importance for interpretability.

In case of protein array data of Lung cancer, **Cox Proportion Hazard** model is a regression model used trained on the differential protein data+clinical survival data for survival prediction in months.

**3. IMPLEMENTATION**

**3.1. Data Preprocessing**

**3.1.1. Data Loading for Colorectal Cancer and Data Collection for Lung Cancer**

For colorectal cancer, set the working directory in RStudio and load library “Biobase” which provides foundational classes and functions to handle biological data.

Read the SDRF file and convert it into an annotated data frame.

Process CEL files using the oligo::read.celfiles() function.

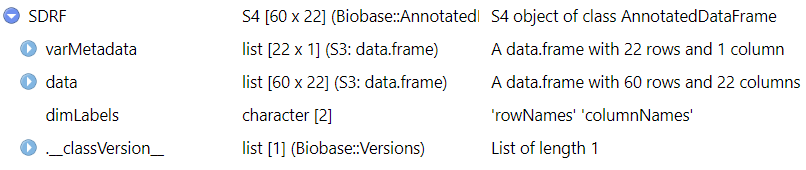


Fig 4: Collected data of colorectal cancer

For Lung Cancer, data is collected and loaded from genomic portal consisting of public TCGA-LUAD(Lung Adenocariconama) and TCGA-LUSC (Lung Squamous Cell Cariconama) data. The data consisted of following samples:-

TCGA-LUAD

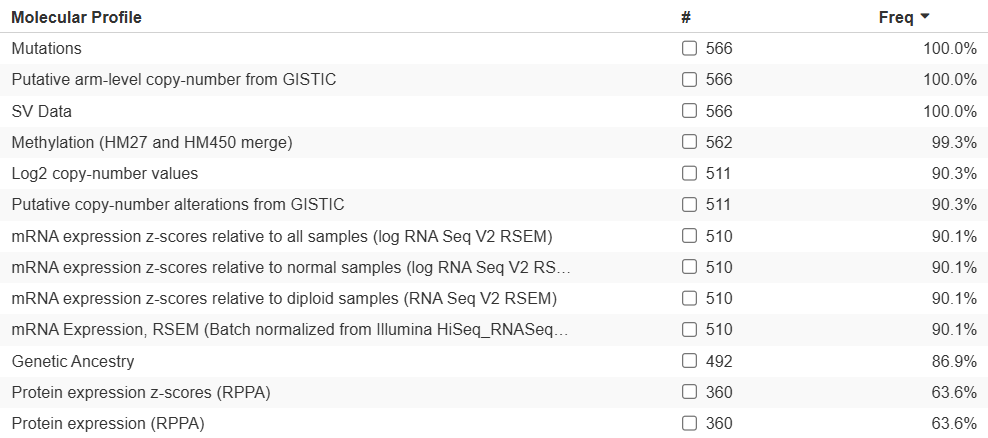


Fig 5: Collected data of TCGA-LUAD

TCGA-LUSC

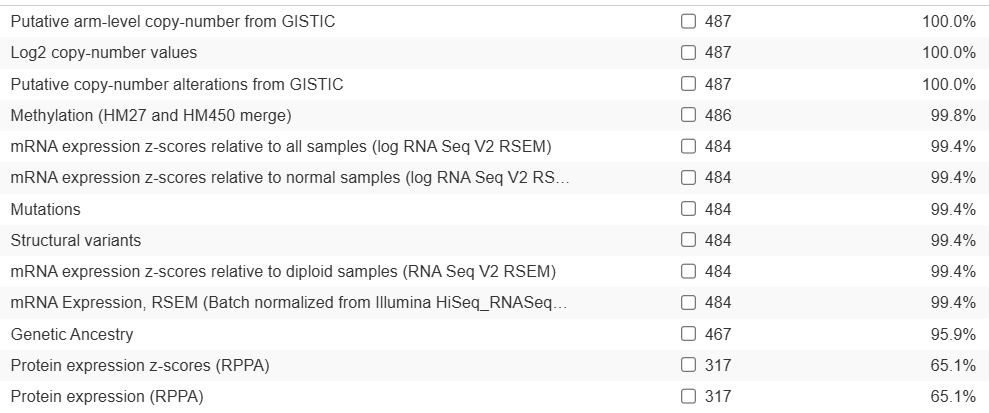


Fig 6: Data collected for TCGA-LUSC

**3.1.2. Normalization and Transformation**

In colorectal cancer, we performed Log2 transformation was applied to stabilize variance across expression values and Quantile normalization (RMA – Robust Multi-array Average) to adjust for technical variations.

