scientific reports

Check for updates

OPEN

A novel approach to topological network analysis for the identification of metrics and signatures in non-small cell lung cancer

Isabella Wu[1]✉ & Xin Wang[2]

Non-small cell lung cancer (NSCLC), the primary histological form of lung cancer, accounts for about 25%—the highest—of all cancer deaths. As NSCLC is often undetected until symptoms appear in the late stages, it is imperative to discover more effective tumor-associated biomarkers for early diagnosis. Topological data analysis is one of the most powerful methodologies applicable to biological networks. However, current studies fail to consider the biological significance of their quantitative methods and utilize popular scoring metrics without verification, leading to low performance. To extract meaningful insights from genomic data, it is essential to understand the relationship between geometric correlations and biological function mechanisms. Through bioinformatics and network analyses, we propose a novel composite selection index, the C-Index, that best captures significant pathways and interactions in gene networks to identify biomarkers with the highest efficiency and accuracy. Furthermore, we establish a 4-gene biomarker signature that serves as a promising therapeutic target for NSCLC and personalized medicine. The C-Index and biomarkers discovered were validated with robust machine learning models. The methodology proposed for finding top metrics can be applied to effectively select biomarkers and early diagnose many diseases, revolutionizing the approach to topological network research for all cancers.

Lung cancer is the deadliest cancer worldwide and is highly fatal, with a 5-year survival rate of $< 21\%$[1]. Non-small cell lung cancer (NSCLC) accounts for about 85% of all lung cancer cases[2]. 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[3], making treatment extremely difficult. This clinical need has driven the search for more effective prevention and early diagnosis strategies, including the identification of effective biomarkers.

By providing insights into the molecular origins and behaviors of NSCLC, biomarkers help identify high-risk patients and targets for personalized medicine and the development of targeted therapies[4,5]. The expansion of the availability and quantity of molecular biological data has created a pressing need for improved computational methods for data analysis. Studies have revealed that genes do not function in isolation, and instead work together. As such, network-based approaches have emerged as powerful tools for investigating gene interactions and understanding their complex relationships[6]. In the past few years, topological data analysis (TDA) has revolutionized the oncology field, becoming one of the most powerful and widespread tools to extract useful information from high-dimensional biomedical data[7,8]. We can use TDA to analyze the interactions between genes and identify essential biomarkers. Network nodes are ranked by scoring methods that reflect varying network features[9,10]. The high-scoring nodes are selected as the most essential genes in the network and further analyzed as potential markers.

Despite recent advances in applying TDA to cancer studies, one critical issue remains largely overlooked—whether geometric and topological connectivity implies functional connectivity in biological networks[11,12]. A comprehensive analysis of the correlation between geometric connectivity and functional connectivity is currently lacking, leading to ineffective use of the powerful methodology. The performance of existing works is not high, largely due to the lack of understanding of the most key aspect of topological analysis—the network scoring

[1]Choate Rosemary Hall, Wallingford 06492, USA. [2]Electrical Engineering, Stony Brook University, Stony Brook 11790, USA. ✉email: iwu24@choate.edu

metrics. They rank the network nodes by different features, evaluating the network in varying aspects to select the high score nodes[12]. Inherently, these scoring metrics are defined geometrically with no clear implications for functional significance. Existing studies often fail to biologically validate the scoring methods they choose to use in biomarker identification for disease prediction[13,14]. They often utilize popular and conventional geometric-based scoring metrics (e.g. Degree)[15–17] without reasoning, and use only a single scoring metric.

Thus, the central goal of this study is to identify the top-performing topological scoring metric (or method) that best captures functional significance in protein networks. More specifically, we focus on designing a metric, or a composition of metrics, that identifies critical biomarkers most efficiently and accurately. To do so, we conducted systematic and comprehensive analysis and validation in two main stages.

First, to investigate the ability of the metrics to select effective biomarkers, we performed detailed functional enrichment and network analyses to identify the top biomarkers for NSCLC diagnosis. To exploit the power of gene interactions, we further explore the concurrent use of multiple biomarkers in NSCLC prediction and develop a biomarker-signature consisting of a small group of top biomarkers. To determine their diagnostic ability, we introduce a method to calculate the area under the receiver operating characteristic curve (AUC) for multiple biomarkers concurrently, known as Integrated AUC. By doing so, we directly evaluate the biomarkers by their functional abilities to diagnose NSCLC. In the second stage, we build off the previous stage to identify the top-performing topological scoring metrics (or methods) by evaluating their ability to select biomarkers with the highest diagnostic capabilities most efficiently and accurately. We propose a novel composite selection index that concurrently considers complementary factors to evaluate the network and produce biologically significant results.

We expect the proposed methods to fundamentally advance topological network research in the cancer field. We unlock a better understanding of how geometric relationships in gene networks relate to functional mechanisms to extract meaningful insights from genomic data. Our proposed methods are not restricted to the use in diagnosing NSCLC, but can be extended to the early diagnosis of other types of cancers. With the widespread use of topological network analysis, we expect the proposed techniques to drastically improve the search for new biomarkers and targets for drug therapy and redefine the usage of TDA in oncology.

Results

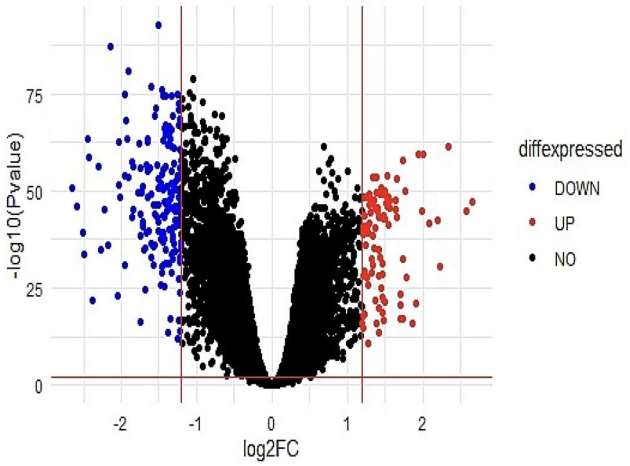
Stage I: Biomarker signature identification. Identifying specific and sensitive biomarkers is critical for early cancer diagnosis[18]. As the first stage of this study, we identified the top biomarkers to serve as a base for us to explore in depth the critical scoring metrics for biomarker identification in the next stage. An overview of this study and its two main stages is presented in Supplementary Fig. S1.

