TopoDX: A Novel Approach to Topological Network Analysis for the Early Diagnosis of Non-Small Cell Lung Cancer

Isabella Wu

Abstract

Lung cancer accounts for about 25% - the highest - of all cancer deaths. My research focuses on non-small cell lung cancer (NSCLC), lung cancer's primary histological form. NSCLC is often undetected until symptoms appear in the late stages, making it imperative to identify more specific and sensitive tumor-associated biomarkers for early diagnosis. Many studies fail to evaluate the effectiveness of their methods, achieving low performance (75-80%). In this study, I first identified significant differentially-expressed-genes (DEGs), or biomarkers, using functional enrichment analysis, novel complex network methods, and AUC to robustly validate the diagnostic performance of biomarkers. Unlike existing studies, I utilized a diverse range of network topological metrics to identify the most comprehensive set of biomarkers. Secondly, I introduced a novel and systematic method to identify the top topological metric in protein networks that is essential for biomarker selection. Compared to selection by conventional metrics, the diagnostic performance is improved 17%. In addition, I proposed a novel composite selection index (C-index) which concurrently considers complementary factors in biomarker selection, greatly increasing the diagnostic performance from 75% AUC to 91%. Finally, I explored a clinicogenomic machine learning model with top biomarkers selected and clinical covariates to enable more accurate diagnosis. The proposed methodology of finding top metrics is effective and general, and can be applied to various cancers. I expect my work to fundamentally advance topological network research, which is widely applied to identify biomarkers. I anticipate my overall research to contribute immensely to personalized medicine, improving the vital search for therapeutic targets.

1 Introduction

Lung cancer is the deadliest cancer worldwide and is highly fatal, with a five-year survival rate of only 18.6%. Non-small cell lung cancer (NSCLC) accounts for about 85% of all lung cancer cases. More than half of lung cancer patients die within one year of being diagnosed because NSCLC is often not detected until the late stages, with the onset of noticeable symptoms. Hence, there is an urgent need to identify more specific and sensitive biomarkers for early diagnosis. Biomarkers provide insights into the molecular origins and behaviors of NSCLC, enabling the development of targeted therapies and the identification of high-risk patients for personalized medicine.

The development of bioinformation technology and analytical strategies provide powerful tools for the analysis and identification of differentially expressed genes (DEGs) as biomarkers. In the search for biomarkers, topological network analysis is one of the most powerful and widespread tools in the field to analyze interactions between genes and identify essential biomarkers in the network. Despite the potential, the performance of existing work is not high. Many studies that utilize network analysis [1,2,3] fail to fully consider the biological significance of their quantitative network methods, and only use a limited number of metrics in the search of

biomarkers. Currently, degree of freedom is the most widely used conventional metric. My preliminary studies discover that other topological scoring methods could greatly outperform degree of freedom. In this work, the terms metric and method will be used interchangeably to describe the topological scoring methods.

The expansion of the availability and quantity of molecular biological data has created a need for improved computational methods for analysis. The aim of this work is to develop novel computational methods to empower the topological network analysis for the efficient and accurate finding of critical biomarkers. My work is divided into three stages.

In the first stage of the work, I perform a detailed analysis of lung cancer cases to identify the important biomarker signature for the diagnosis of NSCLC. The discovery of biomarkers is of critical importance and has attracted a lot of studies. However, to my best knowledge, no previous studies have reached my level of comprehensiveness in analysis. I compiled three datasets to increase the number of samples and ensure greater accuracy with the larger dataset. Functional enrichment analysis was used to find the most significant enriched biological pathways, and the relationship interactions between DEGs were analyzed by protein-protein interaction (PPI) networks in *Cytoscape.* AUC was used to robustly validate the diagnostic performance of the biomarkers. I further explored concurrent use of multiple biomarkers in NSCLC analysis, and developed a 4-gene signature consisting of *AGER*, *CA4*, *RASIP1*, and *CAV1* that were validated using survival and expression analysis in the Cancer Genome Atlas (TCGA) database. These identified DEGs, when utilized together, might be promising biomarkers for NSCLC, and can be considered as possible therapeutic targets for future targeted drug therapy.

In *the second stage* of this study, I introduce a novel and systematic method to identify the top performing topological scoring metric (or method) in protein networks. From the results, I propose a novel composite selection index (C-index) that concurrently considers complementary factors, enabling the selection of biomarkers with greatly increased diagnostic performance. To guide the search of composite index, I introduce two new performance evaluation methods based on AUC. The use of C-index is a breakthrough and is transformative in the usage of networks to identify biomarkers. It helps improve search algorithms and provides scoring functions that lead to vast improvement in the speed and accuracy of identifying disease correlated pathways and genes.

Finally, *in the third stage*, a clinicogenomic machine learning model using the potential 4-gene signature as well as clinical covariates is exploited to achieve high performance and accuracy in the diagnosis of NSCLC. For validation, the model was used to evaluate the metrics discovered in the second stage in a validation cohort.

The proposed methods are not restricted to the use in diagnosing NSCLC, but can be extended to diagnose other types of cancers. The proposed methodologies, once optimized, can fundamentally

advance the topological network research. With the widespread use of topological network analysis, the proposed techniques will drastically increase the accuracy and efficiency in the search for new biomarkers.

2 4-Gene Biomarker Signature Identification

To ensure a greater accuracy in biomarker identification with a larger dataset, I compiled three datasets to increase the number of samples. Both functional and network analysis were used to identify the biomarkers. Unlike previous studies that most commonly use a singular network metric, I took the time to test the the result of each network metric, and performed the analysis one by one. Each metric identified a set of genes, and the use of all metrics ensures the finding of the most comprehensive list of candidate biomarkers. The top genes were then chosen from this comprehensive list based on area under the receiver operating characteristic curve (AUC) score. AUC measures the ability of a classifier to distinguish between disease and control. The higher the AUC, the better the biomarker is able to detect disease. By using every network metric, I analyze the network from all possible aspects and ensure that no significant gene is missed. Utilizing AUC robustly validates that the biomarkers selected perform the best in the dataset.

In addition to the comprehensive identification methodology, I further explore the concurrent use of multiple biomarkers in NSCLC prediction by calculating the performance of their aggregate expression. Many previous studies have only considered the capability of individual biomarkers, and failed to evaluate their combined potential. Cancer can never be caused or predicted by a single mutation or gene. It is caused by multiple genes in a functional or signaling pathway working together in a cascade of mutations to promote tumorigenesis. Thus, I developed a 4-gene biomarker signature that, with the concurrent use of top biomarkers, which greatly increases the diagnostic performance. This first part of my research serves as a base for the following sections, where I explore in more depth the topological network methods (which performs the best in biological networks) to look for critical metrics for biomarker identification. Rather than using all metrics in the biomarker search, as done in the study of this section, the search with new metrics will be much more efficient, and can achieve higher performance at lower complexity.

2.1 Microarray data and detection of DEGs

My study analyzed 526 total NSCLC samples, 446 tumors and 80 control cases, by combining Gene Expression Omnibus (GEO) [4] datasets GSE31210, GSE33356, and GSE50081. The GEO database provides transcriptomic profiles of multiple cancers. By combining the three microarray datasets into one, a more comprehensive representation of the diseased population can be achieved. I used GEO2R analysis to preprocess the data based on tumor vs control groups. An initial pool of 267 statistically significant DEGs (p-value<0.01 and |logFC| > 1.2) were identified for further analysis, of which 93 are upregulated and 174 are downregulated as shown in Fig. 1. The raw gene expression values are normalized using z-score.