**Before normalization**

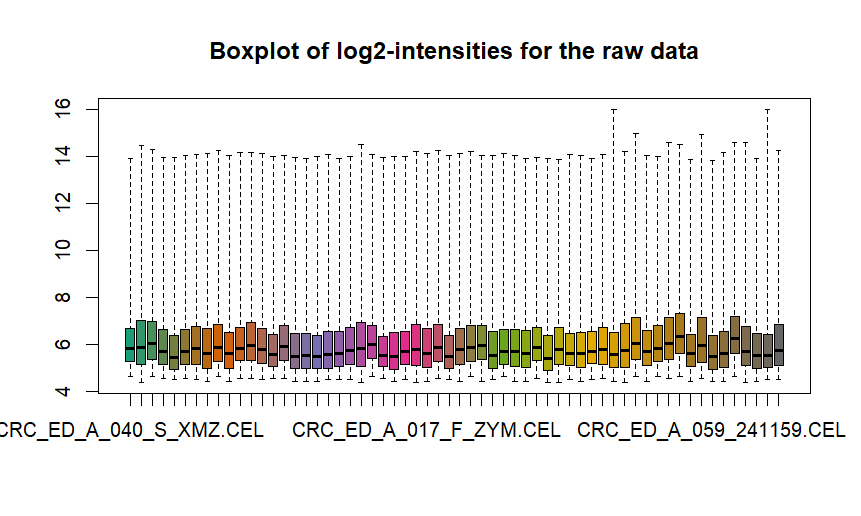


Fig 7: box plot before normalization for colorectal cancer

**After normalization**

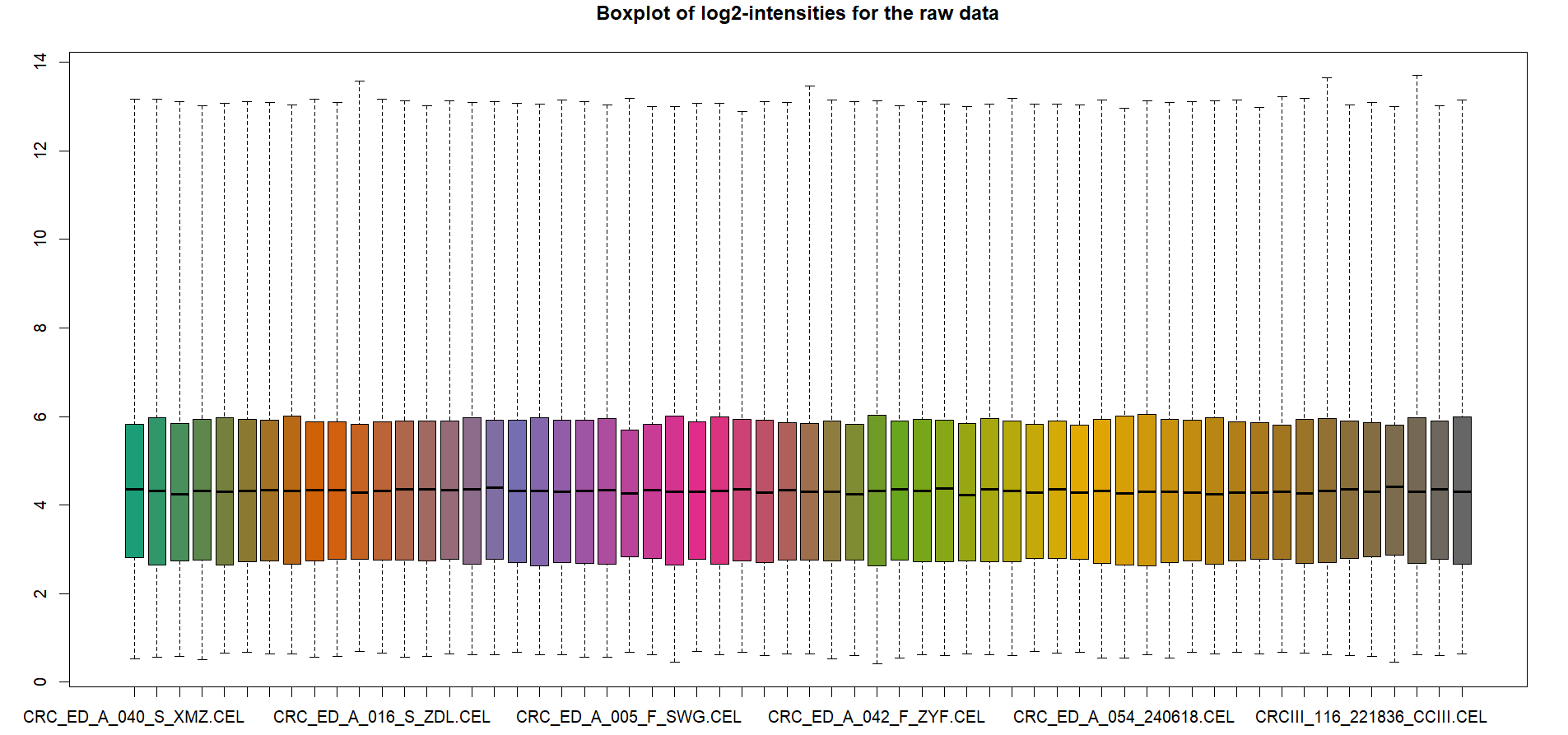


Fig 8: box plot after normalization for colorectal cancer

**PCA plots to visualize clustering of samples.**

**Before normalization**

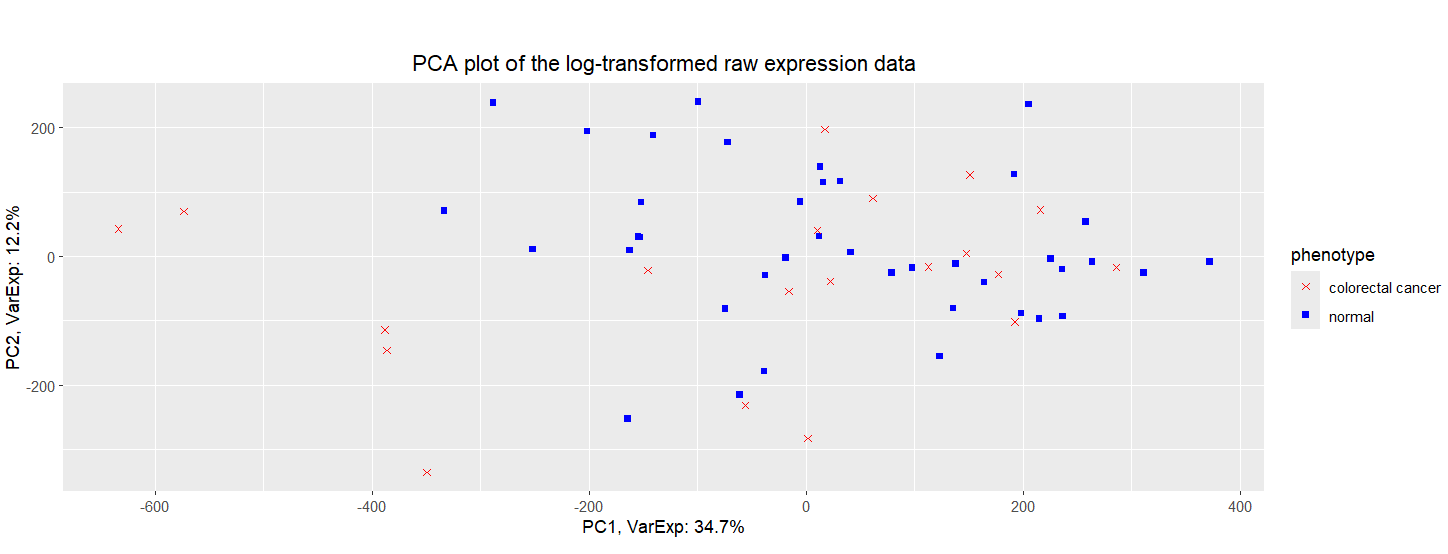


Fig 9: PCA plot before normalization for colorectal cancer

**After normalization**

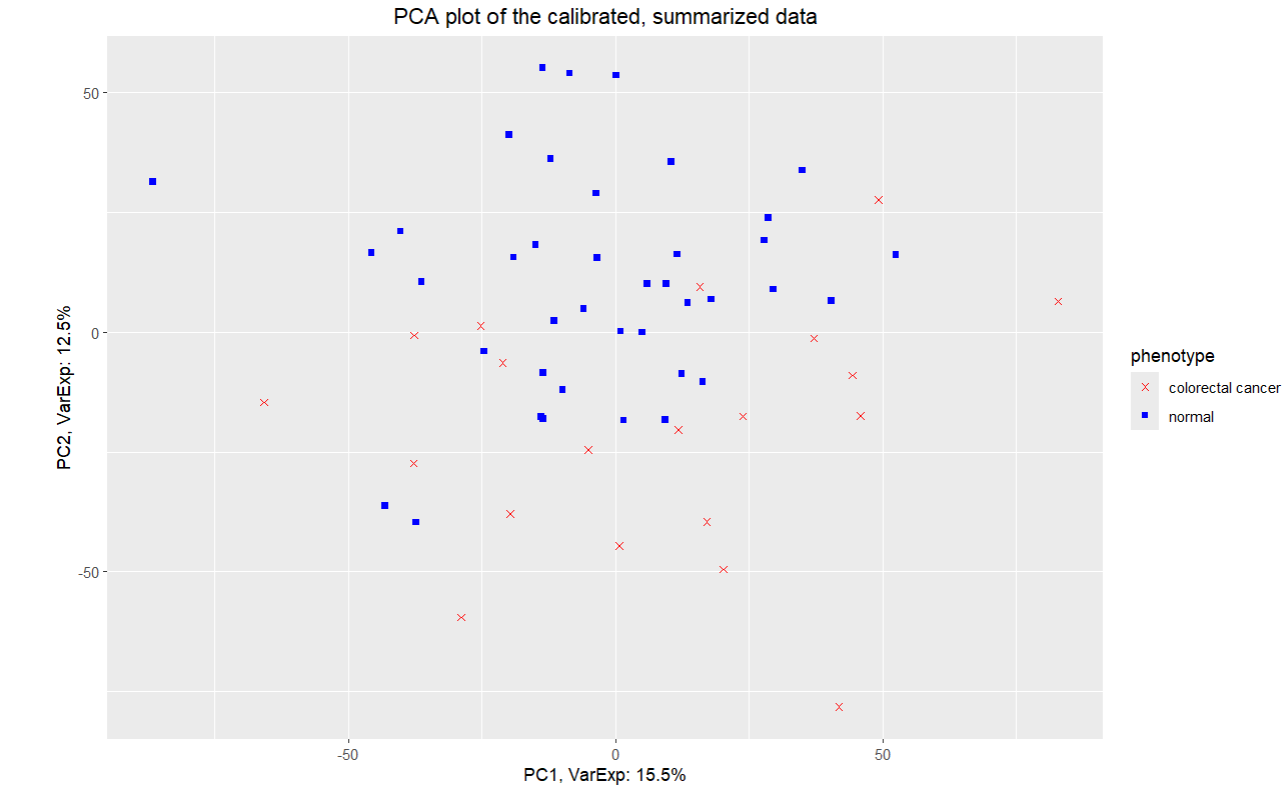


Fig 10: PCA plot after normalization for colorectal cancer

**Heatmap**

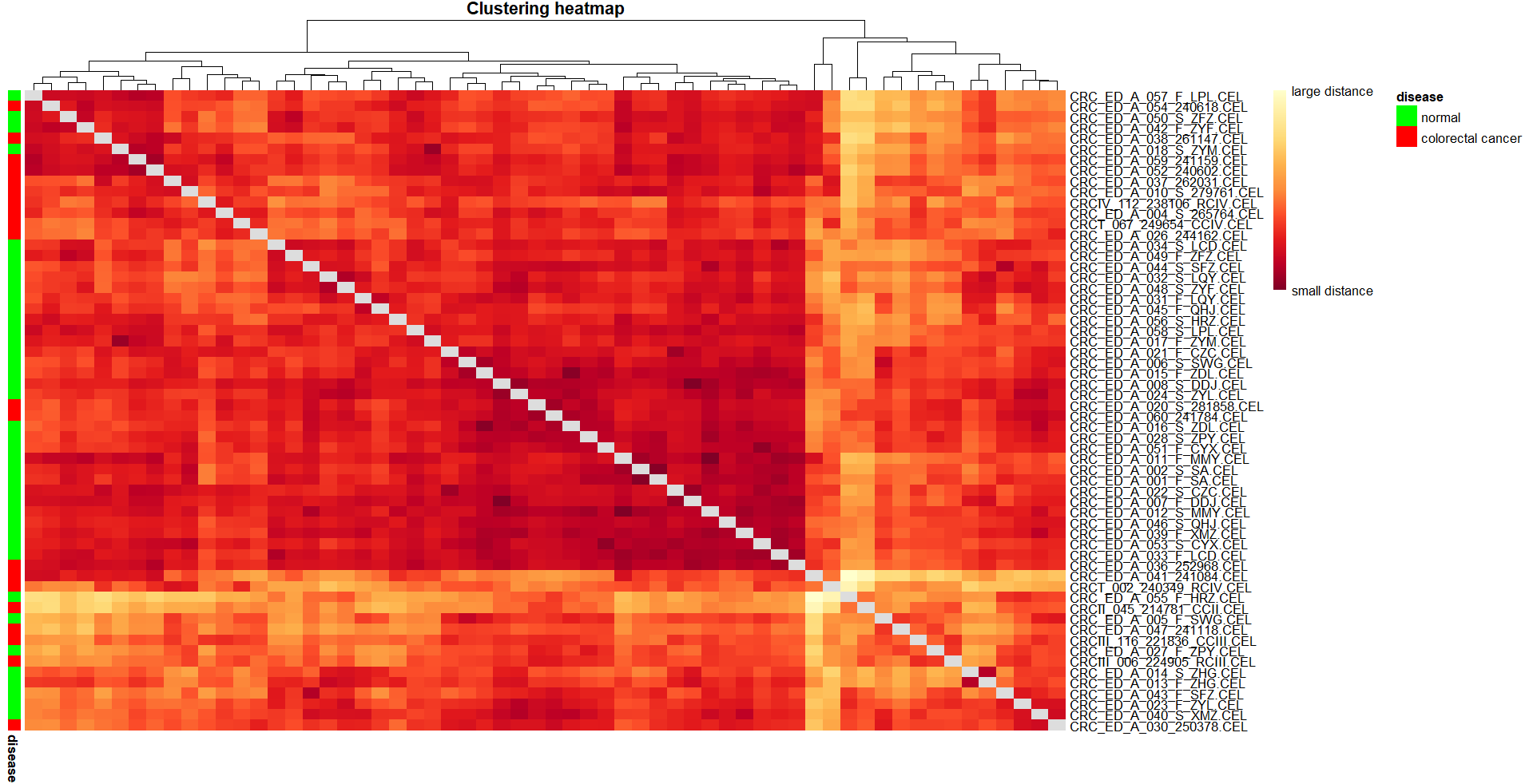


Fig 11: Heatmap for colorectal cancer

The top dendrogram shows hierarchical shows clustering of samples based on their distance. Samples that are close together in the dendrogram have similar expression profiles. In center color, the matrix shows pairwise distances between samples.

* White / light yellow = smaller distances (more similar)
* Red / Orange = larger distances (more dissimilar)

Color on left : Green -> Normal and Red -> colorectal cancer

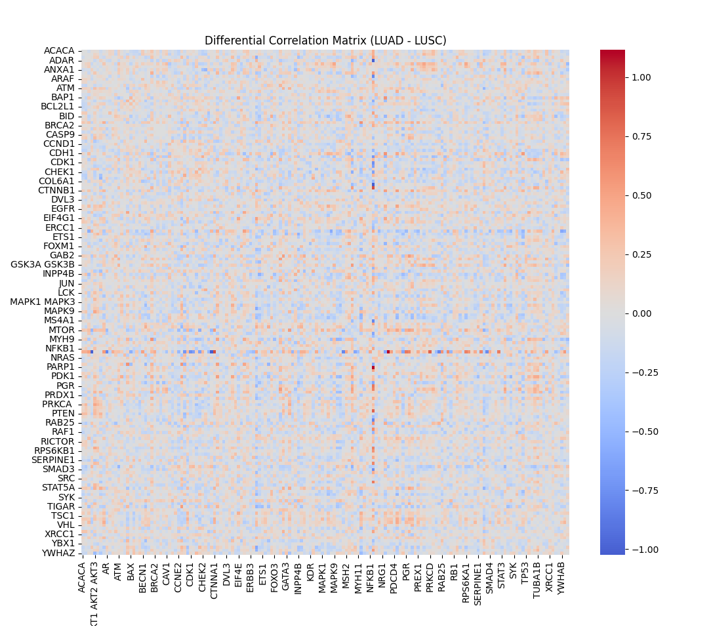


Fig 12: Correlation matrix

In TCGA Lung cancer, correlation heatmap was computed for protein analysis of cancer subtypes-LUAD & LUSC. This correlation heatmap shows the highly correlated protein genes.

RNASeq gene expression log2 ratio values were normalized and standardized. DNA methylation data numerical values were converted into beta values and then later standardized. Copy Number Variations (CNV) log2 ratio values were taken and then standardized.Both CNV and DNA are then combined together to get the complete gene expression multi-omics data.

**3.1.3. Probe annotation and feature selection**

The analysis of microarray data heavily depends on the crucial annotation step known as probe annotation. The process of probe ID mapping to biological information allows researchers to interpret gene expression results. This research analyzed probe IDs by assigning them biological descriptors which included Gene Symbols and Gene IDs as well as Gene Names.

**Gene Symbols:** The genetic identification system utilizes Gene Symbols as standardized abbreviations to distinguish different genes.

**Gene IDs/Hugo Symbols:** Unique numerical or alphanumerical identifiers specific to each gene.

**Gene Names**: The utilized dictionary uses Descriptive names to explain gene functions and their purposes.

In colorectal cancer, accomplishment of the task used the hgu133plus2.db package. This Bioconductor annotation package delivers complete mappings of probes to genes through their symbols together with their IDs and names for our microarray platform.

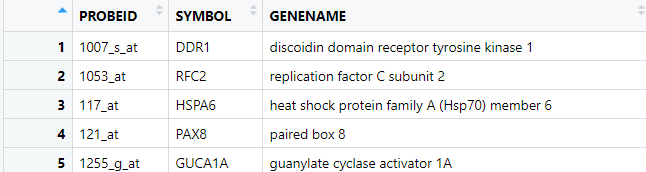


Fig 13: Probe annotation for colorectal cancer

Lung Cancer Gene Symbols and GeneIDs:-

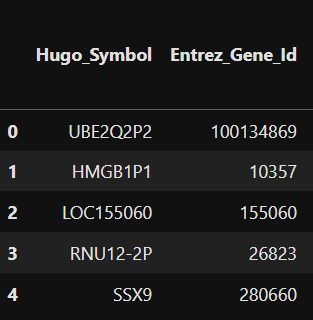


Fig 14: Gene symbols and GeneIDs for Lung cancer

The colorectal dataset underwent a filtering process following the annotation step with parameters that applied strict conditions.

Different gene symbols that remained unannotated were deleted from analysis to achieve full interpretation of downstream probes.

The analysis removed probes which linked to more than one gene entry. The objective behind this step was to achieve unique gene identification which improved the reliability of subsequent differential expression analysis results.

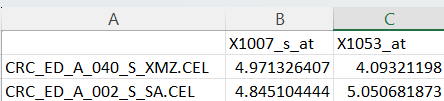


Fig 15: Gene Symbols

In Lung cancer, the genes were filtered using Variance Threshold as the criterion. The gene expressions with lower values than variance threshold were dropped and remaining highly significant gene expressions were selected.

The dataset transformed into a high-quality research base through these data filtering procedures which retained uniquely mapped confident probe data points.

* + 1. **Differentially Expressed Genes**

The objective of differential expression analysis consists in revealing genes that display substantial expression level differences between colorectal cancer samples and normal samples and also revealing the substantial expression level differences in its lung cancer subtypes. The analysis reveals disease molecular processes while detecting possible biomarkers for detection.

* 1. **Methodology**

The research design matrix enabled the evaluation of colorectal cancer expression profiles as a separate category from normal expression profiles through Condition = 0 and Condition = 1 classifications. The statistical model used for analysis is created through this matrix design.

The Limma package enabled differential expression analysis for normalized microarray data which researchers performed. The linear modeling capabilities of Limma gained additional power through eBayes moderation because this technique provides improved stability of gene expression variances. The researcher designed a contrast matrix that detailed the comparison of conditions in order to recognize genes with significant expression changes.

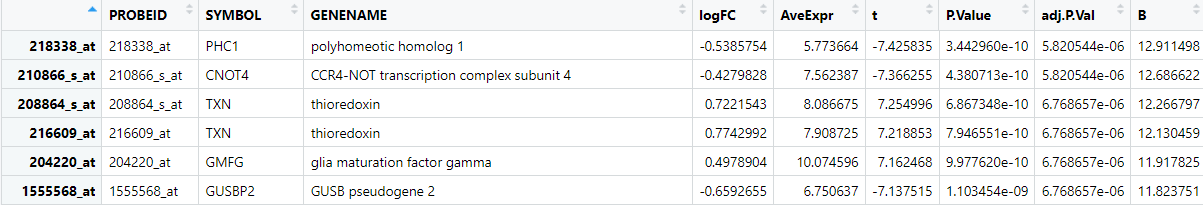


Fig 16: Differential expressed gene (DEGs) for colorectal cancer

In lung cancer, multi-omics data were taken of its subtypes-LUAD and LUSC. The data was used for classification into its subtypes. LUAD was encoded as binary value 0 and LUSC was encoded as binary value 1.



Fig 17: RNA Sequential Gene Expression data

DNA Methylation HM27 and HM450 gene expression data:-

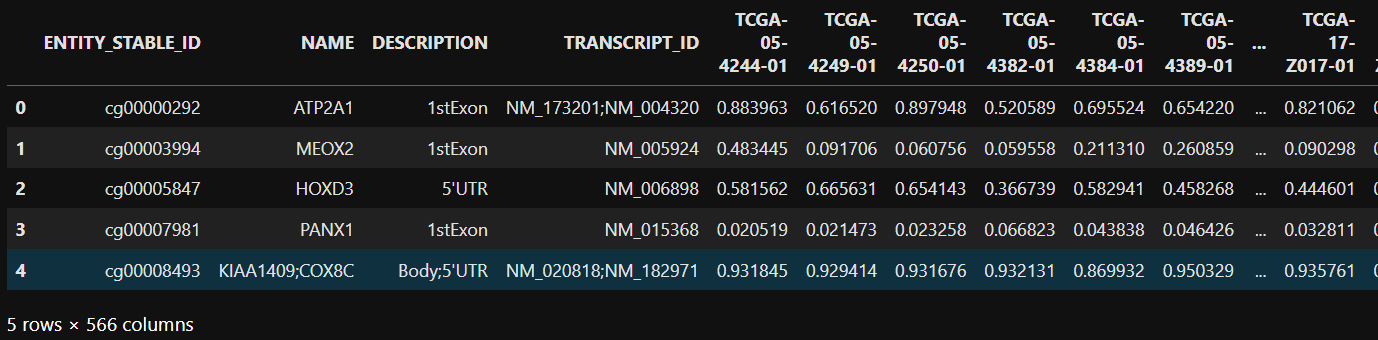


Fig 18: DNA Methylation HM27 and HM450 gene expression data

CNA(Copy Number Variations) gene expression data:-

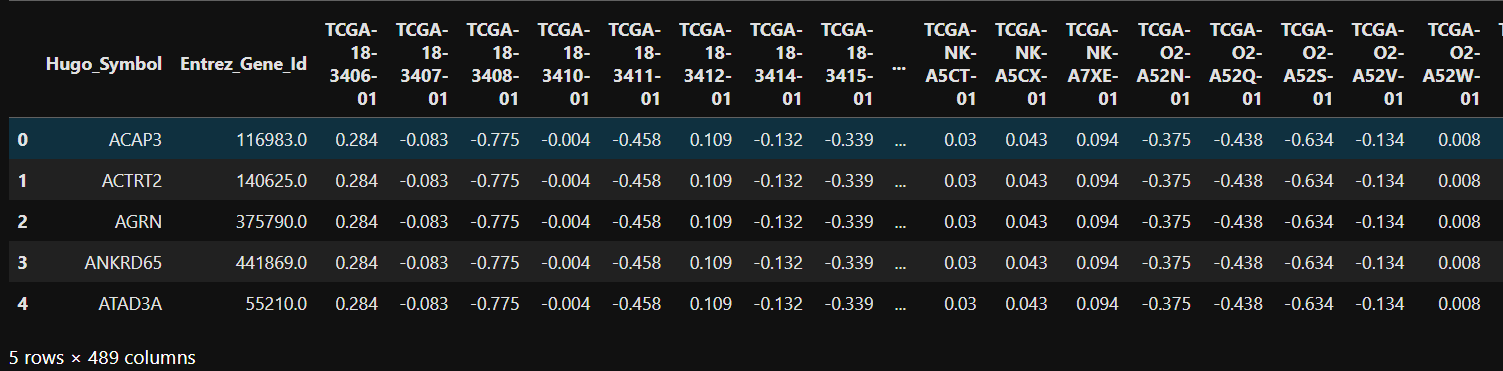


Fig 19: CNA(Copy Number Variations) gene expression data

Reverse Phase Protein Array(RPPA) data:-

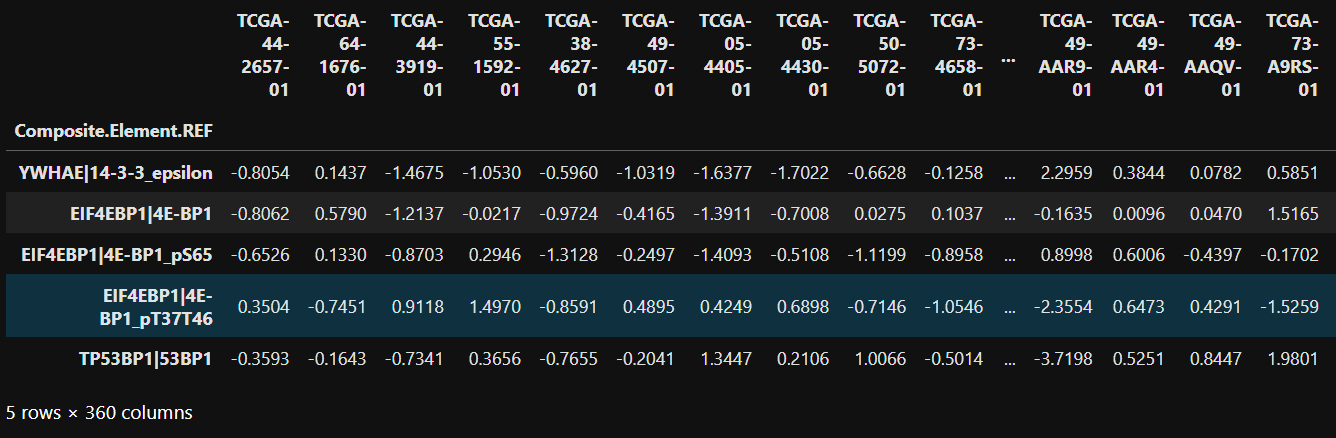


Fig 20: Reverse Phase Protein Array(RPPA) data

We used the clinical data to label the patientIDs into its subtypes.

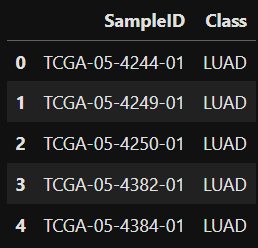


Fig 21: Clinical Sample IDs

* + 1. **Key Statistics**

Total Probes Analyzed: 43,122.

DEG Selection Criteria:

Adjusted p-value (FDR) < 0.05.

Log fold change (logFC) threshold > 0.5.

The applied thresholds enabled the identification of 4,748 DEGs within the dataset which included both upregulated and downregulated genes.

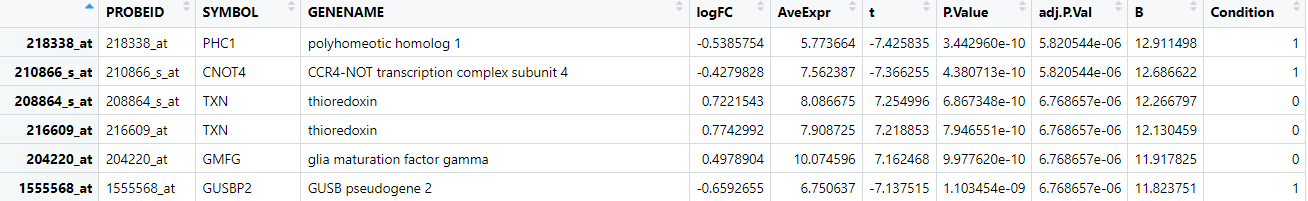


Fig 22: Colorectal Cancer Data

* + 1. **Significant Gene Attributes**

The characterization of the selected DEGs utilized these main annotation elements:

The examination tool PROBEID establishes unique attributes to identify single genes or transcripts.