DEG screening and functional enrichment. In total, we identified 267 differentially expressed genes (DEGs) (p-value < 0.01 and $|\log FC| > 1.2$) from the three datasets GSE31210, GSE33356, and GSE50081, taken from the Gene Expression Omnibus (GEO)[19]. The three datasets are summarized in Supplementary Table S1. 93 were upregulated and 174 were downregulated (Fig. 1a). To investigate the roles that the DEGs play in disease mechanisms, we examined the DEG-related pathways. Through enrichment analyses, we identified 10 enriched GO terms with FDR < 0.05 (Fig. 1c), and their z-score expression values (Fig. 1b).

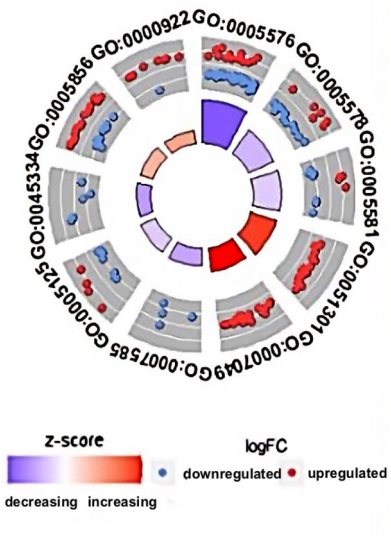
Upregulated DEGs play a role in multiple pathways that promote tumorigenesis, including cell division (GO:0051301), cell cycle (GO:0007049), cytoskeleton (GO:0005856), and spindle pole (GO:0000922). Downregulated genes were significantly involved in weakened tumor defense and disruptions in signal transduction pathways, such as decreased cytokine activity (GO:0005125) and clathrin-coated endocytic vesicles (GO:0045334). Downregulated genes were also enriched in the extracellular matrix (GO:0005576 and GO:0005578), and ECM-receptor interaction (hsa04512, FDR = 0.0152) was identified through KEGG pathway analysis. 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[20]. Our results show that the DEGs are largely connected to disease-related pathways and play potent roles in cancer onset and development.

PPI network analysis. Disease susceptibility and other disease correlated factors are due to the perturbation of an interconnected gene network[21], not single gene mutations in isolation. To explore the interactions between DEGs and identify intrinsic mechanisms of disease, we constructed a protein-protein interaction (PPI) network (Fig. 2a) based on our 267 DEGs to understand the topology of molecular interactions and identify the most essential top-scoring biomarkers in the network. The network was visualized using Cytoscape[22] and further analyzed with the CytoHubba algorithm[23]. The nodes in the center of the network (Fig. 2b) and the cluster modulus (Fig. 2c) are zoomed in, as these significant regions contain highly connected nodes that have great impact on the other nodes in the network. The topological scoring metrics in CytoHubba are divided into two categories, *local* to evaluate individual nodes and *global* to evaluate the network as a whole. 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 (EPC). Often, literature studies utilize only one scoring method[15–17]. To ensure that no essential genes are missed and all possibilities are considered, we created a complete and comprehensive list of candidate biomarkers using all twelve scoring methods.

Each metric was utilized to select 10 top nodes each, with some having a higher cutoff because a few nodes share the same ranking score. Without counting overlapping genes between metrics, we obtained 82 candidate biomarkers (Table 1a) overall. To evaluate the ability of biomarkers in distinguishing between disease and control, we chose to use area under the receiver operating characteristic curve (AUC) score to select the overall top 20



(a) Volcano plot of DEGs. Red represents up-regulated genes and blue represents down-regulated genes.



(b) GO term enrichment analysis of DEGs, statistically significant Gene Ontology terms (FDR<0.05).

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

(c) Significant GO terms and their FDR values. Pink represents Cellular Component (CC), yellow represents Biological Process (BP), and green represents Molecular Function (MF).

Figure 1. GO and KEGG analyses of DEGs.

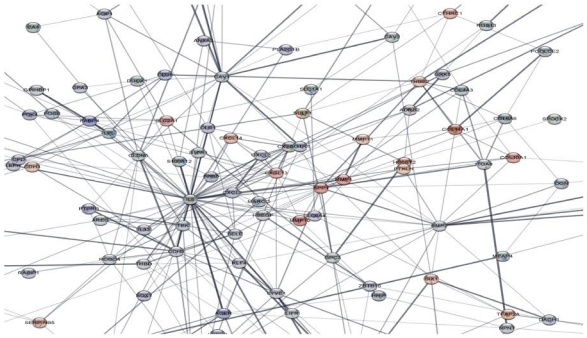
disease-correlated genes in the network (Table 1b). This allows us to directly evaluate the biomarkers by their ability to predict disease.

Disease prediction with multiple biomarkers simultaneously. Cancer is caused by multiple genes in a functional or signaling pathway working together in a cascade of mutations to promote tumorigenesis, and can never be caused or predicted by a single mutation or gene. Rather than only considering the capability of individual biomarkers as commonly done in literature studies, we further explore the concurrent use of multiple biomarkers in NSCLC prediction to vastly increase the diagnostic performance,

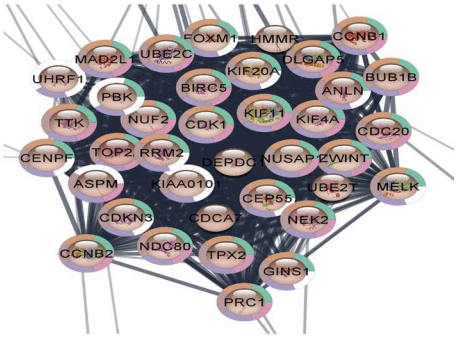
Utilizing multiple biomarkers is more comprehensive, and may better deal with disease heterogeneity and reduce anomalies during prediction.To the best of our knowledge, the incremental usefulness of adding multiple biomarkers from different disease pathways has not been fully evaluated amongst other NSCLC studies. Simply using too many may decrease performance. We need to find the optimal number of biomarkers to use concurrently for the highest performance, as well as a way to evaluate the joint performance of biomarkers.

As each biomarker's expression value quantifies its relationship to the health condition of a subject, we propose the concept of *Integrated AUC* to calculate the AUC of the aggregated expression of biomarkers. In this study, the aggregated expression is defined as the mean expression of the group of biomarkers as it is most suitable, but its definition may be expanded.

(a) Overview of network. Red nodes represent up-regulated genes and blue nodes represent down-regulated genes. The darker the color, the greater the |logFC| value.



(b) A closer view of the center of network



(c) Significant cluster module

Figure 2. PPI network of DEGs and CytoHubba visualization.