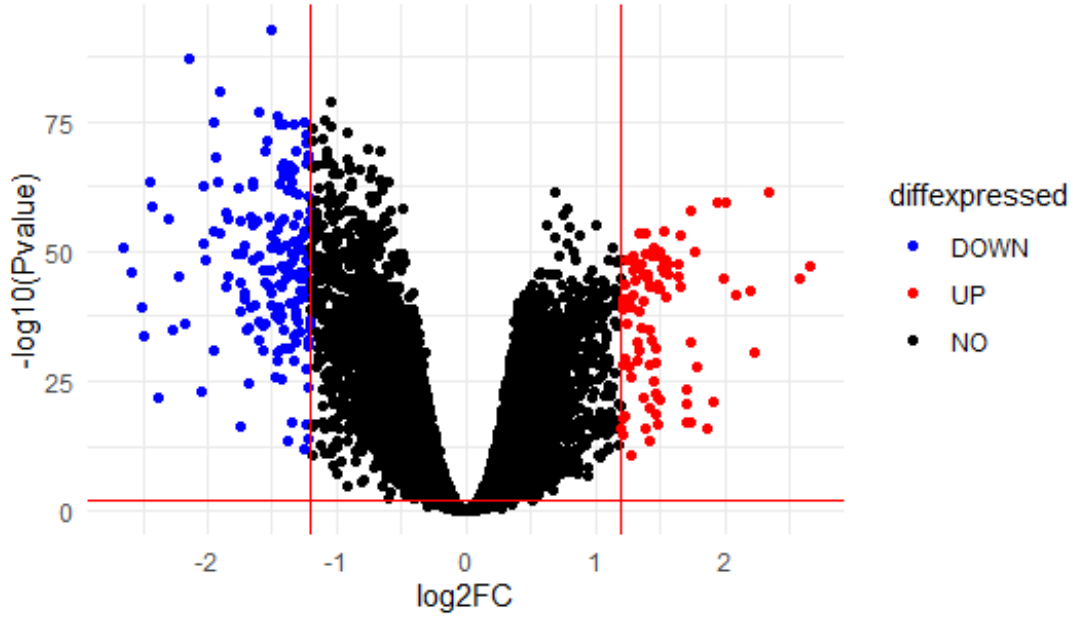


Figure 1: Distribution of differentially expressed genes

2.2 Functional Enrichment Analysis

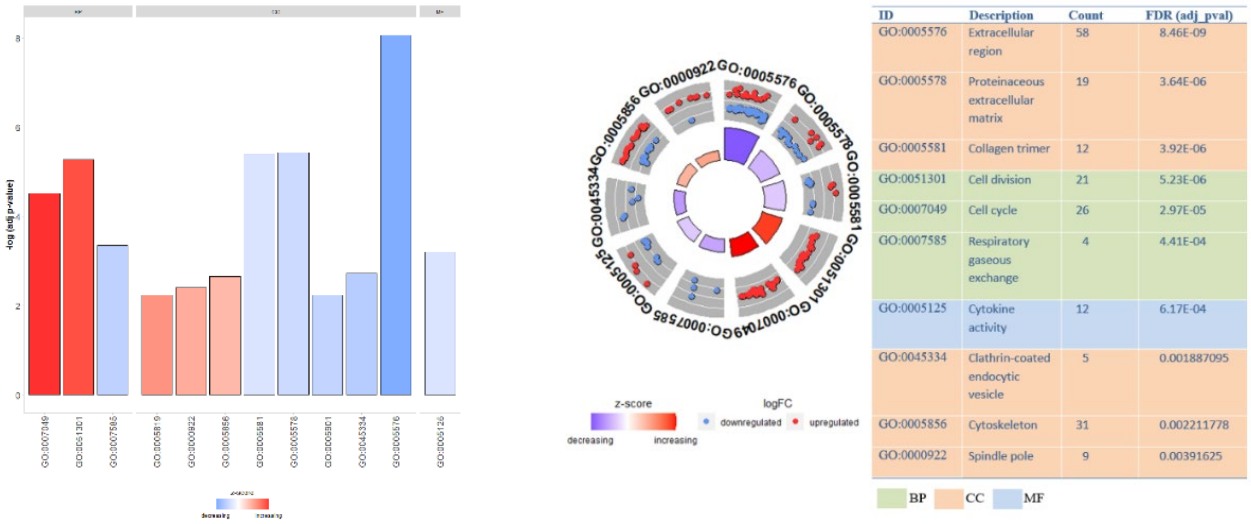
I performed functional enrichment analysis using the DAVID gene functional annotation tool [5]. The GO (Gene Ontology) database provides information about gene function through ontology and includes three main categories: biological process (BP), cellular component (CC), and molecular function (MF). In this study, I analyzed the most significant GO terms using DAVID to identify enriched terms with a threshold value of FDR (adjusted p-value) < 0.05. Statistically significant GO terms were also expressed as a z-score ($zscore = \frac{(upregulated\ genes - downregulated\ genes)}{\sqrt{count}}$), which signifies whether the GO term is more likely to be decreased (negative value) or increased (positive value).

The DEGs were significantly enriched in 16 GO terms (4 biological process, 11 cellular components, 1 molecular function). The GO term enrichment values were visualized in Fig. 2a. The 10 GO terms with the lowest FDR values are additionally presented in a circular visualization using the *GOplot* package (version 1.0.2) in R [6] (Fig. 2b). The lower the FDR value, the more significant the enrichment.

As shown in Figure 2b, GO terms corresponding with upregulated genes lead to increased cell division, including cell division (GO:0050301), cell cycle (GO:0007049), cytoskeleton (GO:0005856), and spindle pole (GO:0000922). Increased cell division promotes tumorigenesis and the development of cancer. GO terms corresponding with downregulated genes play a role in anti-tumor defense, such as decreased cytokine activity (GO:0005125), which functions to inhibit cancer development and progression in the immune system. Downregulated genes in clathrin-coated endocytic vesicles (GO:0045334) may signify a disrupted neurotransmission and signal

transduction pathway that can lead to cancer. The extracellular matrix (GO:0005576 and GO:0005578) is a major component of the tumor microenvironment and plays critical roles in cancer development and progression. Altering the balance of ECM signaling can provide biochemical and biophysical cues that promote cancer cell proliferation, migration and invasion, as well as stiffness of tissues.

String-db [7] analysis reveals one significant KEGG biological pathway, ECM-receptor interaction (hsa04512, FDR= 0.0152). Complications in the ECM-receptor interaction pathway can result in induced cancer progression and development, as ECM-receptors play important roles in tumor shedding, adhesion, and degradation.



ID	Description	Count	FDR (adj_pval)
GO:0005576	Extracellular region	58	8.46E-09
GO:0005578	Proteinaceous extracellular matrix	19	3.64E-06
GO:0005581	Collagen trimer	12	3.92E-06
GO:0051301	Cell division	21	5.23E-06
GO:0007049	Cell cycle	26	2.97E-05
GO:0007585	Respiratory gaseous exchange	4	4.41E-04
GO:0005125	Cytokine activity	12	6.17E-04
GO:0045334	Clathrin-coated endocytic vesicle	5	0.001887095
GO:0005856	Cytoskeleton	31	0.002211778
GO:0000922	Spindle pole	9	0.00391625

Figure 2a and 2b: GO term enrichment analysis of DEGs (FDR < 0.05)

2.3 Network analysis

Disease susceptibility and other disease correlated factors are not the result of single gene mutations in isolation, but are due to the perturbation of a gene network and its interactions. I explore the use of network analysis and other methodology to understand the topology of molecular interactions, and identify the most essential top scoring biomarkers in the network.

To investigate interactions between DEGs and the relationships between genes and diseases, I constructed a protein-protein interaction (PPI) network based on the DEGs using the STRING Interactome database, then inputted the network into *Cytoscape* [8] for further analysis with the *CytoHubba* plugin. The network is shown in Figure 3. Improving on previous studies, I developed a completely comprehensive list of candidate biomarkers by using all twelve *CytoHubba* topological scoring methods to ensure that no essential genes are missed. Each method's top ten scoring nodes were identified based on disease score, and the top 10 were

chosen from this combined list of top-scoring nodes based on AUC value.