The official symbols like (PHC1, CNOT4 and TXN) function as data cross-referencing methods across biological databases.

Each gene received a functional descriptive name as its GENENAME designation.

The measurement of gene expression changes between conditions appears as Log2 fold change in the results.

The expression level of TXN gene increases by 0.77 within colorectal cancer specimens relative to healthy tissue.

The expression levels of PHC1 molecules decreased compared to normal samples based on its negative value of -0.53.

AveExpr: Represents the average expression value for each gene across all samples. Extremely stable genes can be distinguished from those with high expression variability through this metric.

A more powerful indicator showing differential expression evidence exists within the t-statistic and larger absolute values indicate robust statistical evidence.

A P.Value raw measure provides information about the chance that the noticed difference emerged by random occasion.

The Benjamini–Hochberg correction applied through adj.P.Val determined adjusted p-values for controlling the false discovery rate (FDR). Research identified statistical significance using genes that exhibited an adj.P.Val under 0.05.

Log-odds ratios of differential expression act as supplementary data for designating genes as significantly expressed among other candidates.

Genes receive annotations about the conditions where differential expression becomes evident (specified as "Normal" or "Colorectal Cancer").

In TCGA lung cancer data, the significant gene attributes were the TCGA sampleids in RNASeq data, DNA, CNA ,RPPA numerically transformed data.

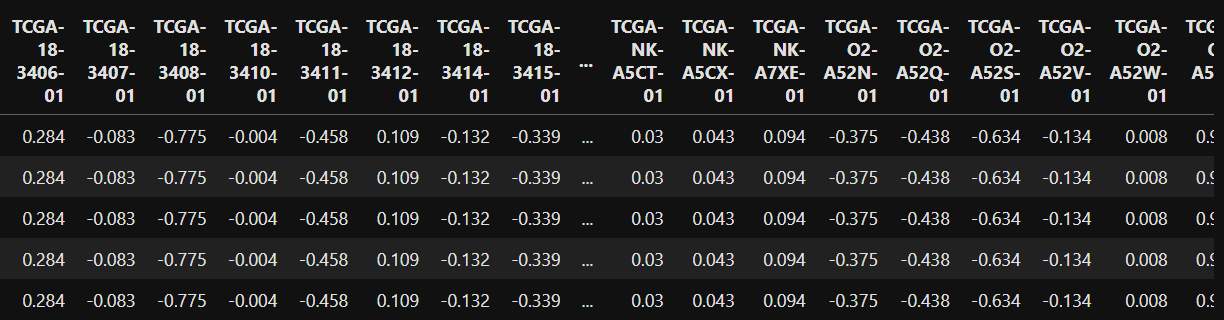


Fig 23: TCGA Sample Patient ids

* + 1. **Results and Visualization of DEGs and Protein Lung Cancer Data:-**

Visualizing recently analyzed genes helps scientists easily understand both their significant statistical values and biological meanings. The study focuses on colorectal genes showing high fold changes combined with low p-values because these elements strongly indicate differential expression and also focuses on the correlation analysis of protein pairs lung cancer subtypes. The implementation of volcano plots creates a visual understanding between statistical significance and fold change evaluation for thousands of analyzed colorectal genes and the implementation of correlation matrix plots for lung cancer protein pairs to determine the strong correlated protein pairs.

* + 1. **Volcano Plot**

The volcano plot divides genes into four distinct groups depending on their statistical significance and expression modifications.

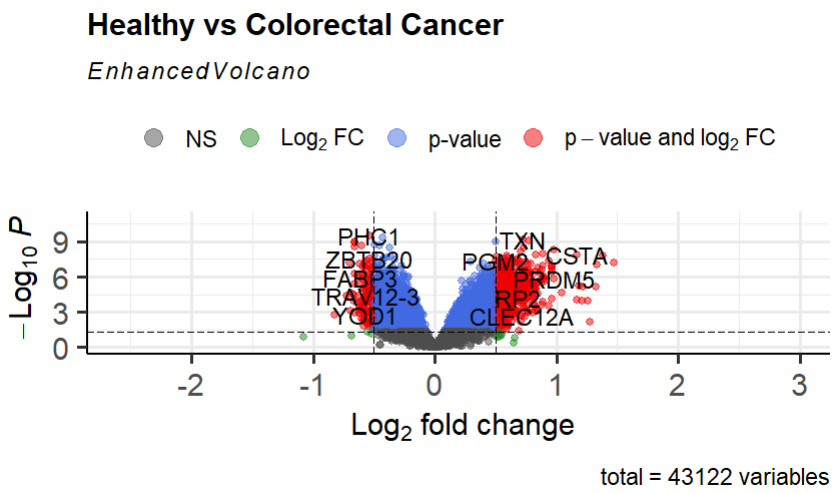


Fig 24: Violin Plot for Healthy v/s Colorectal Cancer

The group of NS (Not Significant, Grey Dots) includes genes which do not pass the specified statistical threshold requirements for analysis.

Genes with meaningful expression changes can be found among the green dots although they did not pass statistical significance tests which would require reconsideration when using less strict cut-off points.

The blue dots represent the statistically significant genes which show minimal expression level changes despite their meaningful contribution to overall biological processes.

Red-dotted genes are significant targets due to their statistical significance (adjusted p-value < 0.05) and logFC > 0.5 threshold fulfillment for future investigation.

* + 1. **Correlation Heatmap**

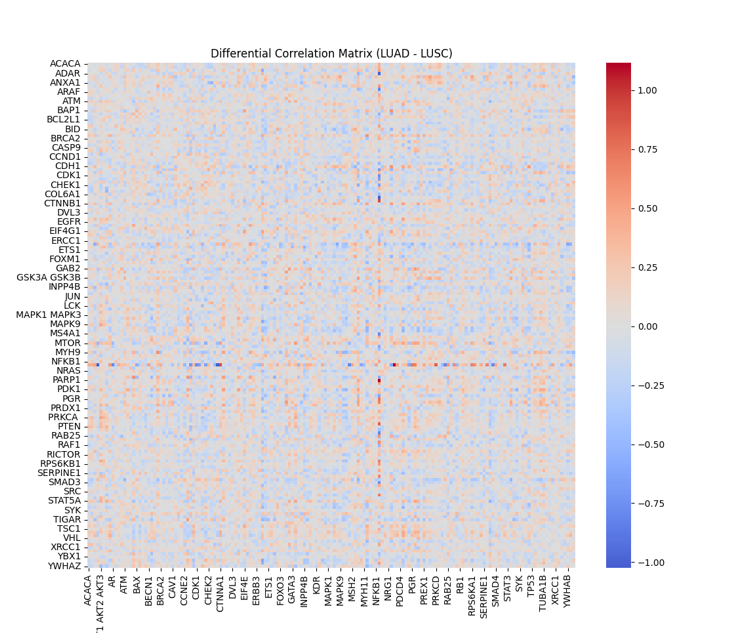
****

Fig 25: Correlation matrix for Protein Pairs of Lung Cancer Subtypes

The correlation matrix was computed using pearsonr library of python scipy statistics module.The above visualization of correlation matrix shows us how much these protein pairs are highly correlated to each other.

The purpose of computing correlation matrix was to compute the p-value matrix of the protein pairs using pearsonr to filter the highly significant protein pairs(edges).

Benjamini Hochberg multiple correction statistical test was carried out using python statsmodel library to control the FDR(False Discovery Rate) of proteins in p-value matrix.

* + 1. **Observations**

Observations:

• A total of 43,122 variables were analyzed.

• Genes such as PHC1, CN0T4, GUSBP2, and SKAP1-AS2 are highlighted as significantly differentially expressed.

• Upregulated genes (right side) and downregulated genes (left side) indicate potential biomarkers or targets for further investigation in colorectal cancer research.

• Up regulated genes show higher expression in the condition of interest compared to the control. It means the gene is more active. They promote growth or prevent cell death.

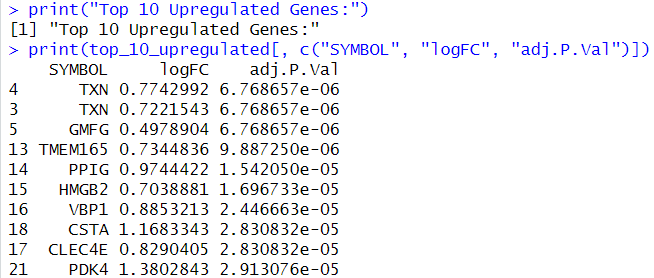


Fig 26: Top 10 unregulated gene symbols

• Down regulated genes show lower expression in the condition of interest. It means the gene is less active. These are the genes that includes tumor suppressors, whose inactivity allows cancer progression.

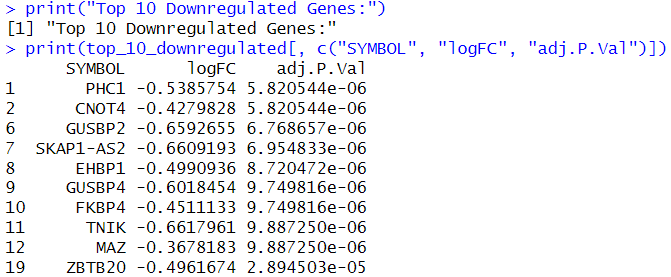


Fig 27: Top 10 Downregulated genes

In case TCGA data, the top differential protein genes were identified after the PPI network graph and KEGG pathway analysis.

* + 1. **Top 10 DEGs**

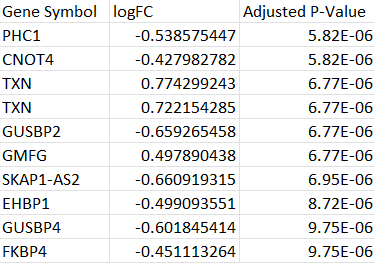
****

Fig 28: Top 10 DEGs

These genes show the most statistically significant and biologically meaningful changes in expression between the normal and cancer samples.

**Significance (Adjusted P-Value):** These genes have the lowest adjusted p-values, meaning the likelihood of observing these expression differences by random chance is very low.

**Magnitude of change (Log Fold Change - logFC):** These genes exhibit the largest positive or negative changes in expression levels. A higher absolute value of logFC indicates a stronger difference in expression between these conditions.

**Biological Insights**

**•** Identified genes involved in colorectal cancer pathways.

• Potential biomarkers for early detection and targeted therapy.

**Gene Enrichment Analysis**

It is a computational method used to determine whether specific genes, identified from an experiment are overrepresented in certain biological pathways, molecular functions or cellular components. This helps researchers interpret the biological meaning of their gene lists in the context of existing biological knowledge.

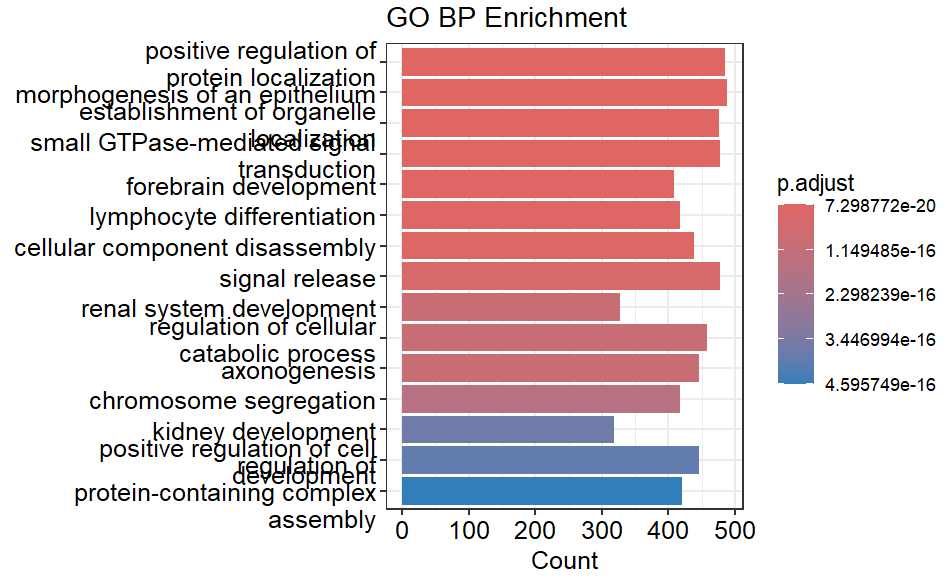
****

Fig 29: GO BP Enrichment Analysis of Genes

The above bar plot shows “Gene Ontology Biological Process” enrichment analysis, showing biological processes that are significantly enriched in the dataset.

In Y-axis, These are the genes that are overrepresented like

• Positive regulation of protein localization

• Morphogenesis of an epithelium

• Forebrain development

• Axonogenesis

• Chromosome segregation

In X-axis, The length of each bar represents the number of genes in dataset that are involved in the respective biological process.

Conclusion, “Positive regulation of protein localization” and “Chromosome segregation” indicates potential functional or regulatory mechanisms influenced by genes.

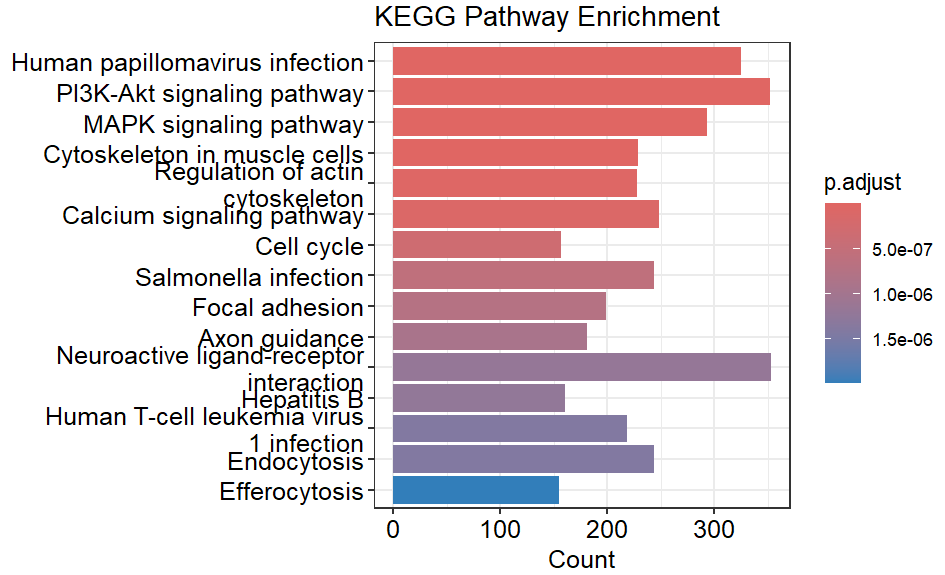
****

Fig 30: KEGG Pathway Enrichment Analysis

The above bar chart shows the KEGG Pathway Enrichment Analysis, highlighting that are significantly enriched in gene set.

In y-axis, These are the biological pathways where the genes in dataset show significantly involvement like

• Human papillomavirus infection

• PI3K-Akt signaling pathway

• MAPK signaling pathway

• Regulation of actin cytoskeleton

• Cell cycle

In x-axis, Each bar’s length reflects the number of genes from the dataset participating in that pathway. “Human papillomavirus infection” pathway has the highest count, suggesting strong gene representation.

Conclusion, pathways such as “PI3K-Akt signaling pathway” and “MAPK signaling pathway” are often associated with key processes like cell growth, survival and differentiation.

Pathways like “Cell cycle” emphasize cellular regulation, which may be crucial in cancer-related studies.

Pathways like “Human papillomavirus infection” suggest potential links to viral mechanisms influencing gene expression.

* + 1. **Protein–Protein Interaction (PPI) Network**

It is used to visualize and analyze the relationships between proteins in the colorectal dataset.

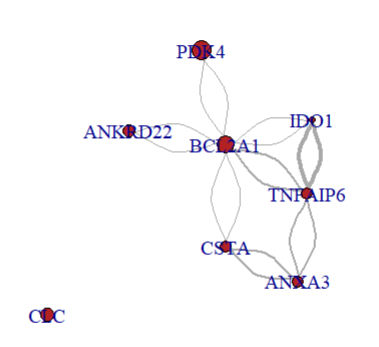


Fig 31: PPI Network

This is the graph for these proteins in the dataset.