After evaluating the performance of several different gene groupings, we found that the top 4 biomarkers *AGER*, *CA4*, *RASIP1*, and *CAV1* together produced the highest Integrated AUC at 0.9238 (Fig. 3a). They make up our **4-gene biomarker signature** for the prediction of NSCLC. Their Receiver Operating Characteristic (ROC) curves are visualized in Fig. 3c. For comparison purposes, we also calculated the AUC of the top 10 disease-correlated genes in the network (from Table 1b) combined. As expected, the 4-gene signature outperformed 10 genes (Fig. 3b). This may be due to the nature of interactions between genes—certain biomarkers may not interact optimally for high performance, or biomarkers outside of the top four may play roles of lower significance. The finding of the biomarker signature and the use of Integrated AUC as an evaluation metric can be expanded on and further explored to improve the prediction accuracy of other types of cancers.

Validation of 4-gene signature by survival analysis and TCGA database. At the beginning of the study, we randomly divided the dataset into 80% for identification of biomarkers and metrics, and 20% validation. To validate the effectiveness of using the 4-gene signature, we compared its performance in the validation data set to each of its 4 individual components, as well as that of the top 10 genes combined. The 4-gene signature outperformed all of its individual genes, as well as the top 10 genes together (Supplementary Table S2), demonstrating that it is less complex and more effective than using more genes, and more powerful than only using individual biomarkers alone.

To validate the effect of the 4-gene signature on NSCLC prognosis, we performed overall survival analysis with these genes using Kaplan-Meier survival plots (Fig. 4a) to examine their impact on patient survival, with a threshold of p-value < 0.01 to determine significance. The low expression of *AGER*, *CA4*, *RASIP1*, and *CAV1* are all associated with poor overall survival, indicating their significant roles in NSCLC prognosis.

To confirm that our results are applicable outside our data set, TCGA data from the GEPIA interactive website was used to verify the identified genes to be effective amongst other NSCLC cases. Figure 4b compares gene expressions from two histological types of NSCLC, lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD), and normal lung tissues. All four genes are significantly downregulated in cancerous patients, indicating that the four-gene signature can be expanded and used for new patients and data. From our GO enrichment analysis, these genes play important roles in the ECM matrix and signalling pathways. They are important for cancer detection and treatment, and can act as therapeutic targets for future drug therapy and personalized medicine.

Stage II: Critical network topological metrics. In the past decade, topological data analysis has grown into a prominent role in the oncology field. The scoring metrics are crucial to topological data analysis as they identify the most influential network nodes. Scoring metrics, however, describe only geometric relationships

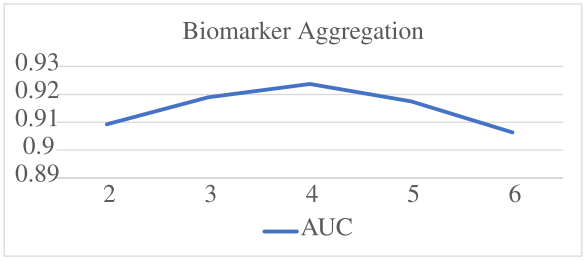
(a) Top ranked nodes found through all 12 topological scoring methods			
Scoring method	**Top-ranked genes (listed in order of disease scores)**		
Degree	IL6, ASPM, TTK, KIF20A, TPX2, CENPF, UBE2C, CCNB1, CDK1, NUF2, KIF11, KIF4A, DLGAP5, TOP2A, MAD2L1, PBK, BIRC5		
MCC	ASPM, BIRC5, BUB1B, CNNB1, CCNB2, 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		
(b) Top 20 genes detected in network analysis based on AUC score			
Gene name	**AUC**	**logFC**	**FDR (adj-pval)**
AGER	0.8979	−2.13691	4.45E−84
CA4	0.8952	−1.50475	1.70E−89
RASIP1	0.8899	−1.20962	3.52E−65
CAV1	0.8741	−1.41913	5.23E−55
FAM107A	0.8731	−1.94635	1.39E−72
CDH5	0.8725	−1.40895	3.69E−63
TEK	0.8724	−1.45729419	2.13E−73
CLDN5	0.8697	−1.21244	1.24E−56
CLIC5	0.8686	−2.03743	4.42E−61
SPP1	0.8647	2.334613	6.97E−60
KIF26B	0.8636	1.293536	2.98E−48
PTRB	0.8604	−1.653827903	1.57E−61
SOX7	0.8560	−1.599491119	2.45E−48
SDPR	0.8542	−1.861487998	2.53E−56
TOP2A	0.8516	2.014367341	4.75E−58
TMPRSS4	0.8459	1.634397285	1.70E−44
AGTR1	0.8446	−1.503517517	5.72E−52
CDCA7	0.8417	1.650737609	6.42E−52
GPX3	0.8416	−1.388971807	9.69E−46

Table 1. Top biomarkers in the PPI network based on all 12 network scoring metrics.

between nodes in the network, not their relation to disease diagnosis. Consequentially, the significance and implications of the scoring methods are vastly overlooked in most biological studies, and a single metric is often employed without reasoning[24]. On the other side of the spectrum, many studies simply use all 12 metrices together, which is also ineffective[25]. The most frequently used metric is Degree, a local metric that may not capture the full extent of gene network interactions. Our foremost goal in the second stage of this work is to thoroughly analyze the performance of these quantitative methods applied to biological networks and identify the top metrics that best capture functional connectivity and biological implications.

We set out to achieve this in two major phases. We first evaluate all 12 topological scoring metrics based on their diagnostic ability for effective biomarker selection. In the search for biomarkers, using a single metric may be inadequate, while using all metrics together is complex and inefficient, and possibly erroneous. Therefore, in the second phase, we further investigate the performance of using multiple metrics concurrently in diagnosis and design a powerful metric. Few previous studies have meaningfully utilized multiple network metrics concurrently to look for biomarkers[17,26]. Our investigations indicate that increasing the analytical coverage of a biological system through optimal pairings of metrics leads to more robust results.