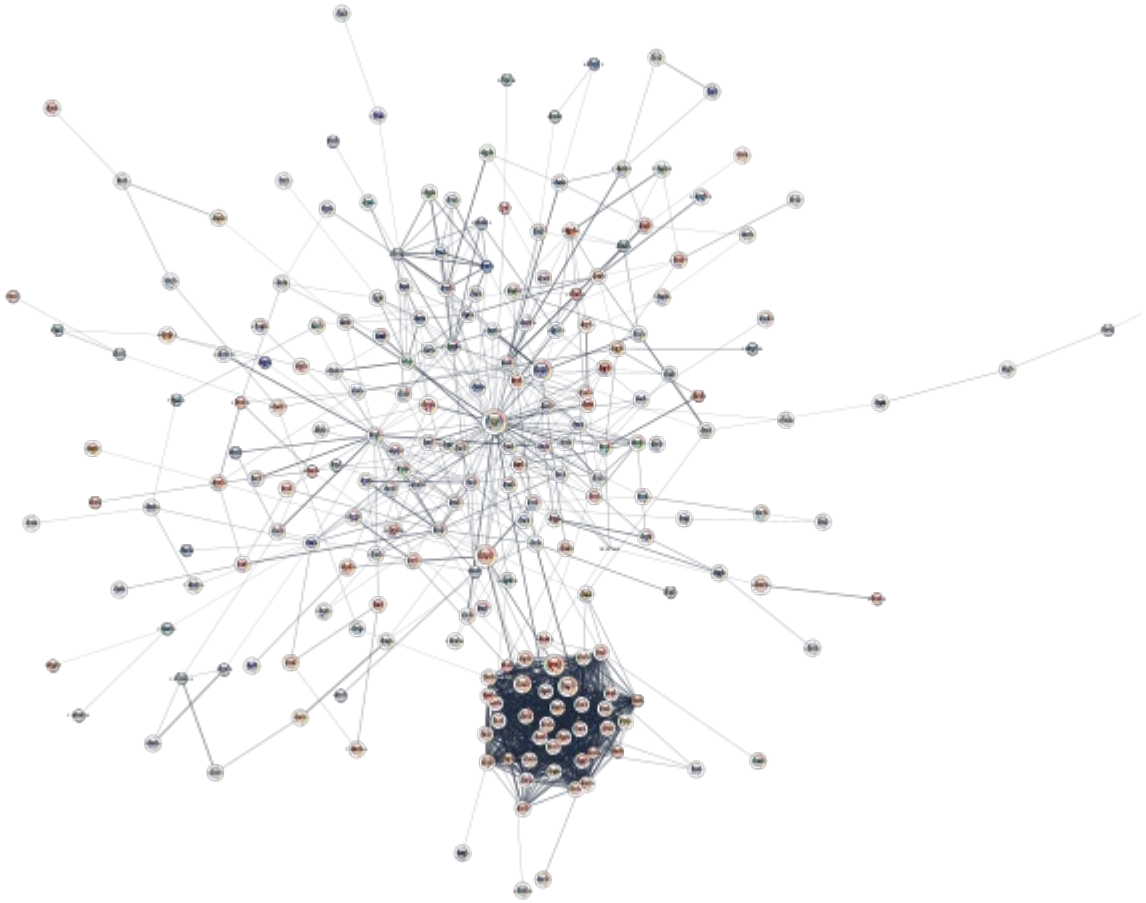


Figure 3: Protein-protein interaction (PPI) network of DEGs in *Cytoscape* (red = downregulated DEG, blue = upregulated DEG)

Depending on the range of nodes involved, the topological scoring metrics in *Cytoscape* are divided into two categories, local and global. The local metrics include Degree, Maximal Clique Centrality (MCC), Density of Maximum Neighborhood Component (DMNC), Maximum Neighborhood Component (MNC), and Clustering Coefficient. The global metrics include Betweenness, Bottleneck, Eccentricity, Closeness, Radiality, Stress, and Edge Percolated Component. Several of the metrics contain overlapping top-ranked nodes, and many have nodes with the same score. This resulted in more than 10 nodes being included in the top ranked nodes for some of the metrics. Without counting the overlapped genes, we obtained 82 candidate biomarkers overall. The ranking methods and their top-ranking nodes are shown in Table 1.

Table 1: Top disease-score ranked nodes found through 12 topological scoring methods

Scoring method	Top-ranked genes (listed in order of disease scores)
Degree	IL6, ASPM, TTK, KIF20A, TPX2, CENPF, UBE2C, CCNB1, CDK1, NUF2, KIF11, KIF4A, DLGAP5, TOP2A, MAD2L1, PBK, BIRC5
MCC	ASPM, BIRC5, BUB1B, CNNB1, CCNB2, CDC20, CDK1, CENPF, DLGAP5, KIAA0101, KIF11, KIF20A, KIF4A, MAD2L1, MELK, NDC80, NUF2, NUSAP1, PBK, RRM2, TOP2A, TPX2, TTK, UBE2C, ZWINT
DMNC	HMMR, CDKN3, CEP55, CENPF, BUB1B, UBE2C, CCNB2, RRM2, NUSAP, NDC80, KIAA0101, TOP2A, ZWINT, CDC20, MAD2L1, PBK, MELK
MNC	IL6, KIF20A, DLGAP5, ASPM, KIF4A, KIF11, NUF2, CDK1, CCNB1, TPX2, BIRC5, TTK
Clustering Coefficient	ABCA3, ACADL, BCHE, CA4, CDCA7, DKK2, FAM107A, FHL1, GINS1, GPX3, HS6ST2, LIFR, PTPRB, RASIP1, SCARA5, SCN7A, SDPR, SFTA3, SOX7, TMPRSS4, WIF1
Betweenness	IL6, BMP2, UBE2C, CAV1, SOX2, SPP1, AQP4, CDH5, NPNT, AGER
Bottleneck	IL6, UBE2C, SOX2, CAV1, BMP2, SPP1, AGER, CDH5, AQP4, NPNT, SFTPD
EcCentricity	KIF26B, ITGA8, HHIP, NPNT, AGER, THBS2, SFTPC, KIF4A,, HMGB3, SCGB1A1, KIF20A, KIF11, SFTPD, IL6, CDH5, CLIC5, GPC3, COL6A6, AGTR1, SFTPB
Closeness	IL6, UBE2C, SOX2, CAV1, SPP1, BMP2, CDH5, DLGAP5, CDK1, CCNB1, BIRC5
Radiality	IL6, SOX2, CAV1, BMP2, SPP1, CDH5, UBE2C, CLDN5, CEACAM5, TEK
Stress	IL6, SOX2, BMP2, CAV1, SPP1, UBE2C, CDH5, SCGB1A1, AQP4, GPC3
EPC	Il6, TOP2A, CDC20, CDK1, UBE2T, ZWINT, CDKN3, NUF2, KIF4A, RRM2, KIF11, NEK2, ASPM, NUSAP1, KIF20A, HMMR, ANLN, DLGAP5, NDC80

To determine whether these top-ranked genes could be potential biomarkers for NSCLC, I calculated the area under the receiver operating characteristic curve (AUC) for each gene using the pROC package in R Studio [9]. The AUC measures the sensitivity and specificity of the biomarkers, evaluating their ability to distinguish between disease and control. The top-ranked ten genes based on AUC were *AGER* (AUC = 0.8982), *CA4* (AUC=0.8954), *RASIP1* (AUC=0.8900), *CAV1* (AUC=0.8744), *CDH5* (AUC=0.8716), *FAM107A* (AUC=0.8711), *KIF26B* (AUC=0.8660), *CLDN5* (AUC=0.8633), *CLIC5* (AUC=0.8633), and *SPP1* (AUC=0.8572). The top 20 genes and their AUC values are shown in Table 2.

Table 2: Top-ranked 20 genes detected in network analysis based on AUC score

Gene name	AUC	logFC	FDR (adj_pval)
AGER	0.8982	-2.13691	4.45E-84
CA4	0.8954	-1.50475	1.70E-89
RASIP1	0.8900	-1.20962	3.52E-65

CAV1	0.8744	-1.41913	5.23E-55
CDH5	0.8716	-1.40895	3.69E-63
FAM107A	0.8711	-1.94635	1.39E-72
KIF26B	0.8660	1.293536	2.98E-48
CLDN5	0.8636	-1.21244	1.24E-56
CLIC5	0.8633	-2.03743	4.42E-61
SPP1	0.8572	2.334613	6.97E-60
SDPR	0.8560	-1.861487998	2.53E-56
PTRB	0.8554	-1.653827903	1.57E-61
CDCA7	0.8464	1.650737609	6.42E-52
AGTR1	0.8432	-1.503517517	5.72E-52
HS6ST2	0.8430	1.995895019	3.59E-44
SOX7	0.8412	-1.599491119	2.45E-48
TMPRSS4	0.8380	1.634397285	1.70E-44
ACADL	0.8378	-1.517576379	1.34E-55
TOP2A	0.8366	2.014367341	4.75E-58
FHL1	0.8359	-1.622980044	1.19E-54

2.4 Disease prediction with multiple biomarkers simultaneously

Rather than just considering the capability of individual biomarkers in NSCLC identification, I further explore the concurrent use of multiple biomarkers. To my understanding, no previous study has done in this way, and many studies fail to consider the complementary use of multiple biomarkers at once. By using multiple biomarkers that work together, my aim is to reduce anomalies and increase overall performance. In order to achieve the goal, I need a way to evaluate their joint performance. To do so, I propose a method that consists of averaging the gene expression values of multiple biomarkers, and then finding the AUC corresponding to this

aggregate expression. This will help reduce the impact of outliers and improve AUC.

Using too many biomarkers, however, can result in decreased performance if the variance is too large. Thus, I first set out to find the ideal number of biomarkers to use concurrently. I aggregated the top biomarker expression values one by one and determined when the AUC reaches its peak. I found that the ideal case is to use the top 4 biomarkers, *AGER*, *CA4*, *RASIP1*, and *CAV1*, together. These four biomarkers make up my proposed **4-gene biomarker signature** for the prediction of NSCLC. The resulting AUC, 0.9238, is higher than any AUC values achieved by using single biomarkers and the AUC achieved by using all ten at once. This is significant, as it confirms my hypothesis that using multiple biomarkers simultaneously increases prediction performance. This concept can be expanded on and further explored in future studies, and will lead to great improvements in prediction accuracy. The AUC comparisons are shown in Figure 4 and Table 3. The ROC curves are visualized in Figure 5.