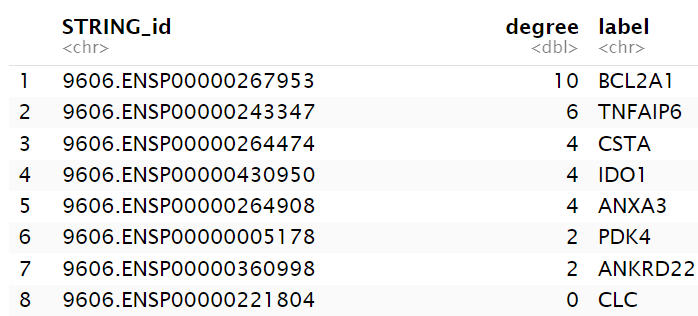


Fig 32: Top Matched Genes from StringDB

We can identify that BCL2A1 as the top hub gene, which has the highest degree 10, indicating that it is highly interconnected with other DEGs.

TNFAIP6 has degree of 6 and CSTA, IDO1, and ANXA3 each had a degree of 4.

High-degree genes are considered biologically important because they represent key regulatory nodes involved in disease mechanisms.

Hub genes such as BCL2A1 and TNFAIP6 can be biomarkers for colorectal cancer identification.

In case of TCGA RPPA cancer dataset, the highly significant correlated protein pairs were matched with the known protein pairs from string Database and then the edges of the Protein pairs were visualized using network graph.

The top hub proteins were derived from the network graph by degree. Degree is the measure of number of edges of a protein in network.

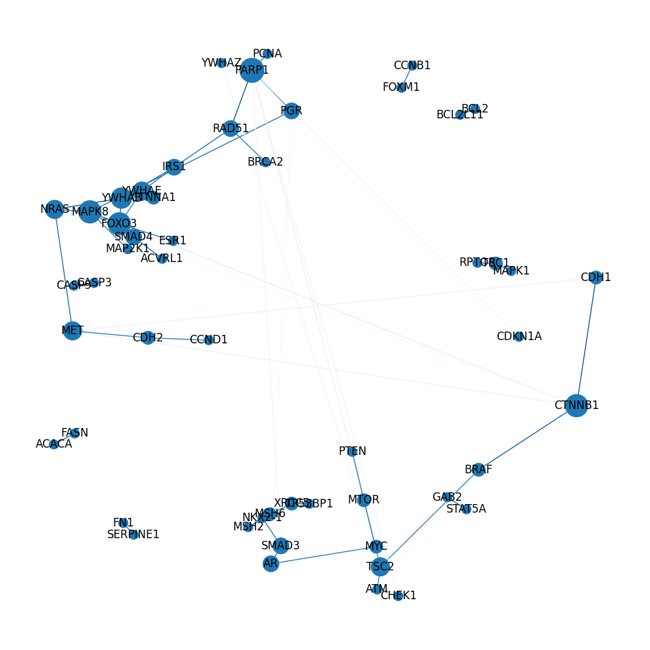
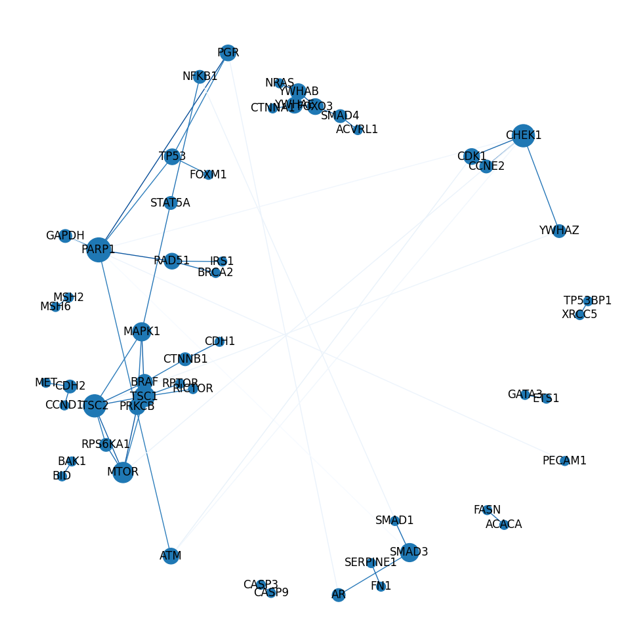
 

Fig 33: Network Graph Visualization of LUAD and LUSC

Top Hub proteins by degree:

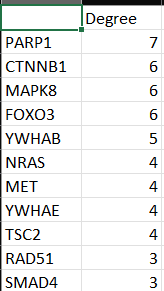
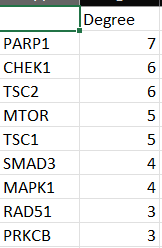
 

Fig 34: Top Hub Proteins by degree

Top Highly Correlated Protein Pairs:-

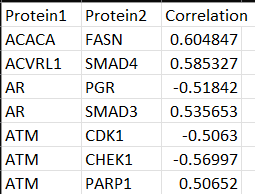
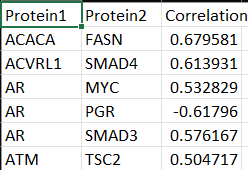
 

Fig 35: Top Highly Correlated Protein Pairs

* + 1. **Centrality Analysis**

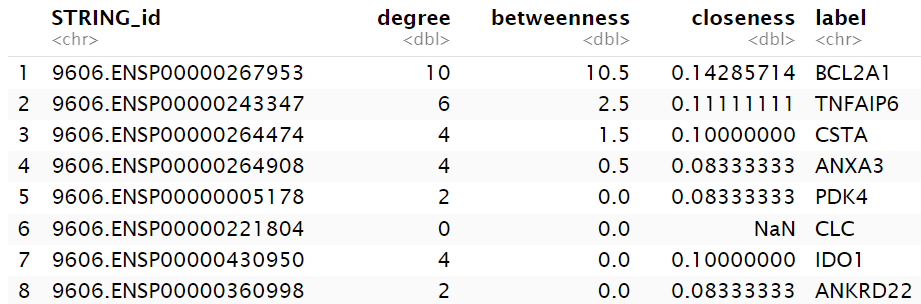
****

Fig 36: Centrality Analysis Of genes

Centrality analysis revealed that BCL2A1 had the highest betweenness centrality (10.5) and degree (10), indicating its role as a key bottleneck and hub in the network.

High betweenness centrality implies that BCL2A1 lies on many shortest paths between other nodes, suggesting it may regulate communication and signal flow within the network. This highlights BCL2A1 as a potentially important regulator among DEGs.

TNFAIP6 also showed moderate betweenness (2.5) and a degree of 6, implying it may also play a significant but less dominant role in network communication.

Nodes like CSTA and ANXA3 showed low betweenness values (1.5 and 0.5) indicating more local connectivity but less global control over network flow.

Genes such as PDK4, IDO1 and ANKRD22 has low degrees and zero betweenness while CLC has a degree of 0. Thus, it is an isolated node with limited influence in the network.

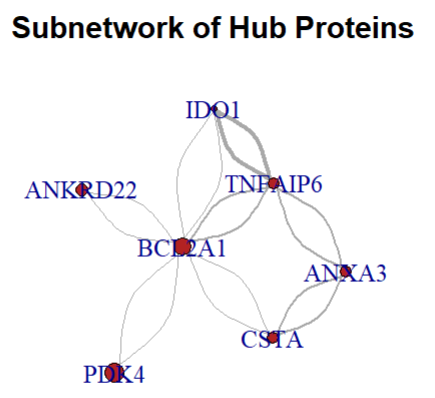


Fig 37: Network graph of top hub proteins

As shown in the figure, BCL2A1 as the core hub in the network, potentially playing a major regulatory role. TNFAI6 and ANXA3 form secondary connections suggesting they may cooperate with BCL2A1o involved in similar biological processes.

* 1. **Gene Ontology Biological Process (BP) enrichment and KEGG pathway enrichment analysis**

To perform enrichment analysis, the hub gene symbols need to converted into standardized identifiers.

The bitr() function from the clusterProfiler package was used to map the hub genes from gene symbols to Entrez UDs, using the org.Hs.eg.db human gene annotation database.

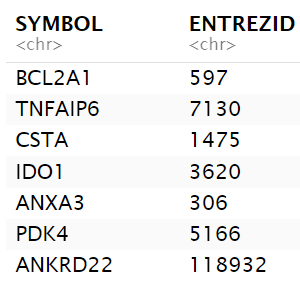


Fig 38: Symbol and GeneId of Proteins

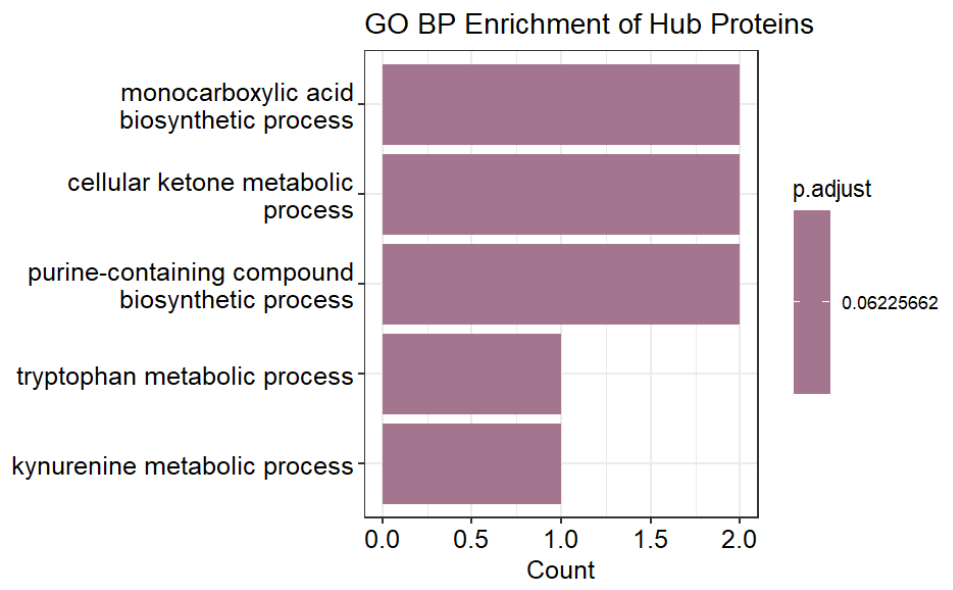


Fig 39: GO BP Enrichment-2

The hub genes were first mapped to Entrez IDs and enrichment was performed focusing on the Biological Process (BP) category. A p-value cutoff of 0.1 and 1-value cutoff of 0.3 were applied to retain terms of potential biological relevance.

The bar plot displays the top five enriched biological processes:

• Monocarboxylic acid biosynthetic process

• Cellular ketone metabolic process

• Purine-containing compound biosynthetic process

• Tryptophan metabolic process

• Kynurenine metabolic process

These processes suggest that the hub proteins are primarily involved in metabolic and biosynthetic pathways, and are particularly related to small molecules and amino acid metabolism.

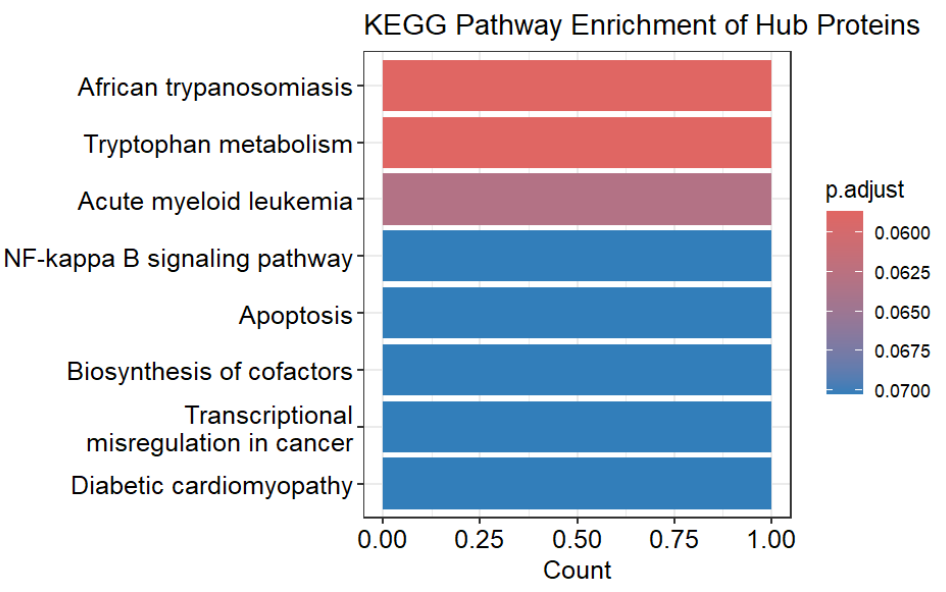


Fig 40: KEGG Enrichment-2

For further investigate the functional roles of the hub proteins, KEGG pathway enrichment analysis was performed using the enrichKEGG() function from the clusterProfiler package.

The bar plot shows top enriched KEGG pathways associated with the hub proteins:

• African trypanosomiasis

• Tryptophan metabolism

• Acute myeloid leukemia

• NF-kappa B signaling pathway

• Apoptosis

• Biosynthesis of cofactors

• Transcriptional misregulation in cancer

• Diabetic cardiomyopathy

NF-kappa B signaling pathway, Apoptosis, Biosynthesis of cofactors, Transcriptional misregulation in cancer are critically involved in immune response, programmed call death and cancer development.

Also, the identification of tryptophan metabolism is consistent with previous findings in the GO enrichment, highlighting its potential role in biological functions of the hub proteins.

For the TCGA RPPA Lung cancer data, KEGG pathway enrichment analysis of top 20 hub proteins was carried out to identify the top differential proteins. The pathway enrichment analysis helps the researchers to determine the drug targets for cancer therapy.

LUSC PPI Enriched data:-

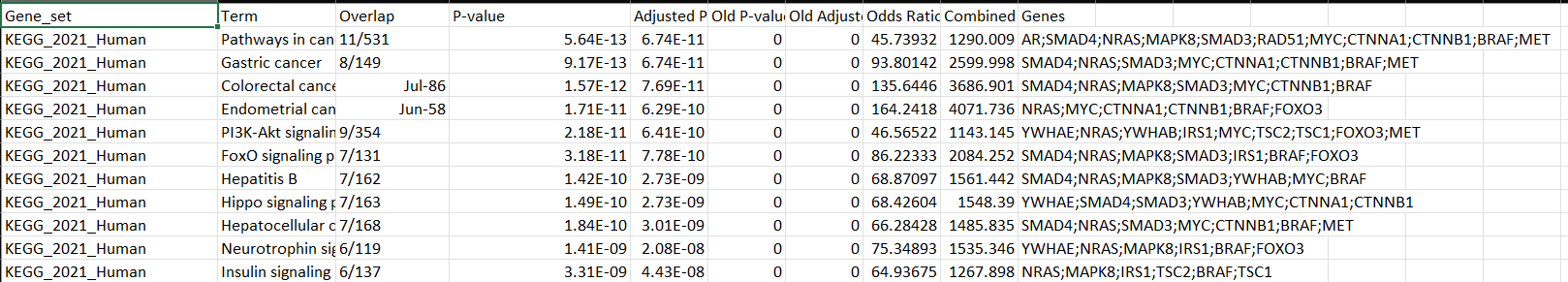


Fig 41: KEGG PPI Enriched data LUSC

LUAD PPI Enriched data:-

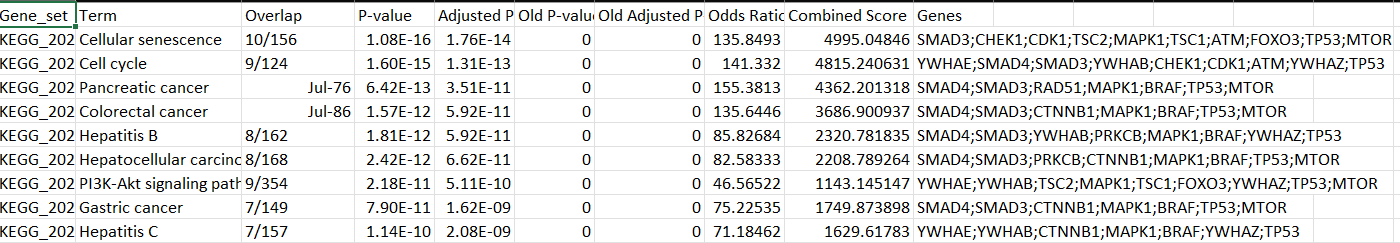


Fig 42: KEGG PPI Enriched data LUSC

From the KEGG pathway analysis, the top most significant recurring genes are TP53, SMAD3, MAPK1, and MTOR in LUAD PPI enriched data as they have lowest p-values and adjusted p-values showing higher significance power for drug target identification.