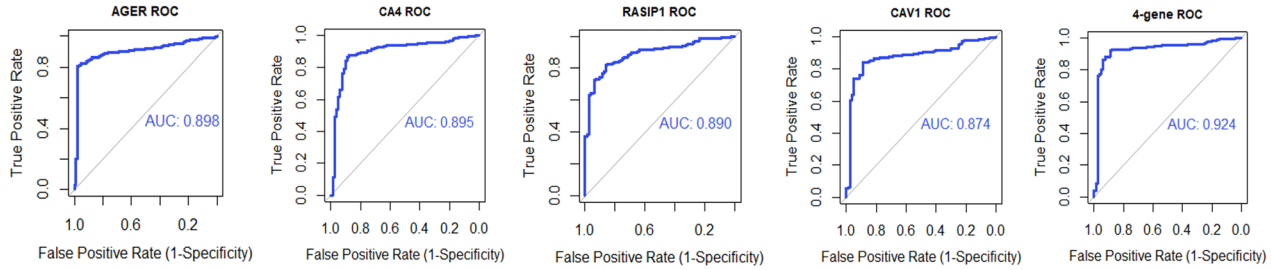
Evaluating the performance of individual metrics. To evaluate the ability of the metrics to identify essential biomarkers, we calculated the Integrated AUC of biomarkers selected by each metric in Table 1a. For compari-



(a) Integrated AUC reached its peak with 4 genes

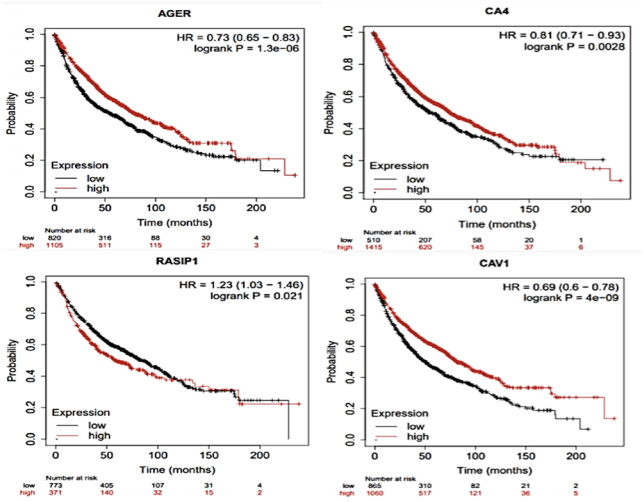
Biomarkers	Integrated AUC
AGER+RASIP1	0.9093
AGER+RASIP1+CA4	0.9190
AGER+RASIP1+CA4+CAV1	0.9238
AGER+RASIP1+CA4+CAV1+FAM107A	0.9175
All Top 10	0.9064

(b) Top biomarker combinations

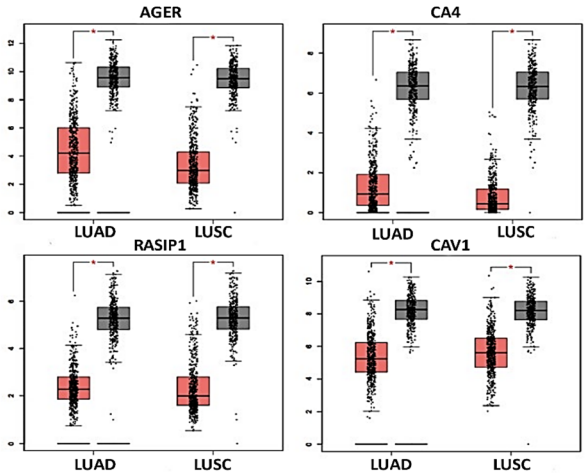


(c) ROC curves of AGER, CA4, RASIP, CAV1, and the 4-gene signature

Figure 3. Performance of 4-gene signature.



(a) Survival plot of the significant genes by Kaplan Meier test. The 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.



(b) Comparison of gene expression between NSCLC and control patients (red = tumor, gray = control). All four genes are downregulated in cancer patients.

Figure 4. Differential expression of 4-gene signature in NSCLC patients.

son, the mean of the AUCs of individual genes selected by each metric was also calculated. However, without considering the diagnostic effects of genes together, the performance of the mean individual AUC is lower, as it cannot effectively capture the interactions between genes and is also more susceptible to outliers. In general, Integrated AUC has higher performance and is more robust, and will be used as the main performance measure in this study.

The top local and global metrics found are Clustering Coefficient (AUC = 0.8770) and Bottleneck (AUC = 0.8609) respectively (Table 2). Compared to the commonly used Degree[27], the AUC of Clustering Coefficient is 11% higher. Rather than simply using Degree to select genes with many connections, these results explore and confirm that other scoring methods better capture the functional features of protein networks. With Integrated AUC, we can validate the connection between the geometric metrics and their functional significance.

Scoring metric	Integrated AUC	Mean individual AUC
Clustering coefficient	0.8770	0.8159
Bottleneck	0.8609	0.8012
Betweenness	0.8517	0.8091
Eccentricity	0.8409	0.8017
Stress	0.8179	0.7919
DMNC	0.8136	0.7860
MCC	0.8114	0.7874
EPC	0.8030	0.7959
Degree	0.7875	0.7877
MNC	0.7748	0.7890
Radiality	0.7904	0.8100
Closeness	0.6366	0.7986

Table 2. Scoring metrics ranked by AUC.

Design of C-Index. In this second stage, we systematically composed metrics to find an ideal set of scoring methods that can adequately capture the interactions in biological networks without incurring high complexity. We evaluated composite performances using Integrated AUC of the top 10 and 20 genes of the superset of the metrics. Furthermore, we assessed the performance of the metrics using a Precision measure, as defined in the Methods section. Precision measures the accuracy of the composite metrics in identifying the top 10 and 20 disease-correlated genes (by AUC score) listed in Table 1b.

To measure both the node-level properties and the network-level properties, we chose to compose the top local and global metrics (which were also the top two metrics overall), Clustering Coefficient and Bottleneck (Fig. 5). 7 out of the top 10 overall genes were correctly identified, resulting in a 70% precision, and the Integrated AUC of the top 10 nodes was 0.9027. Likewise, the top 20 nodes had 70% precision, and Integrated AUC was 0.8765. Already, the composition results in a higher AUC than any of the individual metrics.

To determine whether the composition of Clustering Coefficient and Bottleneck is the most optimal, we also composed the three metrics Clustering Coefficient, Bottleneck, and Eccentricity. Although Betweenness outperformed Eccentricity, we chose to use Eccentricity because by definition Betweenness is essentially the same measure as Bottleneck and is thus not needed in the composite. With three, although the precision increased, the Integrated AUC reduced slightly. We also composed the top four metrics but obtained no improvement in AUC or precision. Considering the vastly increased complexity of adding more metrics, using two metrics together is ultimately more powerful than using three or more.