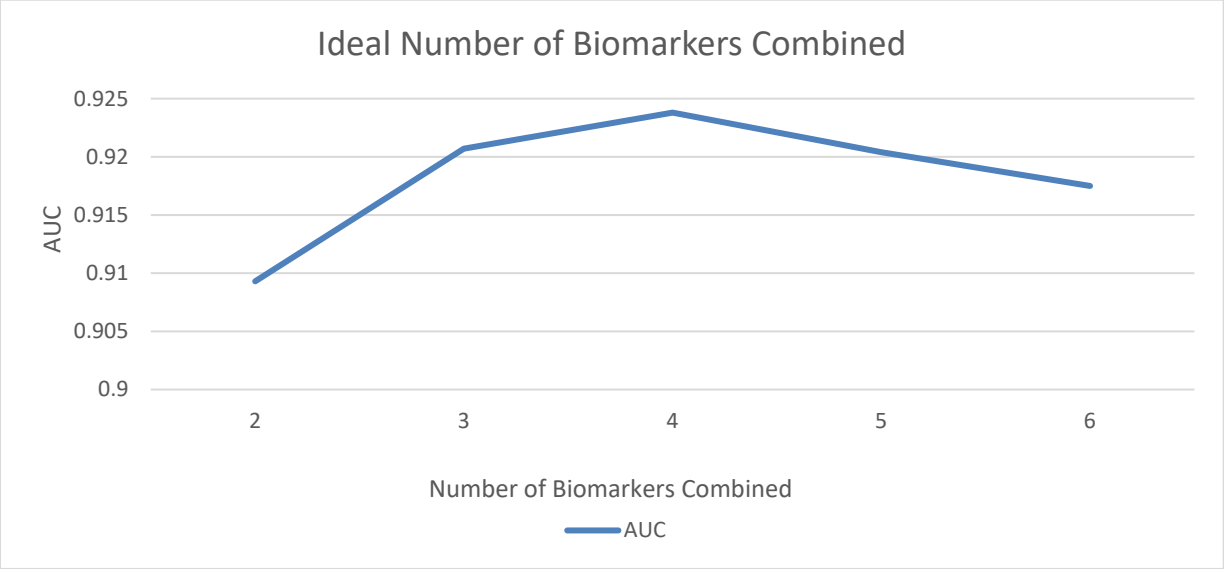


Figure 4: Integrated AUC of combined biomarkers

Table 3: Comparison of integrated AUC between biomarker signatures

Biomarkers	**Integrated AUC**
AGER+CA4	0.9093
AGER+CA4+RASIP1	0.9207
AGER+CA4+RASIP1+CAV1	0.9238

AGER+CA4+RASIP1+CAV1+CDH5+FAM107A	0.9175
All top 10	0.8952

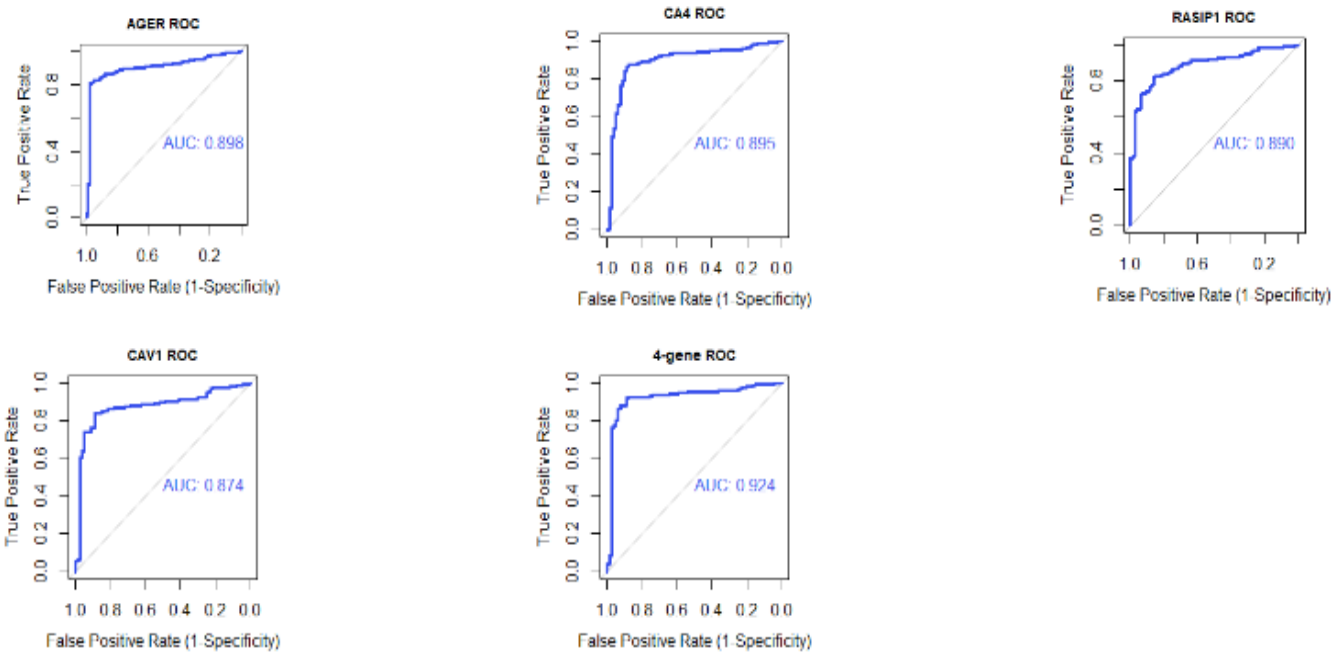


Figure 5: ROC curves of AGER, CA4, RASIP, CAV1, and the 4-gene signature

3 Finding of Critical Network Topological Metrics

Topological networks play a critical role in the field of bioinformatics. Although it is one of the most wide-spread tools used for biomarker identification, many studies haven't explored the full range of the applications of the complex networks. The most conventional, wide-spread network scoring method utilized in biomarker selection is degree of freedom, but as proven in this study, this metric isn't the best performing in biological networks. Thus, in this section, I introduce a systematic method to identify the top topological metric in protein networks.

I performed a systematic study of twelve topological methods in *CytoHubba* to find the metrics that provide the best precision and accuracy. Not only did I find the best individual metrics, I also explored their performances when used concurrently with complementary metrics. In the first stage of study, I obtained a complete and comprehensive list of candidate biomarkers by utilizing all the metrics. However, using so many metrics in practical use will leads to high complexity and inefficiency. Instead, I searched for a combination of the subset of metrics with the best performance, which achieves the comprehensiveness of using all the metrics while vastly

decreasing the complexity and inefficiency. Few previous studies have utilized multiple network metrics concurrently when finding hub genes. Using multiple metrics simultaneously leads to more robust results, as the analytical coverage of a biological system is increased by utilizing several complementary methods that reveal different aspects of the network. To evaluate the performance of the topological metrics, I introduce two new performance evaluation methods based on AUC. The dataset was divided into 80% for identifying the top metrics, and 20% for validation of the top metrics.

3.1 Method performance evaluation with two types of AUC

The local methods I studied include Degree, Maximal Clique Centrality (MCC), Density of Maximum Neighborhood Component (DMNC), Maximum Neighborhood Component (MNC), and Clustering Coefficient. The global methods include Betweenness, Bottleneck, Eccentricity, Closeness, Radiality, Stress, and Edge Percolated Component (EPC). The methods are described in detail in *Cytohubba*'s official article by the creators [10].

To evaluate the performance of the twelve *CytoHubba* metrics, I introduce two new methods to calculate the AUC of multiple genes at once. For the first, the mean AUC of multiple metrics is calculated by taking the average of the AUCs of individual genes. For the second, I first averaged the normalized gene expression values of the individual genes, then found the integrated AUC of the aggregate gene expression. This integrated AUC allows the evaluation of the performance of the group of genes as a whole, rather than analyze each gene individually. In the second type of AUC, a group of genes are used together for cancer prediction.

For comparison purposes, the AUC of the top ten genes overall found in the previous section (*AGER*, *CA4*, *RASIP1*, *CAV1*, *CDH5*, *FAM107A*, *KIF26B*, *CLDN5*, *CLIC5*, *SPP1*) was calculated using both performance evaluation methods. The mean of individual AUCs of genes was 0.87508, and the integrated AUC obtained using the aggregate expressions of genes was 0.8952. As expected, the performance of NSCLC analysis improves with the simultaneous use of multiple genes, as I discovered in the previous section.