Similarly, RAS/MAPK (NRAS, BRAF), TGF-β (SMAD4/SMAD3), Wnt/β-catenin (CTNNB1), and PI3K-Akt-mTOR (TSC1/2, FOXO3) are the top recurring genes in KEGG pathway analysis with lower p-values and showing higher significance for target drug identification.

* 1. **Machine Learning model**

**Random Forest**

Implements random forest to check the importance of each feature in DEGs dataset.

Input features: "LOGFC", "AVEEXPR", "T", "P.VALUE", "ADJ.P.VAL"

Output label: ‘’Condition’’

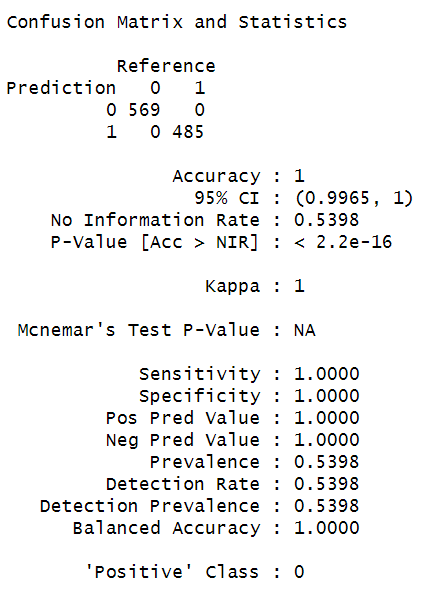


Fig 43: Evaluation Metric-1

**Machine Learning Model for TCGA Cancer Data:-**

Multilayer Perceptron was implemented in the following steps:-

* + - 1. Wrapper function is used from scikeras library to integrate keras models with scikit learn pipelines.
      2. A neural network model was defined using Keras Sequential API which includes:-

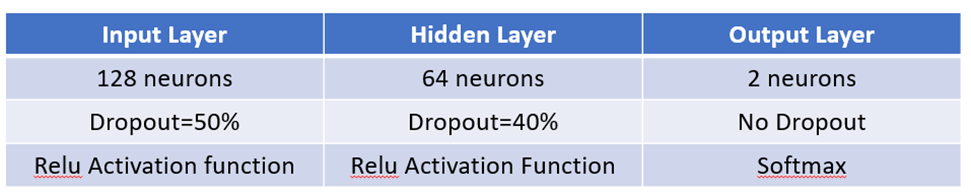


Table 1: Neural Network Model

Learning rate:0.001

Loss function:Log loss function

Optimizer:Adam, Cross Validation: Stratified 10-fold Cross Validation

The model was compiled using Adam optimizer and categorical cross entropy loss (log loss function).

Results:-

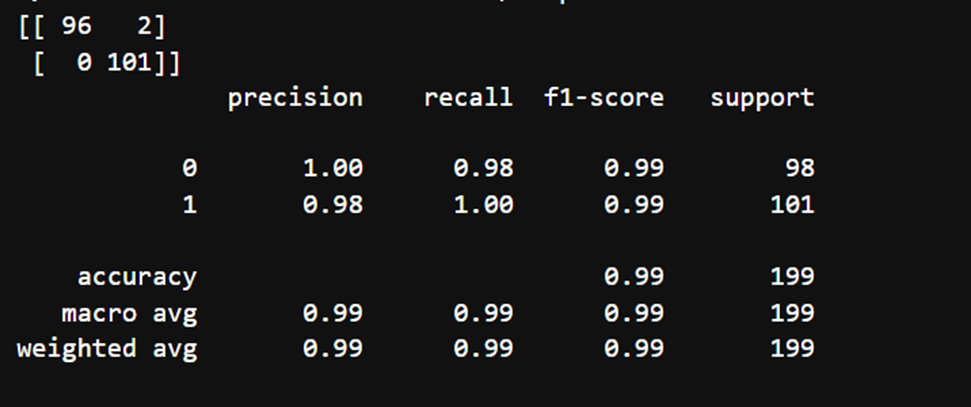


Fig 44: Evaluation Metric-2

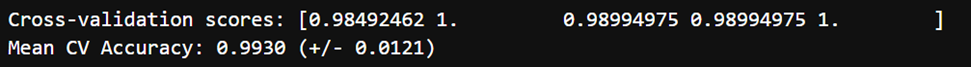


Fig 45: Evaluation Metric-3

The mean cross validation accuracy is 99.30 indicating the best performing model on the validation data with standard deviation of 0.0121.

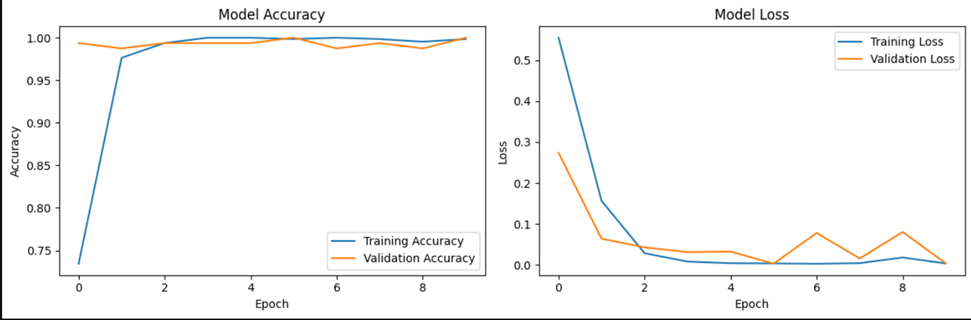


Fig 46: Model Training and Loss

In the given graph, for every iteration model training and validation is increasing and the training and validation loss is decreasing indicating the potential improvement in correctly classifying the NSCLC biomarker instances.

CNA+DNA Model Implementation:-

**Logistic Regression**

Logistic Regression is a statictical model used for binary classification process. It predicts the probability of a class label (LUSC or LUAD) based on input features.

The steps implemented in Logistic Regression model:-

1. Data split into training set and testing set.

The core of the logistic regression was the logistic sigmoid function:-

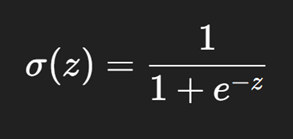


Fig 47: Sigmoid Function

The sigmoid function is used to output the probability range between 0 and 1.

During model training, model learns the best parameters that minimize the loss function such as Categorical Cross Entropy loss function(log-loss function).

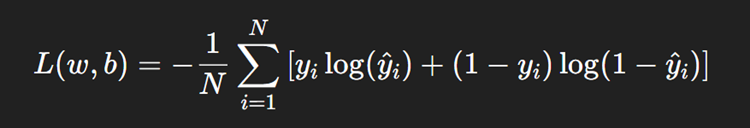


Fig 48: Log loss function

The solver is used in the model i.e. lbfgs which is an optimization algorithm to minimize the loss function.We have trained the model upto 1000 iterations to find the optimal solution.

Evaluation Metrics:-

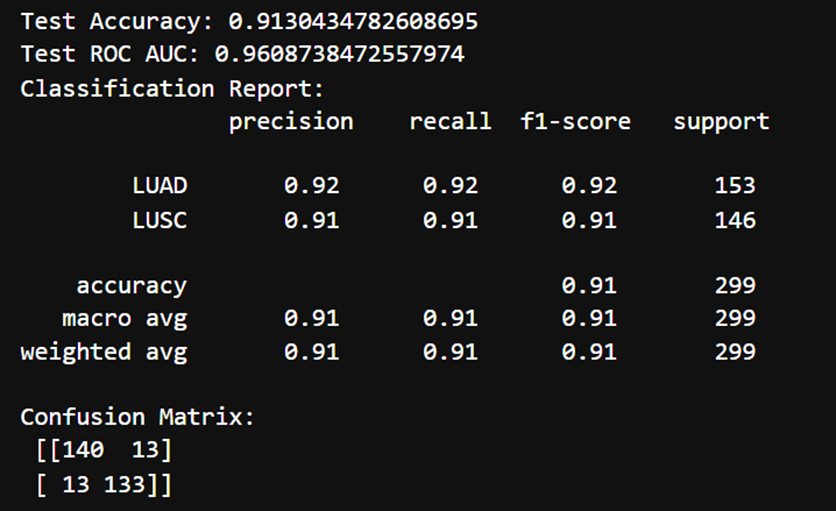


Fig 49: Evaluation Metric

**MLP Classifier:-**

MLP Classifier is a neural network model implemented with backpropagation to correctly classify the NSCLC instances into its two classes LUAD and LUSC.

Architecture:-

Input layer: 50 neurons

Hidden layer: 128 neurons , Activation Function:Relu

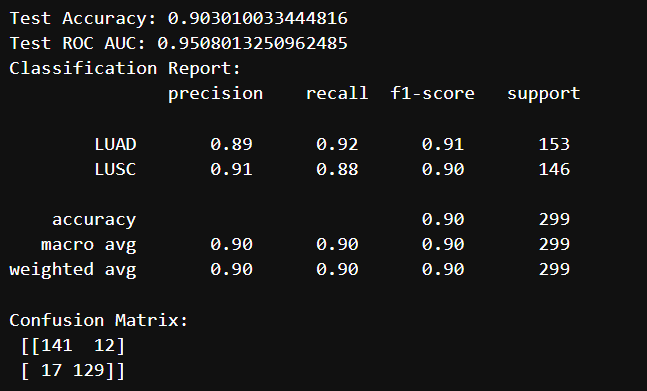
Hidden layer:64 neurons, Activation Function:Relu

Output Layer: 2 neurons, Activation Function: Softmax

Adam optimizer is used to optimize the learning rate of the MLPClassifier. Adam optimizer combines the advantage of both AdaGrad and RMSProp which is efficient for large datasets.

The model is trained on 300 epochs with 5-fold cross validation to assess the generalizability of model.

**Evaluation Metrics:-**

****

**XGBoost Model:-**

XGBoost is a highly efficient and scalable implementation of gradient boosting, an ensemble technique that creates decision trees one-at-a-time to improve model performance. Each tree attempts to reduce the errors from the previous tree. Thus, it is a source of power for classification and regression tasks.

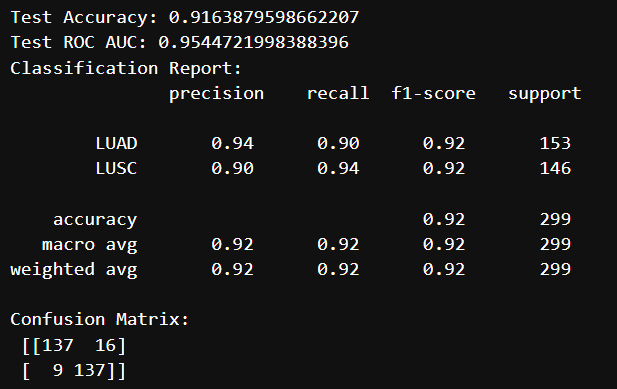
XGBClassifier was implemented from the xgboost library implemented for multi-class classification tasks. It uses multi softmax function that output the labels directly. In XGBoost, we have kept the depth of each decision tree till 6 to capture more complex patterns in the gene expressions data more efficiently. About 100 decision trees were created to enhance the model performance and boosting.

Multi-class logarithmic loss function was used as an evaluation metric that measures the performance of the model. 5-fold cross validation was implemented to assess the model’s performance on the validation data.

**Boosting Mechanism**

The model is trained in additive manner where new trees were fitted on the residuals of the previous trees to gradually reduce the loss. Trees are built and predictions are updated after each iteration using gradient descent.

**Evaluation Metrics:-**

****

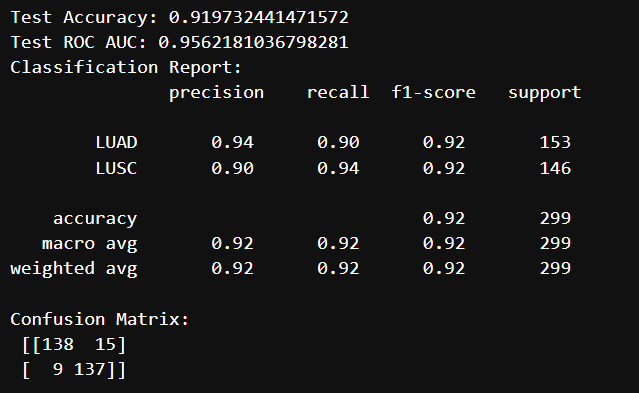
**RandomForest Classifier:-**

Random Forest is an ensemble learning technique in which various decision trees were constructed and their predictions are aggregated together to produce a more robust and accurate model.

Random Forest classifier is implemented using scikit learn ensemble library.We built the ensemble model of 100 decision trees. Limited the depth of each decision tree to 6 helping the model to prevent overfitting.

RandomForest Classifier was implemented with 5-fold cross validation to assess the generalization of the ensemble model.

**Evaluation Metrics:-**

****

**Model Selection:-**

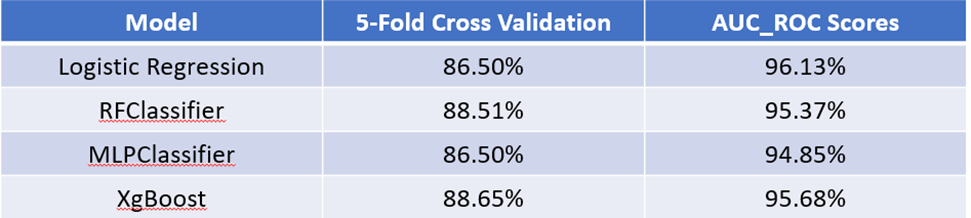
****

Table 2: Model Comparison

As, it can be seen from the results of the model Logistic Regression is the best model as it has better 5-fold cross validation accuracy and highest AUC-ROC scores among all the models.

* 1. **Explainable Artificial Intelligence techniques**

Applying SHAP,

Using fastShap library,

Calculating SHAP values for each features.

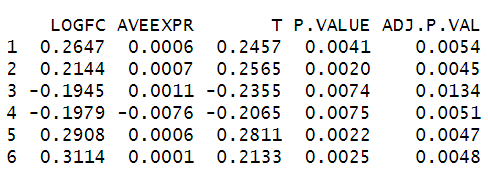
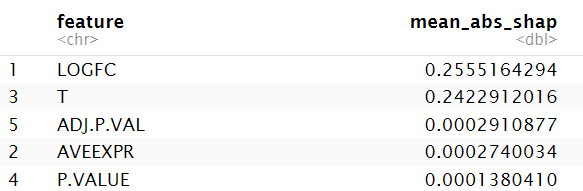


Fig 50: SHAP values

Calculating mean shap values



Plotting Beeswarm plot,

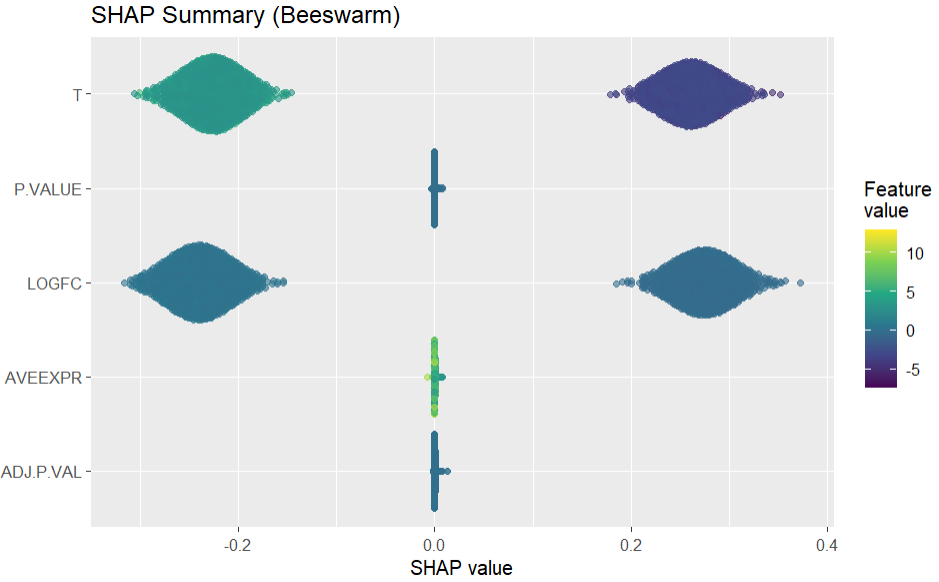


Fig 51: Beeswarm Plot

For t-static, a wide spread of SHAP values is on both sides and higher T values which are green in color push predictions higher and lower T values which are purple in color push predictions lower. So, T is a strong positive contributor to the model. High t-values increases model output. Similarly Log fold change is also strong predictive feature which upregulated genes (positive logFC) increase predicted probability. Whereas P.value, average expression (Aveexpr) and Adj.p.val have very low effect on model prediction.