Our investigations have shown that Clustering Coefficient and Bottleneck can work together to adequately capture the significant biomarkers in the network with the most efficiency, and we propose to compose them as a new metric, **C-Index**. C-Index is defined as

Combined Metrices	Overall Integrated AUC (All genes)	Integrated AUC (Top 10/Top 20)		Precision of Top Genes (Top 10/Top 20)		Gene Union List (Top 10/Top 20)	
Clustering Coefficient + Bottleneck	0.8824	0.9027	0.8765	0.70 (70%)	0.70 (70%)	AGER, RASIP1, CA4, FAM107A, CDH5, CAV1, SPP1, PTPRB, SOX7, SDPR	AGER, RASIP1, CA4, FAM107A, CDH5, CAV1, SPP1, PTPRB, SOX7, SDPR, TMPRSS4, CDCA7, GPX3, HS6ST2, ACADL, FHL1, GINS1, BCHE, NPNT, WIF1
Clustering Coefficient + Bottleneck + Eccentricity	0.8614	0.8906	0.8636	0.80 (80%)	0.85 (85%)	AGER, RASIP1, CA4, FAM107A, CDH5, CLIC5, CAV1, SPP1, KIF26B, PTPRB	AGER, RASIP1, CA4, FAM107A, CDH5, CLIC5, CAV1, SPP1, KIF26B, PTPRB, SOX7, SDPR, TMPRSS4, AGTR1, CDCA7, GPX3, HS6ST2, ACADL, FHL1, GINS1
Clustering Coefficient + Bottleneck + Eccentricity+ Stress	0.8516	0.8906	0.8636	0.80 (80%)	0.85 (85%)	AGER, RASIP1, CA4, FAM107A, CDH5, CLIC5, CAV1, SPP1, KIF26B, PTPRB	AGER, RASIP1, CA4, FAM107A, CDH5, CLIC5, CAV1, SPP1, KIF26B, PTPRB, SOX7, SDPR, TMPRSS4, AGTR1, CDCA7, GPX3, HS6ST2, ACADL, FHL1, GINS1
Degree + Betweenness	0.6178	0.6295	0.6721	0.40 (40%)	0.25 (25%)	AGER, CAV1, CDH5, SPP1, TOP2A, NPNT, CCNB1, CENPF, KIF20A, DLGAP5	AGER, CAV1, CDH5, SPP1, TOP2A, NPNT, CCNB1, CENPF, KIF20A, DLGAP5, UBE2C, KIF11, AQP4, BIRC5, TTK, TPX2, NUF2, ASPM, BMP2, KIF4A

Figure 5. Performance of composite network scoring metrics.

$$\text{C-Index} = N_{\text{top}}(S_{\text{clustering coeff}} \cup S_{\text{bottleneck}}) \quad (1)$$

where N_{top} represents the top genes selected from the union of the sets $S_{\text{clustering coeff}}$ and $S_{\text{bottleneck}}$. These are the sets of top genes identified by Clustering Coefficient and Bottleneck respectively.

C-Index outperforms all individual metrics and other compositions. It can achieve the comprehensiveness of using all metrics while vastly decreasing the complexity and inefficiency. Clustering Coefficient and Bottleneck work together as local and global metrics to better capture the nature of cancer and are capable of identifying essential genes that are prevalent throughout the entire network as well as locally.

Most literature studies use Degree and occasionally Betweenness to find their top biomarkers[28,29]. To investigate whether the combination of these two metrics results in effective biomarker selection, we calculated the performance of their composition. We obtained shockingly poor results: the top 10 genes had 40% precision, and 0.6295 Integrated AUC, while the top 20 genes had only 25% precision, and 0.6721 Integrated AUC. This poor performance is due to several factors that are often overlooked by other studies. Degree, a local metric, only measures how many connections a node has, which may signify that it regulates many other genes. However, these highly connected "hubs" have no great overall influence in cancer-causing pathways because their neighbors are not necessarily interconnected. In contrast, the Clustering Coefficient of a network, also a local metric, measures the tendency of nodes to form densely connected communities. In biological networks, these communities signify functional modules and gene complexes that work closely together and share similar functions. Clustering Coefficient identifies key biomarkers located in significant communities that work together to induce cancer and is a much more robust local metric than Degree in locating essential genes.

Although Clustering Coefficient alone is a powerful metric, it is local and unable to capture the overall topologies of the network. To gain a better view of the essential genes that play important roles in the network as a whole, Bottleneck is used complementary with Clustering Coefficient. Bottlenecks, genes with the highest betweenness centrality, control most of the information flow and interaction between proteins and are key connectors between regulatory pathways that are identified through Clustering Coefficient.

Together, Clustering Coefficient and Bottleneck are capable of considering multiple biological pathways, and improve on the performance of Degree from 78.75 to 88.24%. Not only does C-Index perform better than individual metrics, but it matches the performance of using all 12 metrics together with greatly reduced complexity and cost. The AUC of the overall top 10 biomarkers identified using all 12 metrics was 0.9064, as calculated in Stage I of this study. The AUC of the top 10 biomarkers identified using C-Index was 0.9027. It is as powerful as using all twelve, while much more efficient.

Our results indicate that the proposed C-Index is capable of capturing the most critical group of interactive genes that can early diagnose cancer. Our method of evaluating scoring methods is transformative and a breakthrough in applying topological analysis to biomarker identification.

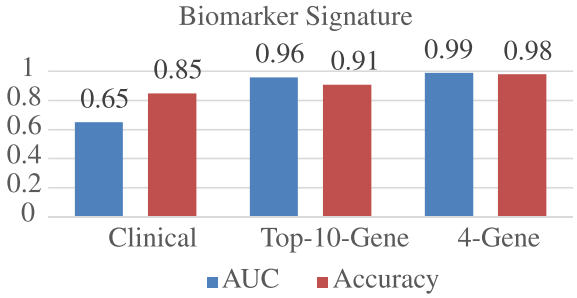
Performance of C-Index in validation set. To validate the C-Index, we calculated the Integrated AUC of the metrics in the 20% validation set. As predicted, C-Index outperformed Degree and each of its components (Supplementary Table S3). Compared to Degree, the performance of using C-Index to select biomarkers increased by 25%. Clustering Coefficient alone improved on Degree by 22.4%. Compared to using all 82 genes found by all 12 metrics, the performance of using only genes found by C-Index is 40% higher. This demonstrates that the metrics that compose C-Index vastly improve on conventional ones, and are capable of identifying a concise list of significant genes most successfully. Benefiting from both metrics with close interactions locally and genes that lie on the critical paths linked globally, C-index can revolutionarily select the most representative group of biomarkers at a low computational cost for accurate diagnosis.

Machine learning validation of biomarkers and metrics. We further validated the above results with a random forest machine learning model to evaluate their capabilities to classify cancer from no cancer in the testing dataset. The model was trained in the 80% dataset, and validated in the 20% dataset. The model considers not only genetic attributes, but also clinical attributes such as age, gender, and smoking status that most impact the development of NSCLC to improve the accuracy of diagnosis. Literature studies often decouple genomic and clinical attributes in cancer prediction. To incorporate both types of attributes and provide a validation method that does not involve Integrated AUC, we design this clinicogenomic model to explore their potential for disease diagnosis.