In the previous section, I used each topological scoring method to find a set of top biomarkers. Now, to evaluate performance, I calculated the mean individual AUC and the integrated AUC of the biomarkers identified with each metric, and used these AUC scores to rank the methods based on performance. As shown in Table 3, in general, using the integrated AUC has higher performance scores than the use of the mean of individual AUCs. This shows that using multiple biomarkers at once to predict cancer can improve the accuracy of diagnosis. The mean of individual AUCs is also more susceptible to outliers, as each biomarker's performance score is directly averaged into this AUC. The top local and global scoring methods using the mean individual AUC value were Clustering Coefficient (AUC=0.8169) and Betweenness (AUC=0.8065) respectively. The top local and global metrics using the integrated AUC were

Clustering Coefficient (AUC=0.8792) and Bottleneck (AUC=0.8700) respectively. The top metric overall was Clustering Coefficient. The performances of all 12 methods are displayed in Table 4.

Compared to Degree, Clustering Coefficient's diagnostic performance is increased 17%. Most of the focus in network studies has been to find highly connected genes, or hubs, using degree. However, this research explores and confirms that other topological scoring methods are better suited for finding relationships between topological properties and functional features of protein networks. This breakthrough is transformative in the usage of networks to identify biomarkers.

Table 4: Scoring metrics ranked by AUC score

Scoring metric	Integrated AUC	Mean Individual AUC
Clustering Coefficient	0.8792	0.8169
Bottleneck	0.87	0.7999
Betweenness	0.8610	0.80654
EcCentricity	0.8442	0.7939
Stress	0.8311	0.7862
DMNC	0.8053	0.7851
MCC	0.8035	0.7845
EPC	0.7948	0.7867
Degree	0.7796	0.7788

MNC	0.7685	0.7803
Radiality	0.7121	0.76747
Stress	0.624	0.7917

3.2 Biomarker Selection with C-Metric

Candidate biomarkers can be searched comprehensively using all twelve topological scoring methods, as done in the first biomarker selection section. Although this increases the true positive rate of NSCLC detection, the search will incur a high computational cost and result in high complexity. On the other hand, using a single scoring method doesn't guarantee comprehensive scoring of the heterogeneous biological network. Thus, finding the ideal number of complementary scoring methods that can adequately analyze network relations while reducing the complexity is a key. In order to gain full understanding of the structure of a network, using local and global metrics concurrently is ideal as it allows us to measure the node-level properties as well as the network-level properties.

In the first biomarker selection stage, I found the top 10 and top 20 ranked nodes overall in the dataset based on AUC. To evaluate the performance of combined methods, I utilized both AUC and precision in predicting the top overall 10 and top 20 ranked nodes. This precision is calculated as $\frac{number\ of\ correctly\ identified\ essential\ genes}{total\ number\ of\ top\ ranked\ genes}$, where the total number of top ranked genes is 10 and 20. To find number of correctly identified essential genes, I took the top 10 and 20 genes from the superset of the genes from each metric, and calculated how many of them matched with the overall top 10 in the dataset.

First, I combined the top local and global metrics, Clustering Coefficient and Bottleneck. 7 out of the top 10 overall genes were correctly identified, resulting in a 70% precision. The mean individual AUC was 0.8716, and the integrated AUC was 0.8918 for the top 10 nodes. Likewise, the top 20 nodes had 75% precision, mean individual AUC 0.8500, and integrated AUC 0.8748. Combining Clustering Coefficient and Betweenness yielded the same group of genes and the same performance scores, as Bottleneck and Betweenness have nearly identical definitions and are interchangeable in this case. The combined metrics and their AUC scores are shown in Table 4. The results indicate that using multiple metrics helps improve the precision of predicting the top ranked biomarkers, proving the significance of my study.

I then combined the top three metrics Clustering Coefficient, Bottleneck, and Eccentricity. The top 10 genes had 90% precision, mean individual AUC 0.8743, and integrated AUC 0.897. The top 20

genes had 90% precision, mean individual AUC 0.8565, and integrated AUC 0.8638. The precision and AUC score increased with three metrices, as shown in Table 5.

To test four metrics, I combined the top four, Clustering Coefficient, Bottleneck, and Eccentricity, and Stress, based on integrated AUC, as well as Clustering Coefficient, Bottleneck, Eccentricity, and Closeness based on mean individual AUC. I found that both resulted in the same combined gene lists as the three metrices Clustering Coefficient, Bottleneck, and Eccentricity, with no improvement in AUC or precision. Thus, I confirm that combining Clustering Coefficient, Bottleneck, and Eccentricity results in the ideal combination with the highest performance score and efficiency. This becomes my **C-index**.

As discussed earlier, previous studies most commonly use degree and occasionally betweenness to find hub genes. To investigate whether the combination of these two results in high accuracy, I calculated the performance of the two metrices combined. The top 10 genes had 50% precision, 0.8344 mean individual AUC, and 0.6077 integrated AUC. The top 20 genes had 25% precision, 0.8063 mean individual AUC, and 0.7041 integrated AUC. Rather than simply taking widely-used metrics, these results indicate that the metrics found are much better suited for identifying biomarkers that can more effectively predict NSCLC.

Table 5: Performance of combined network scoring metrics

Combined Metrices	**Combined AUC (Top 10/Top20)**	**Mean Individual AUC (Top 10/Top20)**	**Precision of Top Genes (Top 10/Top 20)**	**Gene Union List (Top 10/Top 20)**
Clustering Coefficient + Bottleneck	0.8918	0.87157	0.70 (70%)	AGER, CA4, RASIP1, CAV1, CDH5, FAM107A, SPP1, SDPR, PTPRB, CDCA7
	0.8748	0.849875	0.75 (75%)	AGER, CA4, RASIP1, CAV1, CDH5, FAM107A, SPP1, SDPR, PTPRB, CDCA7, HS6ST2, SOX7, TMPRSS4, ACADL, FHL1, GPX3, GINS1, BCHE, NPNT, LIFR
Clustering Coefficient +	0.8918	0.87157	0.70 (70%)	AGER, CA4, RASIP1, CAV1, CDH5, FAM107A, SPP1, SDPR, PTPRB, CDCA7

Betweenness	0.8748	0.849875	0.75 (75%)	AGER, CA4, RASIP1, CAV1, CDH5, FAM107A, SPP1, SDPR, PTPRB, CDCA7, HS6ST2, SOX7, TMPRSS4, ACADL, FHL1, GPX3, GINS1, BCHE, NPNT, LIFR
Clustering Coefficient + Bottleneck + Eccentricity	0.897	0.87432	0.90 (90%)	AGER, CA4, RASIP1, CAV1, CDH5, FAM107A, KIF26B, CLIC5, SPP1, SDPR
	0.8638	0.856545	0.90 (90%)	AGER, CA4, RASIP1, CAV1, CDH5, FAM107A, KIF26B, CLIC5, SPP1, SDPR, PTPRB, CDCA7, AGTR1, HS6ST2, SOX7, TMPRSS4, ACADL, FHL1, GPX3, GINS1
Clustering Coefficient + Bottleneck + Eccentricity + Stress	0.897	0.87432	0.90 (90%)	AGER, CA4, RASIP1, CAV1, CDH5, FAM107A, KIF26B, CLIC5, SPP1, SDPR
	0.8638	0.856545	0.90 (90%)	AGER, CA4, RASIP1, CAV1, CDH5, FAM107A, KIF26B, CLIC5, SPP1, SDPR, PTPRB, CDCA7, AGTR1, HS6ST2, SOX7, TMPRSS4, ACADL, FHL1, GPX3, GINS1
Clustering Coefficient + Bottleneck + Eccentricity + Closeness	0.897	0.87432	0.90 (90%)	AGER, CA4, RASIP1, CAV1, CDH5, FAM107A, KIF26B, CLIC5, SPP1, SDPR
	0.8638	0.856545	0.90 (90%)	AGER, CA4, RASIP1, CAV1, CDH5, FAM107A, KIF26B, CLIC5, SPP1, SDPR, PTPRB, CDCA7, AGTR1, HS6ST2, SOX7, TMPRSS4, ACADL, FHL1, GPX3, GINS1

Degree + Betweenness	0.6077	0.8344	0.50 (50%)	AGER, CAV1, CDH5, SPP1, TOP2A, NPNT, CCNB1, CENPF, KIF20A, DLGAP5
	0.7041	0.806325	0.25 (25%)	AGER, CAV1, CDH5, SPP1, TOP2A, NPNT, CCNB1, CENPF, KIF20A, DLGAP5, UBE2C, KIF11, AQP4, BIRC5, TTK, TPX2, NUF2, ASPM, BMP2, KIF4A

4 Clinicogenomic Model for Diagnosis of NSCLC

In the first sections of this research, I focused on the performance of biomarkers and their ability to diagnose NSCLC, and how to improve the search for them. However, compared to only using genetic information, the usage of a variety of other factors will help improve the accuracy of diagnosis. The most important clinical attributes that impact the development of NSCLC are age, gender, and smoking status.