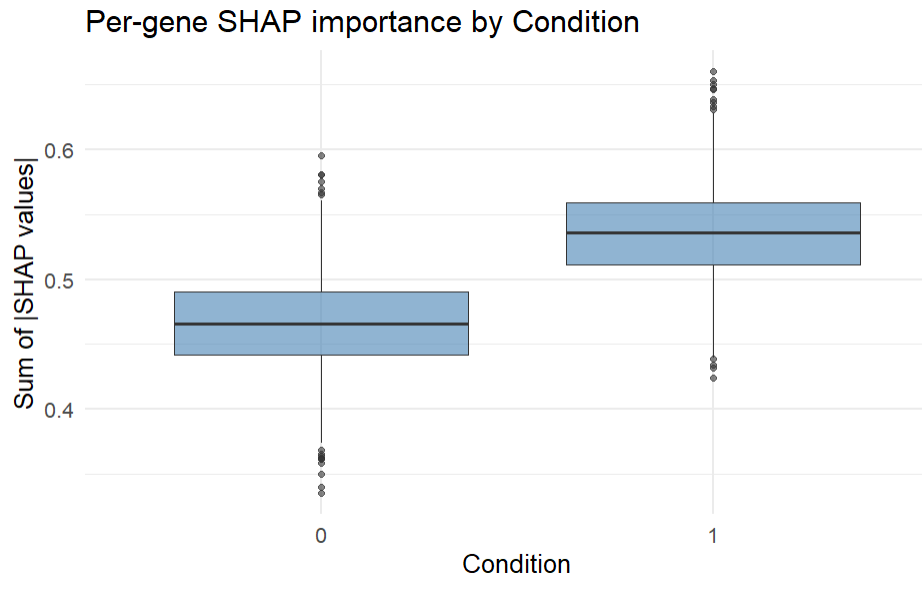
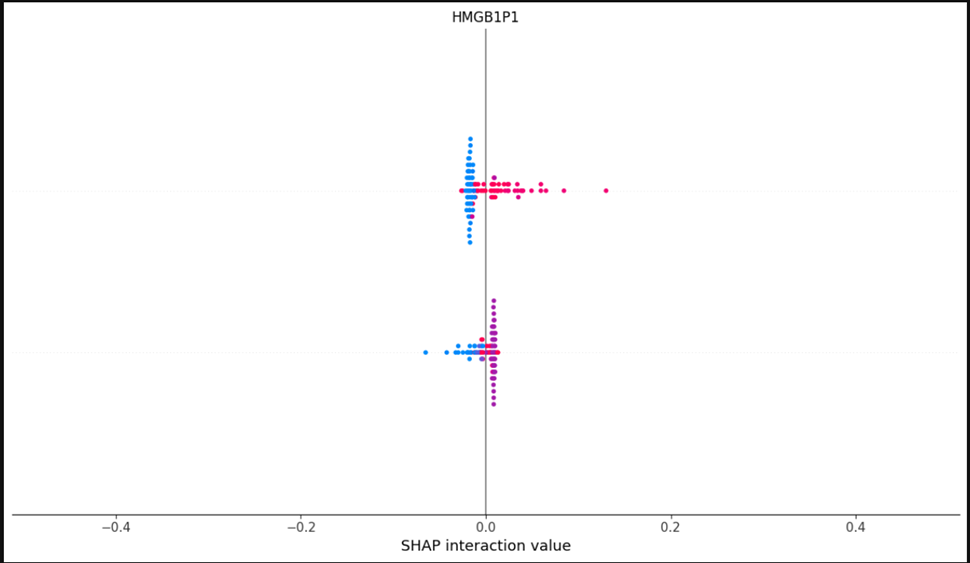


Fig 52: Per Gene SHAP Importance by Condition

This boxplot visualizes the pre-gene SHAP importance by condition 0 vs 1 based on the sum of absolute SHAP values per gene.

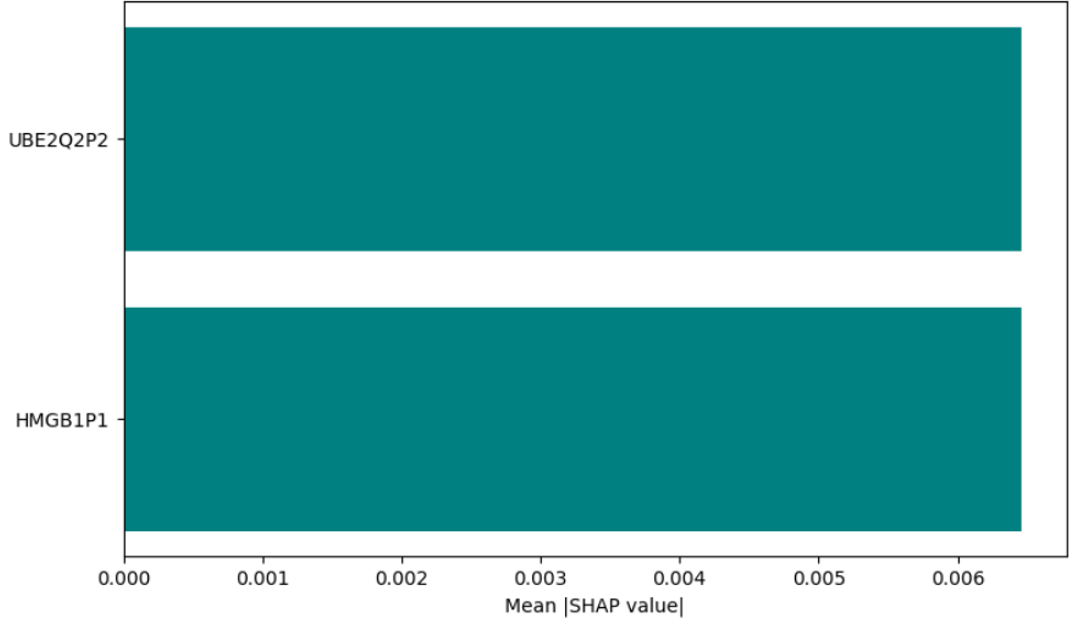
The median SHAP value sum is higher for condition indicating that the model relies more heavily on genes under this condition to make predictions. Condition 1 contributes more distinctive patterns to the model compared to Condition 0. It shows that gene expression patterns under this condition will provide more discriminative information.

**SHAP results for RNA Sequential Model (MultiLayer Perceptron Model):-**



This SHAP plot shows the **interaction effect** of the gene **HMGB1P1.**

The X-axis in this plot shows how much the gene **HMGB1P1** is contributing on the interaction effect on the model’s output.Values closer to 0 in SHAP plot means less contribution. Positive values push the predictions towards LUSC class and negative predictions pushes the predictions towards the opposite class LUAD.The 100 samples of random background data was taken to plot the feature importance.



Out of all the 100 samples taken, top two genes **UBE2Q2P2** and **HMGB1P1** shows the significant contribution in decision making for classifying between NSCLC subtypes.

**LIME Plot for RNA Sequential Model:-**

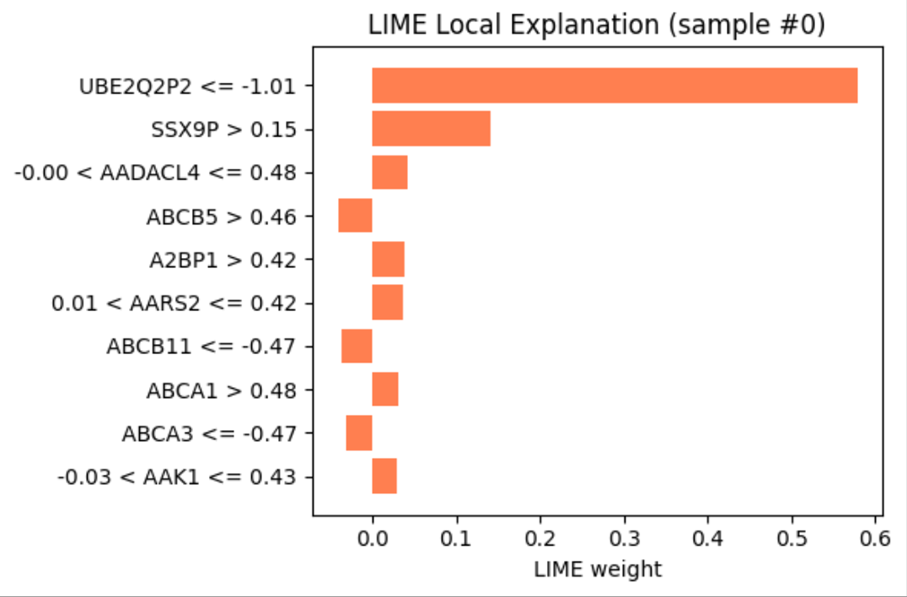
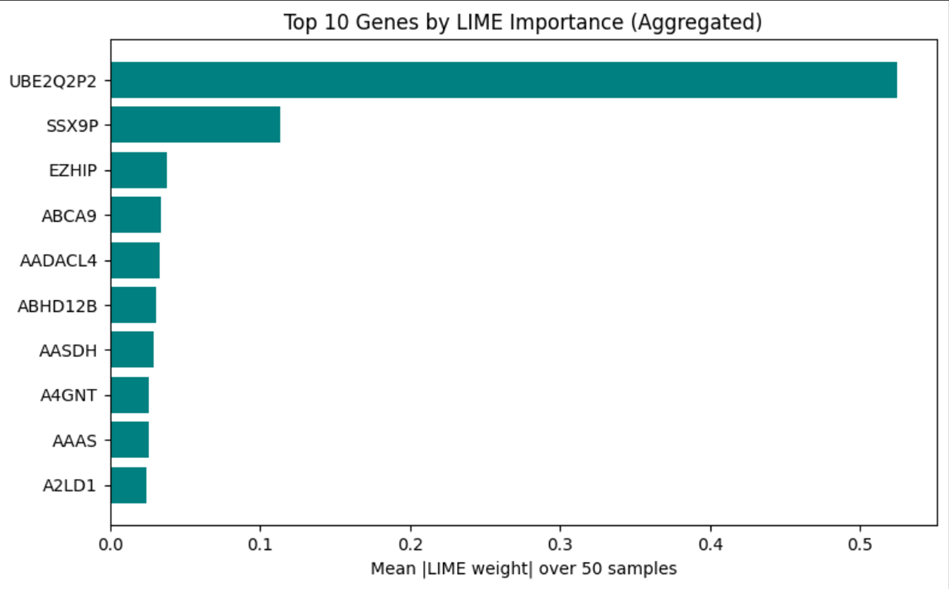


Fig 53: LIME Plot

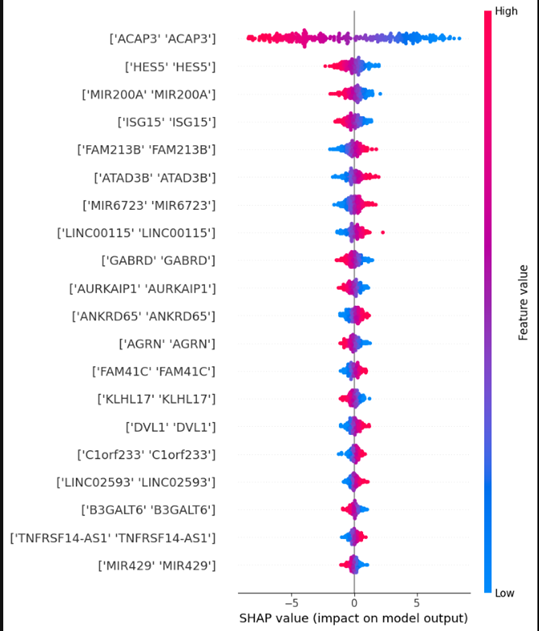
**UBE2Q2P2 and SSX9P** has the highest significant impact on the MLP model’s decision making. On the Y-axis, the range of values of genes represents the range of gene expression values in which the model is showing the high impact in model’s decision making.

Top gene importance absolute mean of LIME Weight:-



LIME approximates the model’s behavior locally by perturbing input samples and observing changes in predictions. The absolute weights (|LIME weight|) reflect feature importance.

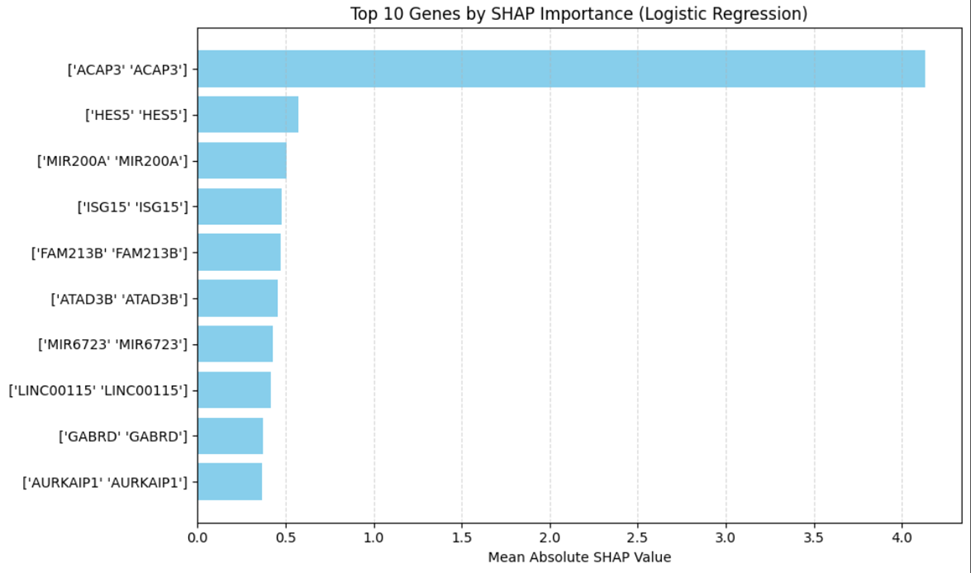
SHAP Results for Logistic Regression:-



The SHAP (SHapley Additive exPlanations) plot shows the impact of individual features ( DNA + CNA genes) on the prediction of a logistic regression model.Each dot in the SHAP plot represents a SHAP value for a gene in a sample of 100 gene data.

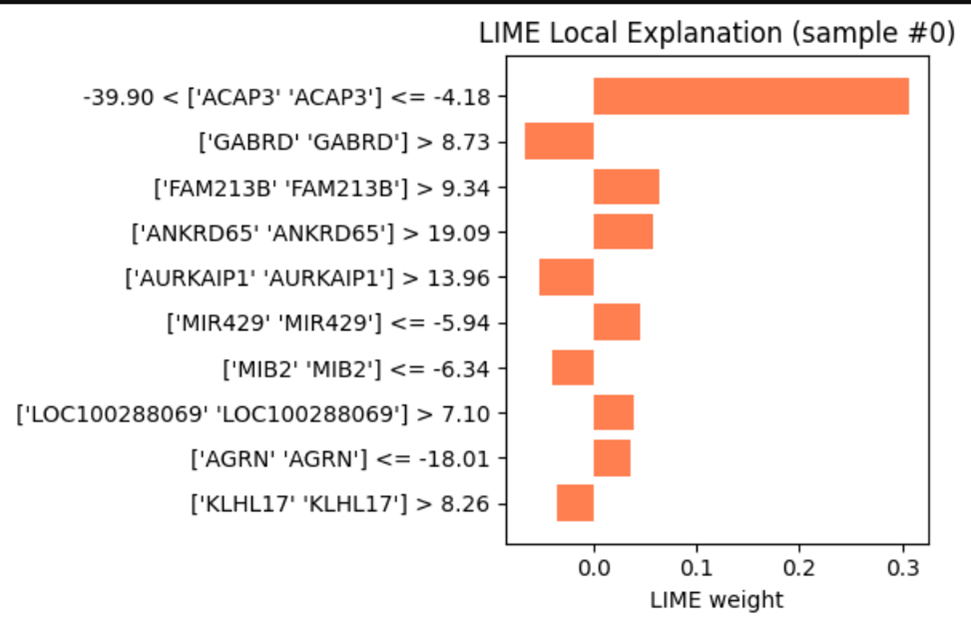
From the interpretation, it can be seen that:-

**ACAP3** and **HES5** are highly impactful because high values of **ACAP3** (pink dots) on the left decreases the prediction and low values of **HES5** (blue dots) on the right increase the prediction.A positive SHAP value increases the prediction and negative SHAP value decreases the prediction.**MIR200A** also correlates with the positive SHAP values hence increasing the model prediction. It may assist in biomarker discovery or therapeutic target identification by revealing which genetic alterations are most responsible for the predictions made by the model.