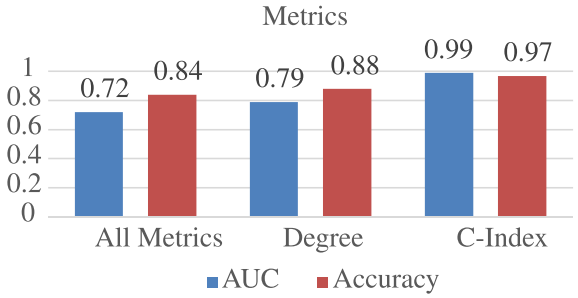
With the addition of the 4-gene signature, the purely clinical model was improved by 51% (Fig. 6a). Compared to using biomarkers alone, the addition of clinical variates also increased performance. The 4-gene model once again outperforms the top-10 gene model, validating the effectiveness of the signature. In addition to validating its ability to distinguish cancer from no cancer, the 4-gene model was further extended to diagnose early vs. late disease stage, as well as NSCLC stages I–IV with a multi-stage Cascading Model design that is described in greater detail in "Methods" (Fig. 6d).

C-Index outperformed the conventional metric by 25% and all the metrics together by 37% (Fig. 6b). Utilizing all metrics suffered from the existence of outliers and inaccuracy. Degree, on the other side of the spectrum, was unable to identify a complete set of genes that could accurately capture the interactions of the network and thus performed poorly.

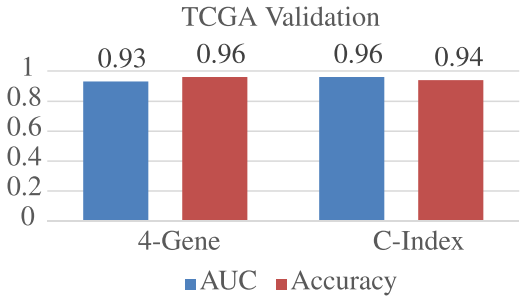
Furthermore, validation was conducted not only on the testing set, but also on the TCGA LUSC and LUAD cancer datasets (Fig. 6c), obtained from UCSC Xena[30], which were processed in the same manner as the GEO dataset. The pretrained 4-gene and C-Index models were assessed on their ability to discern between cancer and non-cancer cases in this new validation cohort. The 4-gene model achieved 96% accuracy, and the C-Index



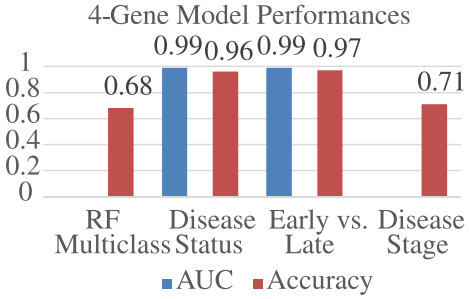
(a) The 4-gene clinicogenomic model outperformed both the clinical model (51% increase) and 10-gene clinicogenomic model (3.5% increase).



(b) C-Index outperformed both the conventional metric (25% accuracy increase) and all 12 metrics together (37% accuracy increase).



(c) The performance of the 4-gene and C-Index models in the TCGA LUSC and LUAD datasets.



(d) The performance of all three stages of the 4-gene Cascading Model in comparison to a multi-stage random forest model.

Figure 6. Machine learning validation of biomarkers and metrics.

model achieved 94% accuracy. Both models exhibited high performance, supporting the efficacy of the metrics and signatures presented in this study.

Discussion

It is vitally important to identify critical biomarkers for exploring the pathogenesis of NSCLC, one of the deadliest cancers in the world. The key to improving prevention and early diagnosis is to find metrics and methods that can guide the effective yet cost efficient search of biomarkers.

There are several important findings from this study. First, through a series of functional, network, and statistical analyses, we identify a 4-gene biomarker signature consisting of *AGER*, *CA4*, *RASIP1*, and *CAV1* that can be further explored as possible therapeutic targets for drug treatment. Second, we prove that the most widely used topological scoring metric, Degree, is not the best suited for biological networks. We instead propose C-Index, a novel composite index that combines Clustering Coefficient and Bottleneck to best capture the interactions in gene networks for high efficiency and performance. Our results solidify the connection between geometric connectivity and functional connectivity. For validating the C-Index and 4-gene signature, we exploited the use of a machine learning model that considers both genomic and clinical factors concurrently.

To the best of our knowledge, this is the first study that comprehensively evaluates all 12 topological network scoring metrics and their effectiveness in identifying cancer-related biomarkers in biological networks. Compared to previous studies that solely relied on popular metrics like Degree[15–17], selected metrics without any validation or reasoning provided, or simply used all 12 metrics together[6,25], which was shown to be inefficient in this study, our study thoroughly evaluates all 12 metrics using our proposed Integrated AUC. Moreover, many previous studies only used one metric or lacked effective metric composition, which is not enough to accurately quantify and characterize the disease network. Our study advances upon these studies by evaluating and validating different metric compositions to identify the most effective one for biological networks: the C-Index.

The two metrics that compose C-Index, Clustering Coefficient and Bottleneck, effectively capture local and global gene interactions in the network. We hypothesize that Clustering Coefficient identifies significant communities that are most likely involved in pathways to promote tumorigenesis. Bottlenecks are critical points in the network that connect the biological pathways identified through Clustering, and may act as key signaling molecules. These two metrics work in tandem to effectively identify genes that work together in pathways to induce cancer. Additionally, we hypothesize that Degree, which only considers the immediate neighborhood of nodes[27], may not be a robust indicator of network topology as it fails to adequately capture the connectedness and centrality of nodes within the network.

This study largely focused on biological implications. Through functional enrichment analysis, we found that the four genes in the 4-gene signature are 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 down-regulation of these four anti-tumor genes in NSCLC patients signifies that a change in their expressions disrupts the tumor microenvironment, promoting tumorigenesis. These results match the significant GO and KEGG pathways identified in the first part of work.

In particular, AGER exhibits significant enrichment in immune signaling pathways. AGER has been found to be downregulated in NSCLC cancer cells, and its overexpression has been shown to suppress cancer cell proliferation, invasion, and migration, while promoting apoptosis[26]. Therefore, its downregulation decreases the effectiveness of these inhibitory effects. Similarly, CA4, which affects the cell cycle and inhibits cell proliferation by downregulating the expression of CDK2, was found to be downregulated in NSCLC cancer cells[31]. Its downregulation prevents the inhibition of tumor cell proliferation. The downregulation of RASIP1, which is involved in GTPase binding and cell-cell attachment, was found to impair cell-cell attachment and possibly promote cell migration in NSCLC patients[32]. Finally, CAV1, a scaffolding protein that may act as both a tumor suppressor and a promoter of metastasis, depending on the type of cancer and stage[33], has been found to be downregulated in many tumors, including NSCLC[34]. The significant biomarkers discovered may be further explored through knock-out trials and analysis of their impacts on cancer prognosis. Results from these trials can be applied to developing drug therapies that target these biomarkers.