In order to incorporate both clinical and genomic attributes, I explored the use of machine learning techniques to create a clinicogenomic NSCLC diagnosis model with high accuracy that incorporates the 4-gene signature (*AGER, CA4, RASIP1, CAV1)* as well as age, gender, and smoking status. A Random Forest (RF) algorithm is used due to its important advantages over other classification models in terms of robustness to overfitting, ability to handle non-linear data, and stability in the presence of outliers.

The Python programming language with libraries such as numpy, pandas, and sklearn were used to construct the RF model. To preprocess the data, gene expression values of the four genes *AGER, CA4, RASIP1,* and *CAV1* were normalized using z-score normalization, ensuring the removal of biases. Ordinal Encoding and One-Hot Encoding were used to convert categorical variables such as smoking status and gender to quantitative data to be used by the algorithm, and then feature scaling was applied. The data was further prepared by splitting the dataset into split into features (the expression of the genes and the clinical attributes) and output (whether the subject has NSCLC or not). The dataset was then randomly split into 80% training set, and 20% validation set.

In the construction of the RF model, the ideal number of decision trees (n_estimators) was found to be 100. The model was fitted to the features of the training set data, and was then used to predict whether a patient had NSCLC or not in the validation dataset. To evaluate the model, confusion matrices, accuracy, and AUC were found.

5 Results

5.1 Validation of Biomarkers

To validate the 4-gene signature and the effects of the genes on NSCLC development, I performed overall survival analysis to validate the top ten genes using Kaplan-Meier survival plots (Figure 6) to examine the correlation of these genes and patient survival. The overall survival (OS) plots were based on 1925 lung cancer patients from GEO and TGCA (The Cancer Genome Atlas database) datasets. The hazard ratio (HR) with 95% confidence intervals and log rank P-value are calculated and indicated on graphs, I used a threshold of p-value < 0.01 is used to determine significance.

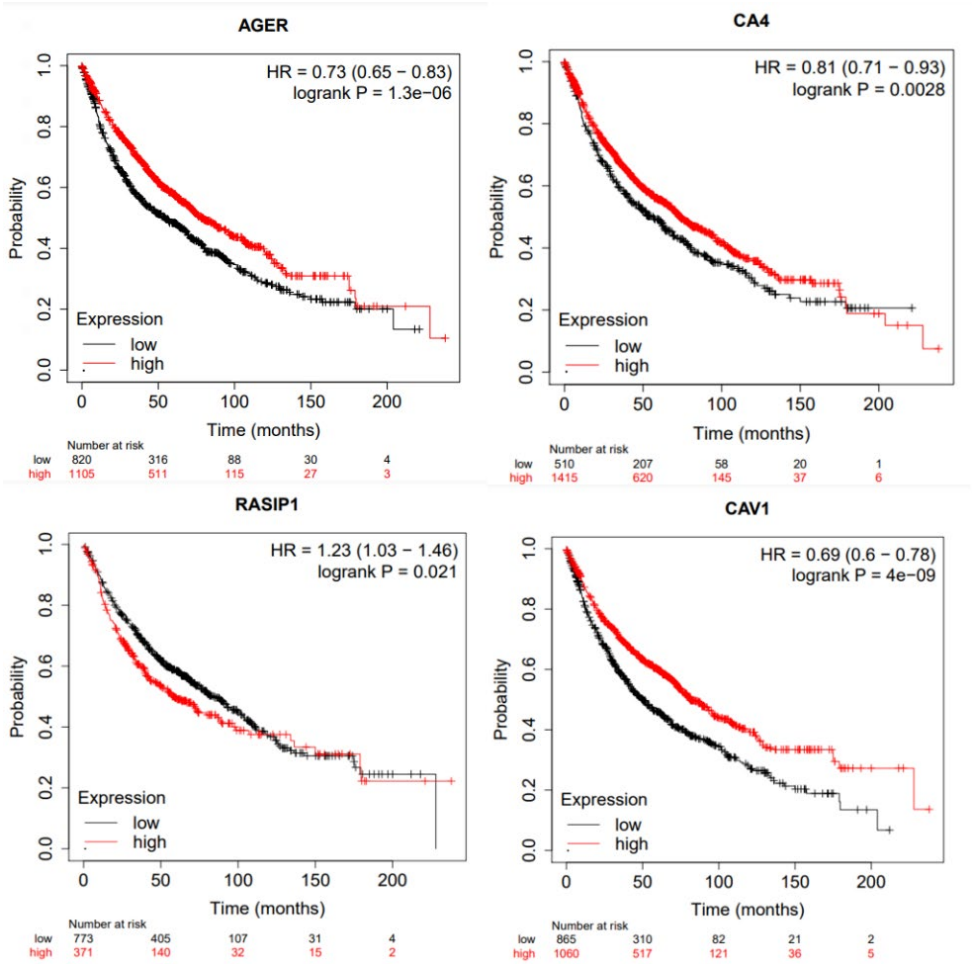


Figure 6: Kaplain-Meier survival plots for AGER, CA4, RASIP1, and CAV1 (red = high expression of gene, black = low expression of gene)

To confirm that the results are applicable outside of my data set, the GEPIA interactive website [11] allows for customizable functions based on TGCA data. For verification of the identified genes, I used GEPIA to plot a gene expression plot (Figure 7) to evaluate gene expression between lung squamous cell carcinoma (LUSC), lung adenocarcinoma (LUAD), and normal lung tissues.

LUSC and LUAD are the two histological types of NSCLC. I find that the expression level of the four genes is significantly different in NSCLC tumor tissues compared to normal lung tissues, indicating that the four-gene signature can be expanded and used in other data sets.

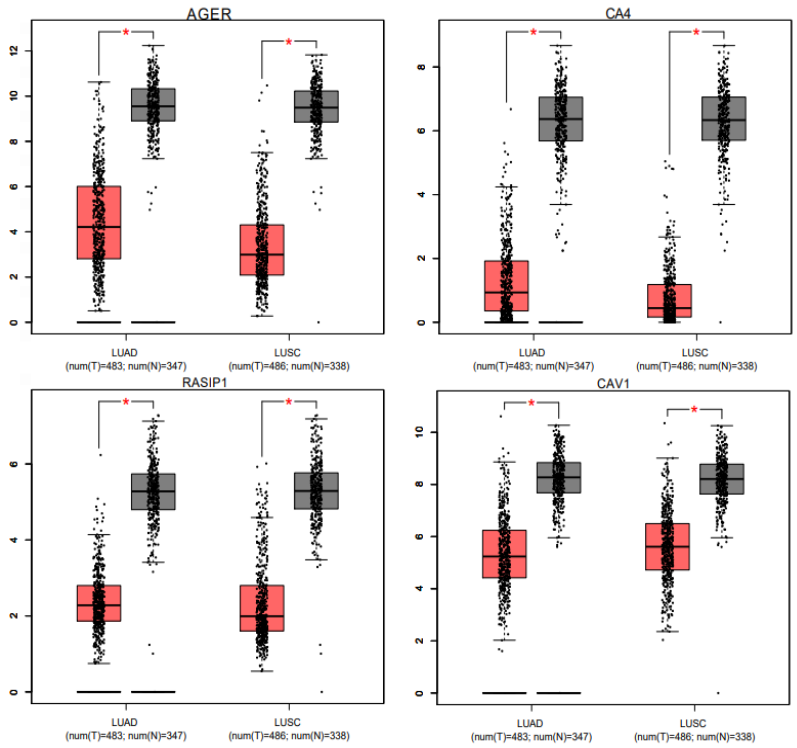


Figure 7: Comparison of gene expression between NSCLC and control patients (red = tumor, gray = control)

As shown in Fig. 6 and 7, low expression of AGER (log-rank P 1.E–06), CA4 (log-rank P 0.0028), RASIP1 (log-rank P 0.021), and CAV1 (log-rank P 4E-9) are associated with poor overall survival, indicating their significant impact on NSCLC prognosis. All four genes play a significant role in ECM receptor and signaling pathways, preventing the spread of tumorigenesis by decreasing proliferation and increasing apoptosis of cells. This matches with the significant GO and KEGG pathways identified in the first part of this work. Underregulating the expression of these genes leads to defected protection against the spread of cancer.

5.2 Validation of Metrics

To validate the metrics found in section 3, I calculated the integrated AUC of the metrics with the 20% validation set. The performance of Degree and C-index (Clustering Coefficient+Bottleneck+Eccentricity), as well as Clustering Coefficient, Eccentricity, and Bottleneck on their own, is shown in Table 6. As shown, the C-index performed better than Degree and each of its individual components. Compared to Degree, the performance of C-index increased by 20%, a significant improvement. Additionally, Clustering Coefficient alone improved on Degree by 17%. This validates my findings, demonstrating that my metrics vastly improve on the

conventionally used one.