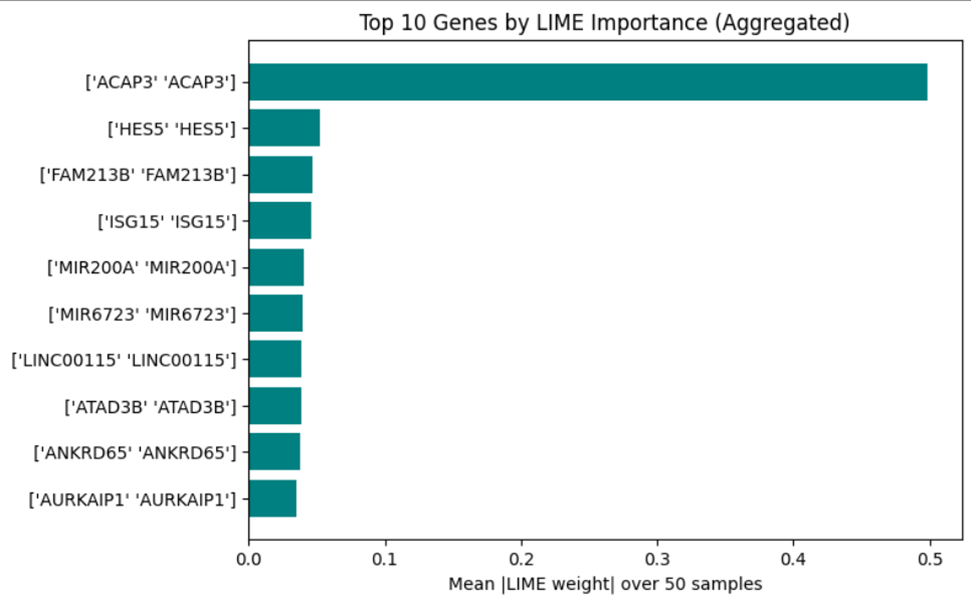
Top 10 genes making significant impact on the logistic regression model’s predictions.

**LIME Importance for Logistic Regression Model:-**

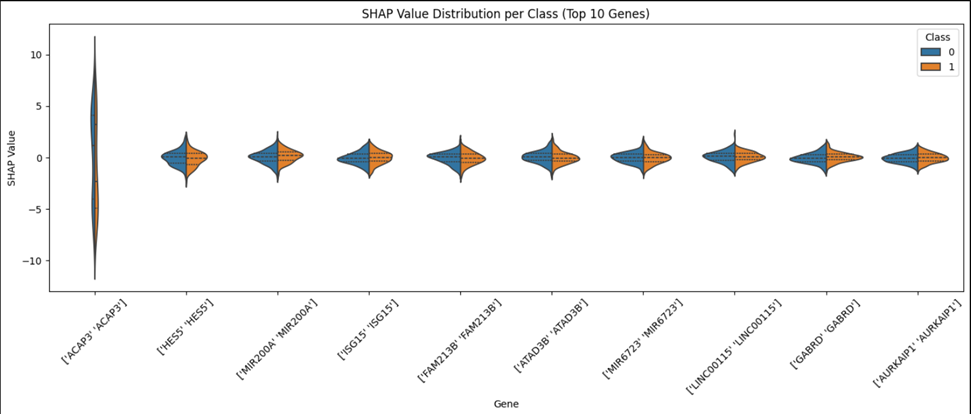
****

**ACAP3 and GABRD** have the significant impact on the model’s predictions within the specific range of DNA and CNA expression values.

Top Genes by LIME weights:-



**Violin Plot:-**

****

Violin plot is showing the the distribution of gene level contributions to Logistic Regression model’s predictions separated by cancer subtypes: LUAD(class 1) and LUSC(class 0) tumor cells.

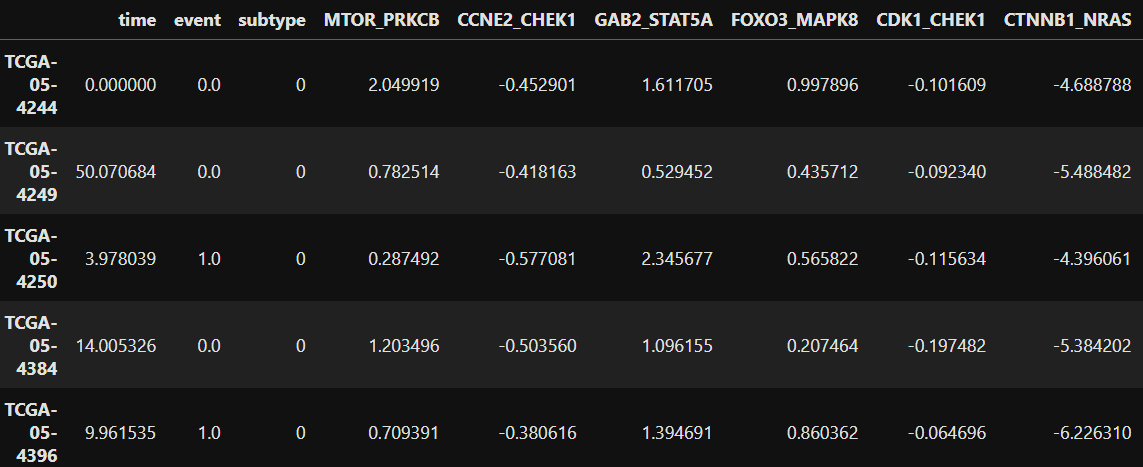
* 1. **Survival Analysis of top differential Proteins:-**

The Cox Proportional Hazards (CoxPH) model is a semi-parametric survival analysis method that can be utilized to assess the impact of multiple group of variables against the time to the occurrence of an event (e.g. death, relapse).

Here, this method is applied to look at time to survival based on a model constructed using the expression levels of the top differentially expressed genes as well as event occurrence, survival time, and subtype as variables.

For the input data into the model, following features were taken from the meta data and TCGA dataset :-

* Top Differential Gene expression values(calculated previously from PPI network analysis)
* Time of Survival in months
* Event
* Subtype:0(LUAD) and 1(LUSC)

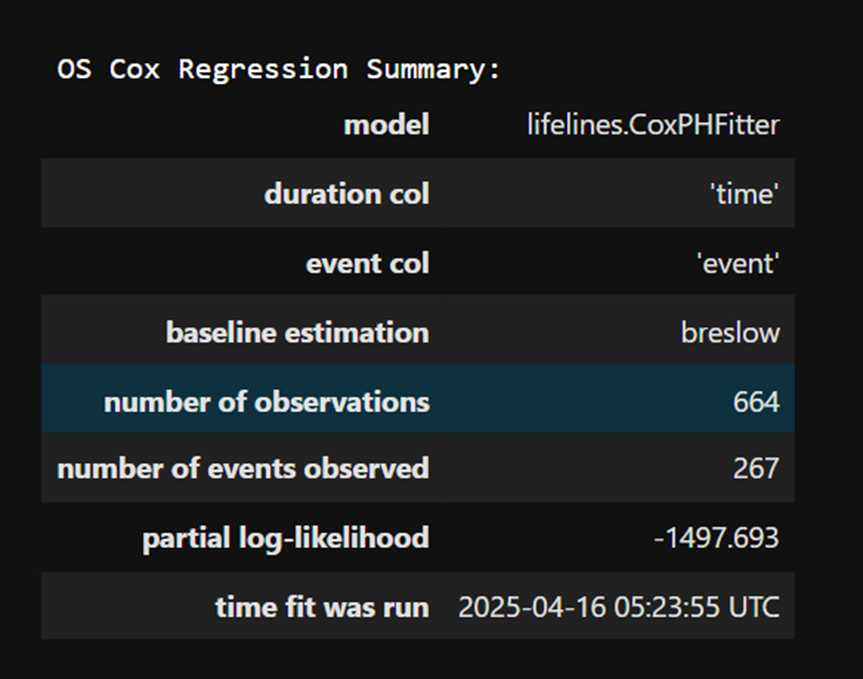


Occurrence of event depends upon the type of survival.There are two types of survival:-

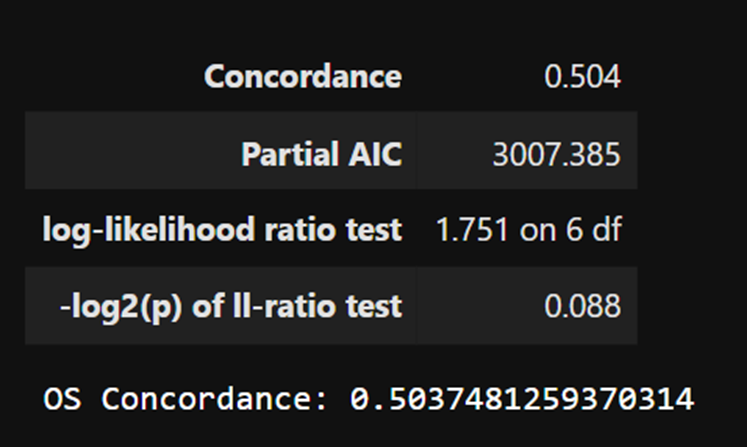
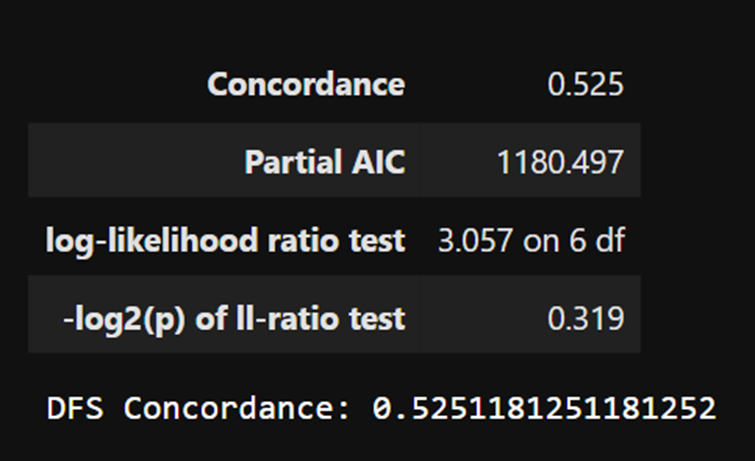
**Overall Survival(OS):**It is the length of time from diagnosis until death from any cause.

**Disease Free Survival(DFS):**The length of time after successful treatment during which a patient survives without any signs or symptoms of cancer.

Cox Regression model is fitted on two types of data i.e. Overall Survival data and Disease Free Survival data.

Concordance ratio of Cox Regression model:-

Concordance ratio shows how well the model predicts the order of event times.

Higher the concordance better is the model in survival prediction.

**KM Plot:-**

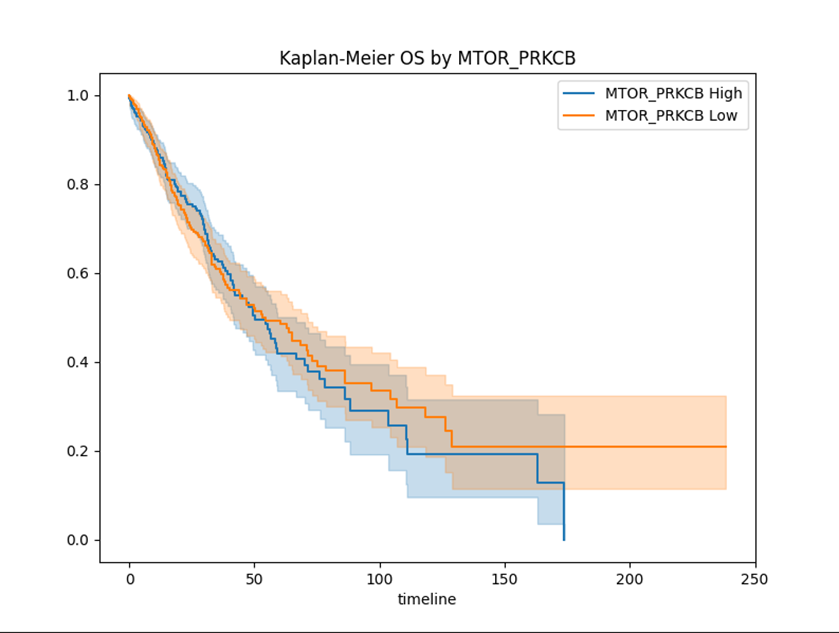


Fig 54: KM Plot

This Kaplan-Meier overall survival (OS) plot visualizes survival probabilities over time for two groups stratified by expression levels of the MTOR\_PRKCB axis — derived from a Cox regression model.

X-axis (timeline): Time (in months) from diagnosis or treatment.

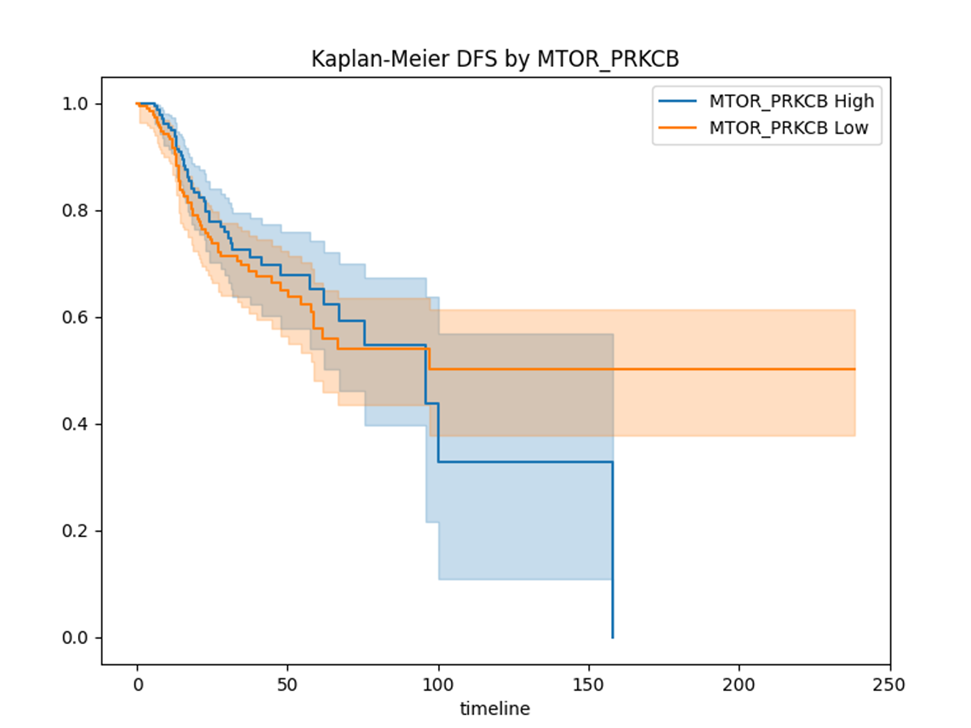
Y-axis (Survival Probability): Proportion of patients that are alive (1.0 = 100% alive).

Two curves, blue and orange:

Blue (**MTOR\_PRKCB High**) - patients with high combined expression of **MTOR** and **PRKCB**.

Orange (**MTOR\_PRKCB Low**) - patients with a low expression of the **MTOR\_PRKCB** pair.

The shaded areas are confidence intervals for the survival estimates.This can highlight the MTOR\_PRKCB axis as a potential prognostic biomarker or therapeutic target.



X-axis (timeline): time after treatment or diagnosis.Y-axis (Survival Probability): the chance of being disease-free.

Blue curve: patients with high **MTOR\_PRKCB** expression.

Orange curve: patients with low **MTOR\_PRKCB** expression.

Shaded areas: confidence intervals (CI).

Biological Implications:-

High **MTOR\_PRKCB** expression is associated with faster recurrence or progression, aligning with its possible role in tumor growth or resistance. Low expression suggests better prognosis, with longer periods free from disease.