This study has shown that using multiple biomarkers and metrics concurrently greatly improves performance. This is because genes work together in pathways that lead to tumorigenesis, and single genes cannot cause cancer without having numerous regulatory effects on other genes through signal transduction pathways. Cancer is a complex disease caused by the interaction of multiple environmental factors and genes. It is the combined effect of all these genes in the pathway together that leads to cancer onset. The 4-gene biomarker signature and biomarkers selected by C-index accurately capture the nature of cancer. With further validation and refinement in other cancer datasets, they are promising for the study of biomarkers and all cancers. The C-Index greatly increases the efficiency and accuracy of future biomarker searches, allowing for the low-cost identification of biomarkers with great diagnostic capability. The findings from this study provide an experimental foundation for further exploration of the usage of PPI networks to diagnose cancers. Most importantly, our results indicate that the conventional method of approaching TDA in oncology is greatly ineffective.

Topological analysis is a powerful method to analyze biomedical data. Our work lends itself to further exploration in settings other than biomarkers. Possible future directions include predicting treatment responses, cellular architecture determination, tumor segmentation, and other applications of cancer data. Our research can be expanded to other types of cancers and more datasets, and our gene signature and C-Index need to be further extensively validated. The proposed methodology of finding top metrics can be extended to effectively and efficiently select biomarkers in various types of cancers, not just NSCLC, which helps to fundamentally advance the topological network research for the continuous pursuit of cancer prevention.

Methods

Stage I: Biomarker signature identification. *Dataset.* In total, we analyzed 547 NSCLC samples consisting of 467 lung tumor samples and 80 normal lung samples from the Gene Expression Omnibus (GEO) database[19], a national repository of genetic information databases. In this study, we retrieved and combined three gene expression profiles (GSE31210, GSE33356, and GSE50081) to ensure greater accuracy and comprehensiveness. The datasets are summarized in Supplementary Table S1. Finally, the merged dataset was randomly divided into 80% for identification of top biomarkers and metrics, and 20% for validation.

Data preprocessing and DEG identification. GEO2R analysis was performed to detect DEGs in NSCLC tumor samples compared with normal lung samples. An initial pool of 267 statistically significant DEGs (p-value < 0.01 and $|\text{logFC}| > 1.2$) were identified for further analysis and classified as up or down-regulated. The raw gene expression values are normalized using a z-score.

Functional enrichment analysis. We performed functional enrichment analyses using the Database for Annotation, Visualization and Integrated Discovery (DAVID) gene functional annotation tool[35,36] to identify significant Gene Ontology (GO) terms with FDR < 0.05. GO terms are biological annotations that signify functional characteristics. They are divided into three main categories: molecular function (MF), cell composition (CC), and biological processes (BP). The most significant GO terms were analyzed using DAVID to identify enriched terms with a threshold value of FDR (adjusted p-value) < 0.05. Similar in function to the adjusted-p-value, the lower the FDR, the more significant the enrichment. Statistically significant GO terms were also expressed as a z-score expression

$$\text{z-score} = \frac{N_{upregulated} - N_{downregulated}}{\sqrt{count}} \tag{2}$$

where $N_{upregulated}$ and $N_{downregulated}$ represent the number of upregulated and downregulated genes respectively. This expression value signifies whether the GO term is more likely to be downregulated (negative value) or upregulated (positive value). We visualized the top 10 GO terms and their z-score expressions using the GOplot package (version 1.0.2) in R[37].

String-db[38] analysis was also utilized to identify significant Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The KEGG database contains genomic information, functions, recognized pathways, and networks with higher-order functional information of various organisms[39].

PPI network construction. We constructed a protein-protein interaction (PPI) network based on the DEGs using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) Interactome database[38]. It is a biological database and web resource of known and predicted protein-protein interactions. The network was then input into *Cytoscape*[22] for further analysis with the CytoHubba[23] plugin. Using all 12 topological analysis methods, we identified our list of candidate biomarkers composed of the top genes selected by each method.

AUC performance evaluation and Integrated AUC. We evaluated the diagnostic value and functional significance of the biomarkers using AUC, which measures the trade-off between sensitivity and specificity. AUC was calculated for each gene using the pROC package in R Studio[40], which was also used to visualize the ROC curves. Similarly, to evaluate the performance of multiple biomarkers at once, Integrated AUC was calculated by first aggregating (in this study, we use mean) the expression values of the biomarkers, and then evaluating the AUC of that aggregated expression. To identify our 4-gene signature, we systematically aggregated the top biomarker expression values one by one and determined when the AUC reaches its peak.

Survival analysis and validation of biomarkers. The Kaplan–Meier plotter database (http://kmplot.com) is an online tool used to analyze the associations between the identified hub genes and overall survival. The overall survival (OS) plots were based on 1925 lung cancer patients from GEO and TGCA (The Cancer Genome Atlas database) datasets. For the RASIP1 gene, however, the dataset used for analysis was restricted to only samples that were tested with the HGU133 Plus 2.0 microarray, as it was the only probe set available for RASIP1. Therefore, the Kaplan-Meier for RASIP1 was limited to 1144 patients. The hazard ratio (HR) with 95% confidence intervals and log-rank p-value were calculated, and a p-value < 0.05 was used to indicate a statistically significant difference.

The expression levels of the hub genes and their association with survival were additionally assessed using the web-based GEPIA database (http://gepia.cancer-pku.cn/)[41] with the settings of $p < 0.05$ and $|\text{Log2FC}| > 1$.

Stage II: Critical network topological metrics. *Evaluation of composite metrics.* To evaluate composite metrics, we first fused, or supserset, the genes identified through each individual metric. Then, Integrated AUC of the top 10 and 20 genes of the superset was evaluated to ensure a fair assessment of each composition and to eliminate any bias surrounding using more or less number of biomarkers.

In addition to AUC, to quantify the ability of each metric to correctly identify the biomarkers that were included in the overall top 10 and top 20 ranked biomarkers found in Table 1, we also define a Precision metric as follows:

$$\text{Precision} = \frac{N_{correct}}{N_{total}}, \tag{3}$$

where $N_{correct}$ and N_{total} represent the number of correctly identified genes and total number of top ranked genes respectively.

To find the number of correctly identified genes ($N_{correct}$), we took the top 10 and 20 disease genes from the supersets of the genes selected by the composite metrics and calculated how many of them matched with the overall top 10 and 20 disease-correlated genes from Table 1b. These overall top 10 and 20 genes were identified by AUC score in the "PPI network analysis" section.