Table 6: AUC performance of scoring metrics in validation set

Scoring metric	Integrated AUC
Degree	0.7596
Clustering Coefficient	0.8860
Bottleneck	0.8675
Eccentricity	0.8357
C-Index	0.9106

5.3 Random Forest Confusion Matrices and AUC

The accuracy and confusion matrix of the clinicogenomic RF model are shown in Fig. 8b. The accuracy and performance of a purely clinical model (age, gender, and smoking status) without the 4-gene signature are shown in Figure 8a. Clearly, the addition of the 4-gene signature had a great impact on the diagnostic capability of the model, improving accuracy by 13%.

The clinogenomic model achieved a high overall accuracy that was greater than 0.95, demonstrating the proficient performance of the classification model. As shown by the confusion matrix, the sensitivity (True Positive Rate) of the model was 0.96, while the specificity (True Negative Rate) was 0.8. Thus, the model seems better at predicting true positives (samples with NSCLC) than true negatives (control samples). One possible explanation is that the data is imbalanced, as there are ≈ 90 samples in the test set that have NSCLC, but ≈ 20 samples without NSCLC. The low false positive and false negative rates show that the model accurately avoids misdiagnosis. The relatively high sensitivity and specificity of the model shows that the RF model can accurately identify and predict NSCLC.

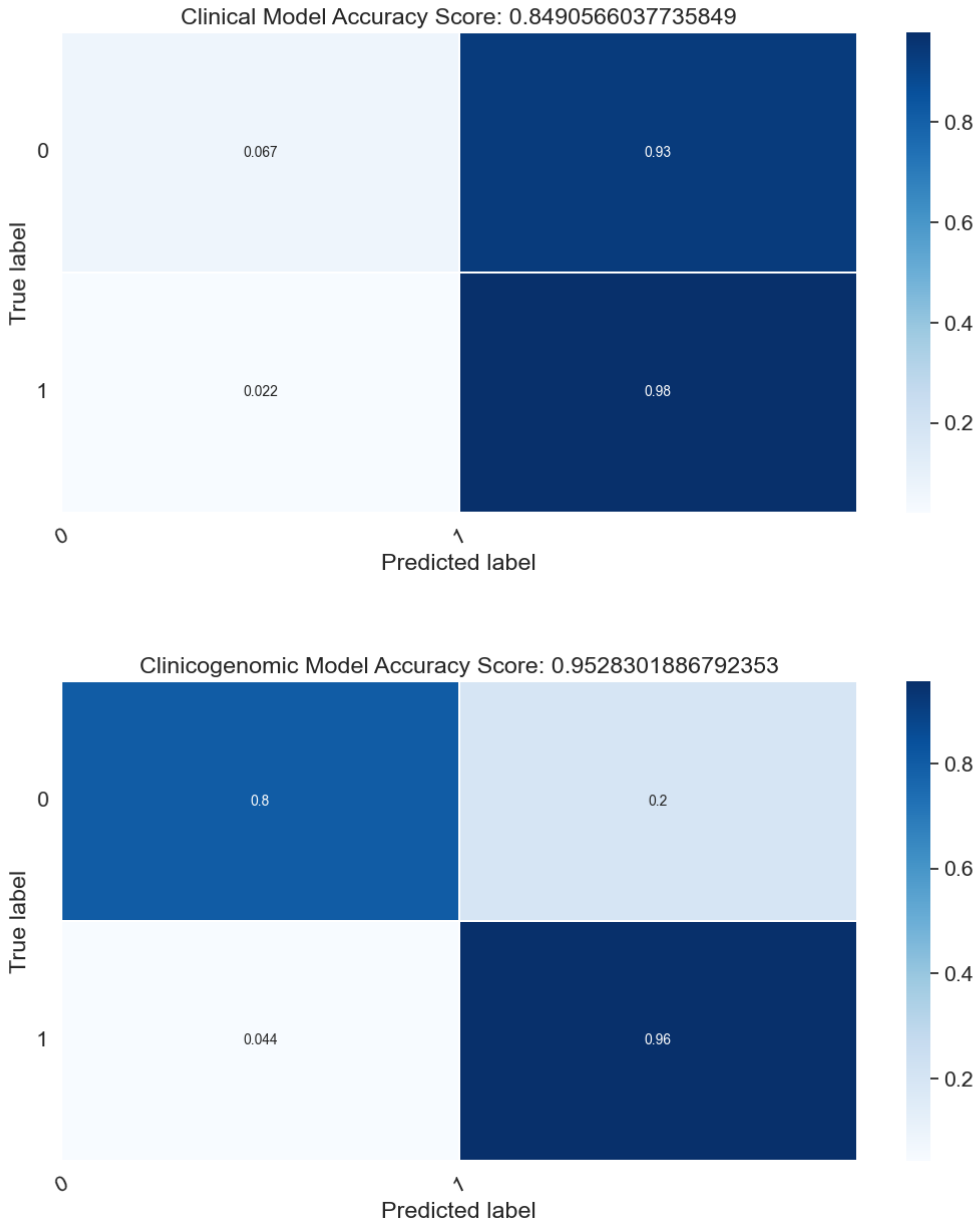
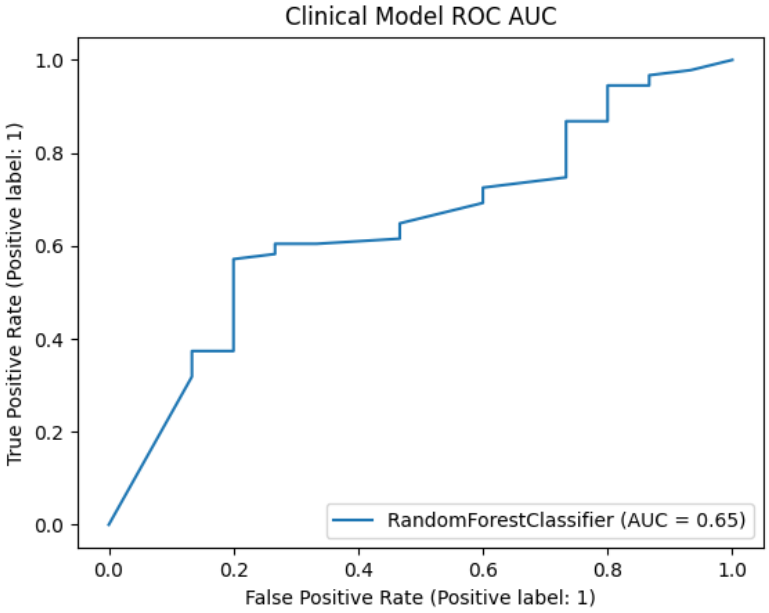


Figure 8a and 8b: The confusion matrix and accuracy score of the testing dataset for the purely clinical model (age, sex, smoking status) and the clinicogenomic model (age, sex, smoking status, and 4-gene signature)

In addition to the confusion matrix, the ROC AUC score of the two models is shown in Figure 9a and b. The extremely high AUC score of the clinogenomic model (0.97) as compared to the low AUC of the purely clinical model (0.65) shows that the addition of genetic information from the 4-gene signature greatly improves the model. The clinicogenomic model is performing at high rates and is accurate in all aspects when diagnosing NSCLC.



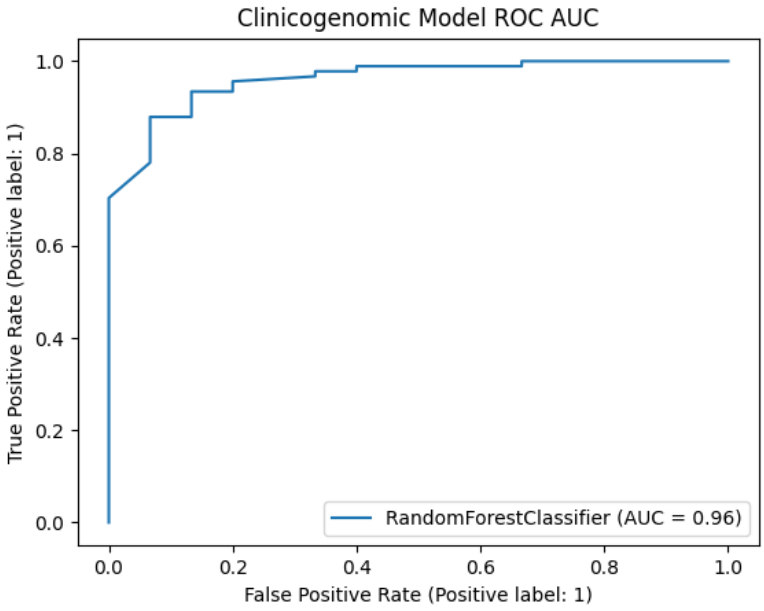


Figure 9: ROC AUC score for purely clinical model and clinicogenomic model

6 Discussion

NSCLC, the main histological type of lung cancer, is one of the deadliest cancers in the world. Thus, it is important to identify effective biomarkers for the exploration of the pathogenesis of NSCLC.