**Risk Score Plot:-**

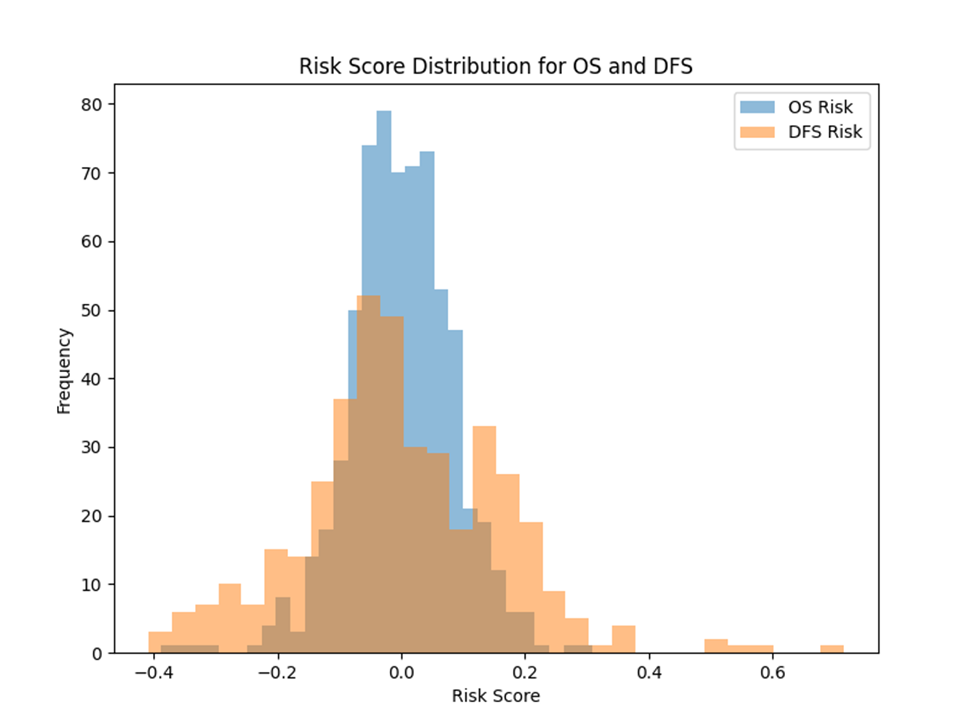
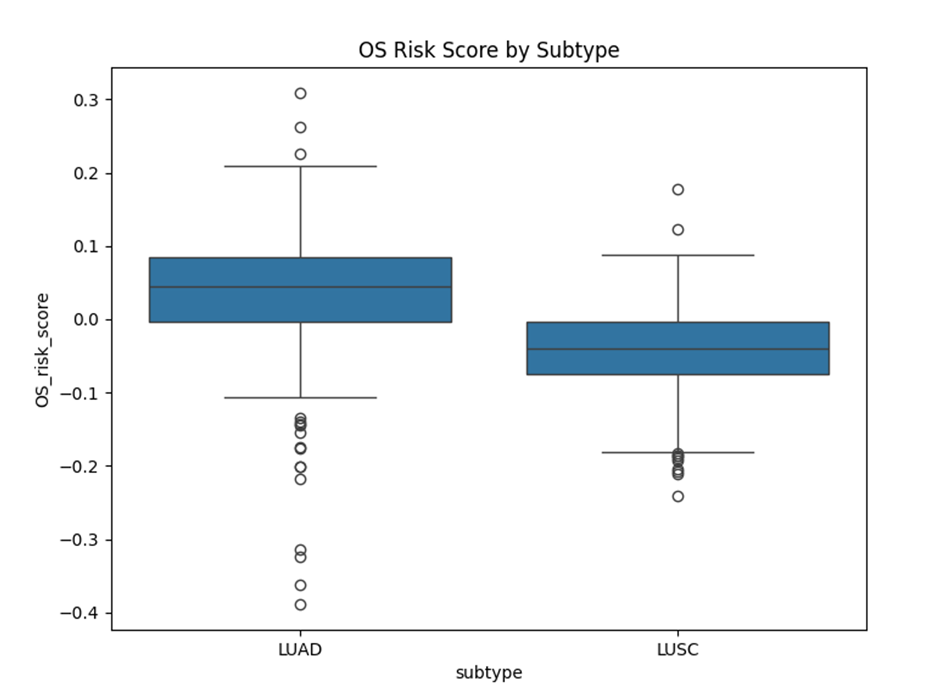


Fig 55: Risk Score Plot

This is the risk score density plot for OS v/s DFS.

The Frequency on the Y-axis indicates the proportion of patients falling in each risk bin and the X-axis shows us the risk scores in each bin.Higher the risk rate less probability of survival rate.

**Boxplot:-**

****

This is the plot for Overall Survival risk score by Subtype LUAD and LUSC. The outliers above the range -0.1 to 0.1 indicates high risk of survival. And the outliers below the range has lower risk of survival. The data points within the box are in the moderate range of risk.

**Survival Probability Plot:-**

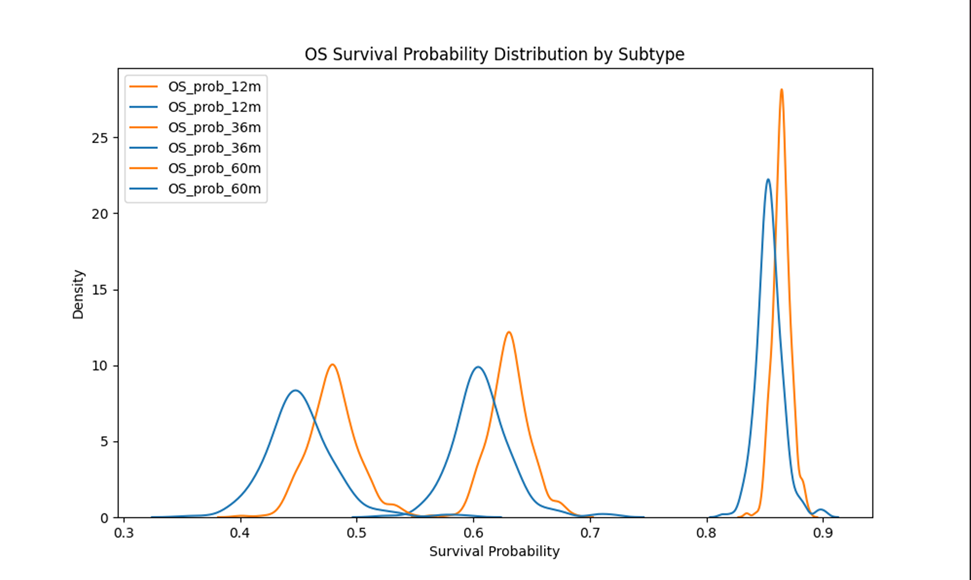


Fig 56: KDE Plot

This plot visualizes the predictions made by cox regression for three checkpoints of months:12 months, 36 months and 60 months for the two subtypes of cancer: LUAD(blue) and LUSC(orange). The Y-axis represents density which is the relative likelihood of a given survival probability occurring in each subtype.

At 12 months:

LUSC (orange) has peaks shifted to the leftward direction which suggests the lower density of survival rate.LUAD (blue) has peaks at ~0.42 and ~0.6 and represents a blend of moderate risk and improved outcomes.

At 36 and 60 months:

LUSC shows peaks which shift to the right , which points towards improved long-term survival.LUAD has distributions closer to ~0.6 and ~0.85, with more stable outcomes of survival.

1. **LIMITATIONS AND FUTURE ENHANCEMENTS**
   1. **Limitations**

Although the research was successful in detecting differentially expressed genes (DEGs) and classifying NSCLC instances into its lung cancer subtypes, studying their biological meaning using statistics, enrichment, survival analysis and machine learning techniques, a number of limitations exist:

**Dataset Size and Diversity**: This study is based on the colorectal cancer dataset (E-MEXP-3756) from ArrayExpress and its context may not be representative of diverse clinical conditions and genetic backgrounds. These results may be different with heterogeneous and larger datasets.The diversity and large size of the multi omics data such as RNA Sequential data, methylation data ,CNV data and RPPA data makes the preprocessing difficult and computationally expensive.

**Microarray technology** platforms have been known to lack dynamic range and sensitivity in comparison to other emerging technologies such as RNA-Seq, which may be capable of providing finer resolution insights.

**Annotation Bias**: Probe annotation depends on existing gene databases. Probes that map to out-of-date or poorly annotated genes can cause inaccuracies.

**Lack of Experimental Validation:** The identified biomarkers and pathway analyses are computationally derived. Without wet-lab validation (e.g., qPCR or Western blotting), biological relevance remains hypothetical.

**Limited clinical metadata**: This dataset does not contain elaborate clinical variables (e.g., tumor stage, age of patient), which limits one's capability to conduct stratified analyses or to correlate clinical outcomes with gene expression.

**Unstable:** LIME's explanations can easily vary with trivial input perturbations, raising a question about reproducibility.

**Black-box developments:** Deep learning models (e.g., MLPs) may still have unexplainable decision paths, even after going through SHAP/LIME.

**Survival Analysis:**The Cox Regression model relies on the assumption of proportional hazards, which may not be true for every identified biomarkers (e.g., time-varying effects of immune-related genes).

**KEGG Analysis:**KEGG Pathways are curated but are not tissue-specific; pathways related to lung cancer (e.g., alveolar repair) may be excluded.

**PPI Analysis:**PPI Networks highlight hub proteins but are less likely to identify context-specific interactions (e.g., tumor-microenvironment crosstalk).

* 1. **Future enhancements**

Multiple frontiers of research are put forward to increase the translational utility and robustness of the results

**Cross-Platform Validation:** Involving more than one set of datasets from different platforms, e.g., RNA-Seq, may cross-validate and enhance the reproducibility of the DEGs identified to promote applicability in multiple technical and biological contexts.

**Clinical metadata integration:** By incorporating detailed clinical variables, we will be able to perform survival analysis and model prognosis and stratified biomarker discovery, and so enhance the clinical validity and utility of the biomarkers.

**Experimental Validation:** These hub genes including BCL2A1 and TNFAIP6 need to be validated by experimental methods (e.g., qPCR and Western blotting) to verify the biological importance and therapeutic significance of these genes.

**Advanced machine learning** may provide enhanced feature extraction and predictive performance with the use of methods from deep learning that can be used to build more reliable diagnostic and predictive models.

**Multi-Omics Integration:** Integrating other omics layers like proteomics and metabolomics will give a complete and multidimensional insight into the pathogenesis and development of colorectal cancer.

**5.CONCLUSION**

This project includes the identification of biological processes and signaling pathways that were majorly implicated in presence of disease and its subtypes using Gene enrichment and pathway studies.

Machine learning and deep learning models such as Random forest, Logistic Regression and Multi Layer Perceptron model were trained to predict the subtype of NSCLC and in the protein–protein interaction network analysis which revealed hub genes such as BCL2A1 and TNFAIP6 to be promising candidates playing a key regulatory role in cancer development.

In order to improve interpretability, SHAP (SHapley Additive exPlanations) values and LIME (Local Interpretable Model-Agnostic Explanations) were employed in ranking the most significant contributing features to the predictions of subtypes of NSCLC instances and also in representation of most significant influencing features with the use of T-statistic and log fold change. These features were used as important biomarkers and emphasized the importance of genes that modulated the microenvironment and regulated the cell cycle and the immune response.Survival Analysis also played the significant role in making survival predictions using Cox Regression.

Overall, the research illustrates the power of bioinformatics and explainable AI in unearthing biomarkers of early disease detection and identification of various areas of further investigation and clinical translation.

**6. REFERENCES**

[1]Dwivedi K, Rajpal A, Rajpal S, Kumar V, Agarwal M, Kumar N. Enlightening the path to NSCLC biomarkers: Utilizing the power of XAI-guided deep learning. Comput Methods Programs Biomed. 2024 Jan;243:107864. doi: 10.1016/j.cmpb.2023.107864. Epub 2023 Oct 20. PMID: 37866126. Link: https://pubmed.ncbi.nlm.nih.gov/37866126/

[2] Lyu, F., Gao, X., Ma, M., Xie, M., Shang, S., Ren, X., Liu, M., & Chen, J. (2023). Crafting a Personalized Prognostic Model for Malignant Prostate Cancer Patients Using Risk Gene Signatures Discovered through TCGA-PRAD Mining, Machine Learning, and Single-Cell RNA-Sequencing. *Diagnostics (Basel, Switzerland)*, *13*(12), 1997. <https://doi.org/10.3390/diagnostics13121997>

[3] Xu, Y., Xu, Q., Yang, L., Liu, F., Ye, X., Wu, F., Ni, S., Tan, C., Cai, G., Meng, X., Cai, S., & Du, X. (2013). Gene expression analysis of peripheral blood cells reveals Toll-like receptor pathway deregulation in colorectal cancer. *PloS one*, *8*(5), e62870. <https://doi.org/10.1371/journal.pone.0062870>

[4] Novielli, P., Romano, D., Magarelli, M., Bitonto, P. D., Diacono, D., Chiatante, A., Lopalco, G., Sabella, D., Venerito, V., Filannino, P., Bellotti, R., De Angelis, M., Iannone, F., & Tangaro, S. (2024). Explainable artificial intelligence for microbiome data analysis in colorectal cancer biomarker identification. *Frontiers in microbiology*, *15*, 1348974. <https://doi.org/10.3389/fmicb.2024.1348974>

[5] Radhakrishnan, S. K., Nath, D., Russ, D., Merodio, L. B., Lad, P., Daisi, F. K., & Acharjee, A. (2025). Machine learning-based identification of proteomic markers in colorectal cancer using UK Biobank data. *Frontiers in oncology*, *14*, 1505675. <https://doi.org/10.3389/fonc.2024.1505675>

[6] Foersch, S., Glasner, C., Woerl, A. C., Eckstein, M., Wagner, D. C., Schulz, S., Kellers, F., Fernandez, A., Tserea, K., Kloth, M., Hartmann, A., Heintz, A., Weichert, W., Roth, W., Geppert, C., Kather, J. N., & Jesinghaus, M. (2023). Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nature medicine*, *29*(2), 430–439. <https://doi.org/10.1038/s41591-022-02134-1>

[7] Wani, N. A., Kumar, R., & Bedi, J. (2024). DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. *Computer methods and programs in biomedicine*, *243*, 107879. <https://doi.org/10.1016/j.cmpb.2023.107879>

[8] Kanan, M., Alharbi, H., Alotaibi, N., Almasuood, L., Aljoaid, S., Alharbi, T., Albraik, L., Alothman, W., Aljohani, H., Alzahrani, A., Alqahtani, S., Kalantan, R., Althomali, R., Alameen, M., & Mufti, A. (2024). AI-Driven Models for Diagnosing and Predicting Outcomes in Lung Cancer: A Systematic Review and Meta-Analysis. *Cancers*, *16*(3), 674. <https://doi.org/10.3390/cancers16030674>

[9] Li, Y., Wu, X., Yang, P., Jiang, G., & Luo, Y. (2022). Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis. *Genomics, proteomics & bioinformatics*, *20*(5), 850–866. <https://doi.org/10.1016/j.gpb.2022.11.003>

[10] R. Alizadehsani *et al*., "Explainable Artificial Intelligence for Drug Discovery and Development: A Comprehensive Survey," in *IEEE Access*, vol. 12, pp. 35796-35812, 2024, doi: 10.1109/ACCESS.2024.3373195

[11] Wang, J., Liu, G., Zhou, C., Cui, X., Wang, W., Wang, J., Huang, Y., Jiang, J., Wang, Z., Tang, Z., Zhang, A., & Cui, D. (2024). Application of artificial intelligence in cancer diagnosis and tumor nanomedicine. *Nanoscale*, *16*(30), 14213–14246. <https://doi.org/10.1039/d4nr01832j>

[12] Dolezal, J. M., Wolk, R., Hieromnimon, H. M., Howard, F. M., Srisuwananukorn, A., Karpeyev, D., Ramesh, S., Kochanny, S., Kwon, J. W., Agni, M., Simon, R. C., Desai, C., Kherallah, R., Nguyen, T. D., Schulte, J. J., Cole, K., Khramtsova, G., Garassino, M. C., Husain, A. N., Li, H., … Pearson, A. T. (2023). Deep learning generates synthetic cancer histology for explainability and education. *NPJ precision oncology*, *7*(1), 49. <https://doi.org/10.1038/s41698-023-00399-4>

[13] Lobato-Delgado, B., Priego-Torres, B., & Sanchez-Morillo, D. (2022). Combining Molecular, Imaging, and Clinical Data Analysis for Predicting Cancer Prognosis. *Cancers*, *14*(13), 3215. <https://doi.org/10.3390/cancers14133215>

[14] Prelaj, A., Galli, E. G., Miskovic, V., Pesenti, M., Viscardi, G., Pedica, B., Mazzeo, L., Bottiglieri, A., Provenzano, L., Spagnoletti, A., Marinacci, R., De Toma, A., Proto, C., Ferrara, R., Brambilla, M., Occhipinti, M., Manglaviti, S., Galli, G., Signorelli, D., Giani, C., … Pedrocchi, A. L. G. (2023). Real-world data to build explainable trustworthy artificial intelligence models for prediction of immunotherapy efficacy in NSCLC patients. *Frontiers in oncology*, *12*, 1078822. <https://doi.org/10.3389/fonc.2022.1078822>

[15] Lotter, W., Hassett, M. J., Schultz, N., Kehl, K. L., Van Allen, E. M., & Cerami, E. (2024). Artificial Intelligence in Oncology: Current Landscape, Challenges, and Future Directions. *Cancer discovery*, *14*(5), 711–726. <https://doi.org/10.1158/2159-8290.CD-23-1199>