In the first stage of this work, I identified 267 differentially expressed genes (DEGs), with KEGG pathway enrichment analysis showing that the DEGs were significantly enriched in pathways relating to ECM-receptor interactions and signaling. These pathways are all closely related to cancer, affecting many cellular processes including motility, proliferation, regulation of gene expression, and cell survival. These pathways have been found to be significantly associated with many types of cancers.

To further screen DEGs and explore the relationships between genes and diseases, I constructed a complex protein protein-interaction (PPI) network in *Cytoscape*. The PPI network enables recognition of functional modules, prediction of protein functions, and identifies important pathways. The biological network assesses the interactions between DEGs and the intrinsic mechanisms between genes and diseases. To ensure a completely comprehensive study, all 12 topological scoring methods were utilized in *Cytoscape* to identify the top DEGs in the network. The performance of the DEGs was evaluated using AUC, and in the end, I identified a 4-gene biomarker signature consisting of *AGER, CA4, RASIP1,* and *CAV1* that was the highest performing in the entire dataset. These biomarkers are significantly enriched in GO terms of receptor activity, immune response, extracellular matrix, and signaling activity, which all play an important role in regulating proliferation, differentiation, and apoptosis. The significant downregulation of these four genes in cancerous patients signifies that a change in the expression of these genes disrupts the ECM, promoting tumorigenesis. These four genes can be further explored as possible therapeutical targets for future drug treatment.

To confirm the relatability of my methods, the 4-gene signature was validated in the TCGA database, as well as with Kaplan-Meier survival analysis. Additionally, the clinicogenomic diagnostic model that utilized both the 4-gene signature and clinical attributes (age, sex, smoking status) in the third stage, shows the significant improvement that the 4-gene signature provides on a pure clinical model.

In the second stage, through an innovative investigation of network topological scoring methods, I prove that the most conventionally used method, degree, does not achieve the highest performance, and I instead propose a novel composite index (C-index), that combines clustering coefficient, bottleneck, and eccentricity for the strongest performance. These three metrics focus heavily on gene interactions in the network, and locate communities of genes or genes that have high connectivity with others.

The clustering coefficient of a network measures the tendency of the nodes to cluster together. High clustering means that the network contains densely connected communities of nodes. In biological networks, these topological clusters reflect biological function. These communities reflect functional modules and gene complexes that work closely together and share similar functions. Rather than finding individual nodes with high connectivity like degree, clustering coefficient identifies significant communities that work together to induce cancer. These communities are most likely involved in pathways that promote tumorigenesis. In biological networks, my results indicate that clustering coefficient is much more robust in locating essential genes.

As shown by Table 5, the two global metrics Betweenness and Bottleneck produced a very similar list of top scoring nodes, with a one node difference (SFTPD). This is due to the similarity of their definitions. Betweenness measures the number of shortest paths going through a certain node. Bottlenecks are genes with the highest betweenness centrality, and control most of the information flow in the network. These bottleneck nodes represent critical points of the network. In biological interaction networks, bottlenecks are key connectors in protein networks that participate in several regulatory pathways that control the interactions of a large number of proteins. My results show that in biological networks and applications, betweenness (or bottleneck) is a better predictor of essentiality than degree.

In a biological network, eccentricity essentially measures node proximity. A node with higher eccentricity means that all other nodes are in proximity. In a biological network, the eccentricity evaluates the ease of a gene to be functionally reached by all other genes in a network. Genes with the highest eccentricities are easily influenced by other genes, and conversely influence other genes with more ease. This indicates that these genes play important roles in functional pathways that may promote cancer. Although previous studies have largely used degree of freedom to find "hub" genes, I have confirmed that the C-index (clustering coefficient, bottleneck, and eccentricity) produce much more robust results.

Throughout this research, both when searching for high performing biomarkers and in my topological network study, I have proven that using multiple biomarkers and methods concurrently greatly improves performance. This is because genes work together in pathways that lead to tumorigenesis, and singular genes cannot cause cancer without having numerous regulatory effects on other genes through signal transduction pathways. Cancer is a complex disease cause by the interaction of multiple environmental factors and genes. It is the combined effect of all of these genes in the pathway together that leads to cancer onset. With further validation and refinement in other cancer datasets, the 4-gene biomarker signature and C-index may be transformative in the study of biomarkers and all cancers, and provide an experimental foundation for further exploration of the usage of PPI networks to diagnose cancer.

7 Conclusion

In this work, I have conducted a three-stage investigation for the early diagnosis of non-small cell lung cancer: 1) performing comprehensive analysis of a merged dataset to identify a 4-gene biomarker signature that achieves a 92% AUC diagnostic performance, 2) introducing a novel and systematic method to identify the top topological metric in protein networks that is essential for biomarker selection, and 3) exploiting the use of a clinicogenomic machine learning model with top biomarkers selected and clinical covariates for NSCLC diagnosis, with 95% accuracy and 96% AUC. More specifically, I proposed a novel composite selection index (C-index) that can concurrently consider complementary factors in biomarker selection to greatly increase accuracy of NSCLC diagnosis, and the performance studies show the increase of AUC from 75% to 92%. This demonstrates the effectiveness of the proposed method in finding critical biomarkers for accurate yet low-complexity diagnosis of NSCLC. The proposed methodology of finding top metrics can be extended to effectively and efficiently select biomarkers in various types of cancers, fundamentally advancing topological network research and the continuous pursuit of cancer prevention.

8 References

1. Liu, X., Liu, X., Li, J., & Ren, F. (2019). Identification and Integrated Analysis of Key Biomarkers for Diagnosis and Prognosis of Non-Small Cell Lung Cancer. *Medical Science Monitor*, *25*, 9280–9289. https://doi.org/10.12659/msm.918620
2. Long, T., Liu, Z., Zhou, X., Yu, S., Tian, H., & Bao, Y. (2019). Identification of differentially expressed genes and enriched pathways in lung cancer using bioinformatics analysis. *Molecular Medicine Reports*. https://doi.org/10.3892/mmr.2019.9878
3. Wang, L., Qu, J., Liang, Y., Zhao, D., Rehman, F. U., Qin, K., & Zhang, X. (2020). Identification and validation of key genes with prognostic value in non-small-cell lung cancer via integrated bioinformatics analysis. *Thoracic Cancer*, *11*(4), 851–866. https://doi.org/10.1111/1759-7714.13298
4. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, *41*(D1), D991–D995. https://doi.org/10.1093/nar/gks1193
5. Huang da W, Sherman BT, Lempicki RA: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009, 37:1–13.
6. Walter W, Sanchez-Cabo F, Ricote M. GOplot: an R package for visually combining expression data with functional analysis. Bioinformatics. 2015;31:2912–4.
7. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 2011;39:D561–8.

8. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27, 431–432, https://doi.org/10.1093/bioinformatics/btq675 (2011).
9. Robin, X., Turck, N., Hainard, A. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77 (2011). https://doi.org/10.1186/1471-2105-12-77
10. Chin, CH., Chen, SH., Wu, HH. et al. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol 8, S11 (2014). https://doi.org/10.1186/1752-0509-8-S4-S11
11. Tang Z, Li C, Kang B, et al. GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. 2017;45:W98–102.
12. Chen, X., Wang, M., & Zhang, H. (2011). The use of classification trees for bioinformatics. WIREs Data Mining and Knowledge Discovery, 1(1), 55–63. https://doi.org/10.1002/widm.14
13. Wendl, M. C., Wallis, J. W., Lin, L., Kandoth, C., Mardis, E. R., Wilson, R. K., & Ding, L. (2011). PathScan: a tool for discerning mutational significance in groups of putative cancer genes. Bioinformatics, 27(12), 1595–1602. https://doi.org/10.1093/bioinformatics/btr193
14. Sanchez-Palencia, A., Gomez-Morales, M., Gomez-Capilla, J. A., Pedraza, V., Boyero, L., Rosell, R., & Fárez-Vidal, M. E. (2010). Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. International Journal of Cancer, 129(2), 355–364. https://doi.org/10.1002/ijc.25704
15. Ma, Q., Xu, Y., Liao, H., Cai, Y., Xu, L., Xiao, D., Liu, C., Pu, W., Zhong, X., & Guo, X. (2019). Identification and validation of key genes associated with non-small-cell lung cancer. Journal of Cellular Physiology, 234(12), 22742–22752. https://doi.org/10.1002/jcp.28839