Machine learning validation and Cascading Model design. The primary focus of this study was to improve the diagnostic capability of early cancer detection. To validate our findings, we evaluated the diagnostic performances of the metrics and signatures with clinicogenomic random forest (RF) models in the 20% dataset. Among various classification models, including artificial neural networks, XGBoost models, SVM, and decision tree models, the RF models exhibited the best performance. RF was additionally found to have advantages over other classification algorithms in terms of robustness to overfitting, ability to handle non-linear data, and stability in the presence of outliers, as previously reported[42].

Our work above focused on the performance of biomarkers 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. Some clinical attributes that impact the development of NSCLC are age, gender, and smoking status. In order to incorporate both clinical and genomic attributes, we explored the use of machine learning techniques to adequately consider multiple factors at once in cancer prediction to design a clinicogenomic model.

The 4-gene and C-Index models were further validated in an external cohort, the TCGA LUAD and LUSC datasets. Phenotype and gene count files were obtained from the UCSC Xena database, and underwent similar preprocessing and normalization procedures as the GEO datasets. Subsequently, the pretrained models were assessed using this new dataset.

In addition to validating our models, we sought to expand the 4-gene model into a Cascading Model that not only leverages the top biomarkers for accurate prediction of cancer status, but also has the capability to differentiate between early and late stages of cancer and lung cancer stages I–IV. Our goal is to develop a precise clinicogenomic diagnostic model that can utilize the selected top biomarkers to accurately predict NSCLC disease stage. In our initial study, we extended the RF model to a multi-class model, which as expected, exhibited high accuracy in classifying cancer from non-cancer cases. However, the multi-class model showed lower accuracy in classifying cancer stages, likely due to limited data availability for stages III and IV.

To more accurately identify the stage of cancer, we propose a multi-stage Cascading Model with 3 stages, depicted in Supplementary Fig. S2. The model evaluates the diagnostic capability of the 4-gene model. It first

classifies the data into cancer or no cancer, then further classifies those with cancer into early vs. late stages of cancer, and finally classifies early and late one step further into cancer stages I–IV. In the 4-gene model, the first classification cancer status had an average accuracy of 0.9553 and AUC of 0.98605. The second classification, early vs. late stages, had an average accuracy of 0.9716 and AUC of 0.9902. Cancer stages I–IV classification had an average accuracy of 0.7137, which may be due to the minor difference between the four cancer stages of patients. The difference of clinical attributes between cancer and no cancer and early vs. late is a lot greater than the difference between the four stages. The model may have difficulty distinguishing the two sides of the boundary. However, as exemplified in the performance study, the Cascading model greatly improves the ability to accurately differentiate between multiple stages of cancer, and is one of the first capable of accurately predicting early vs. late cancer stages.

Data availibility

The datasets GSE31210, GSE33356, and GSE50081 are available online from the GEO database.

Code availability

https://github.com/iwu24/NSCLCcascadingmodel.git.

Received: 17 October 2022; Accepted: 13 May 2023
Published online: 22 May 2023

References

1. Lu, T. *et al.* Trends in the incidence, treatment, and survival of patients with lung cancer in the last four decades. *Cancer Manag. Res.* **11**, 943–953. https://doi.org/10.2147/CMAR.S187317 (2019).
2. Gridelli, C. *et al.* Non-small-cell lung cancer. *Nat. Rev. Dis. Primers* **1**, 15009 (2015).
3. Zappa, C. & Mousa, S. A. Non-small cell lung cancer: Current treatment and future advances. *Transl. Lung Cancer Res.* **5**, 288–300 (2016).
4. Sawyers, C. L. The cancer biomarker problem. *Nature* **452**, 548–552 (2008).
5. Villalobos, P. & Wistuba, I. I. Lung cancer biomarkers. *Hematol. Oncol. Clin. North Am.* **31**, 13–29 (2017).
6. Maharjan, M., Tanvir, R. B., Chowdhury, K., Duan, W. & Mondal, A. M. Computational identification of biomarker genes for lung cancer considering treatment and non-treatment studies. *BMC Bioinform.* **21**, 218 (2020).
7. Masoomy, H., Askari, B., Tajik, S., Rizi, A. K. & Jafari, G. R. Topological analysis of interaction patterns in cancer-specific gene regulatory network: Persistent homology approach. *Sci. Rep.* **11**, 16414 (2021).
8. Lum, P. Y. *et al.* Extracting insights from the shape of complex data using topology. *Sci. Rep.* **3**, 1236 (2013).
9. Miryala, S. K., Anbarasu, A. & Ramaiah, S. Discerning molecular interactions: A comprehensive review on biomolecular interaction databases and network analysis tools. *Gene* **642**, 84–94 (2018).
10. Rabadán, R. *et al.* Identification of relevant genetic alterations in cancer using topological data analysis. *Nat. Commun.* **11**, 3808 (2020).
11. Loughrey, C., Fitzpatrick, P., Orr, N. & Jurek-Loughrey, A. The topology of data: Opportunities for cancer research. *Bioinformatics* **37**, 3091–3098 (2021).
12. Winterbach, W., Van Mieghem, P., Reinders, M., Wang, H. & de Ridder, D. Topology of molecular interaction networks. *BMC Syst. Biol.* **7**, 90 (2013).
13. Ni, M. *et al.* Identification of candidate biomarkers correlated with the pathogenesis and prognosis of non-small cell lung cancer via integrated bioinformatics analysis. *Front. Genet.*https://doi.org/10.3389/fgene.2018.00469 *(2018).*
14. Li, Z. *et al.* Identification of key biomarkers and potential molecular mechanisms in lung cancer by bioinformatics analysis. *Oncol. Lett.* **18**, 4429–4440 (2019).
15. Islam, R. *et al.* Identification of molecular biomarkers and pathways of NSCLC: Insights from a systems biomedicine perspective. *J. Genet. Eng. Biotechnol.* **19**, 43 (2021).
16. Tu, H., Wu, M., Huang, W. & Wang, L. Screening of potential biomarkers and their predictive value in early stage non-small cell lung cancer: A bioinformatics analysis. *Transl. Lung Cancer Res.* **8**(6), 797–807 (2019).
17. Zhu, Y. *et al.* Identification of potential circular RNA biomarkers in lung adenocarcinoma: A bioinformatics analysis and retrospective clinical study. *Oncol. Lett.* **23**, 144 (2022).
18. Henry, N. L. & Hayes, D. F. Cancer biomarkers. *Mol. Oncol.* **6**, 140–146 (2012).
19. Edgar, R., Domrachev, M. & Lash, A. E. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
20. Henke, E., Nandigama, R. & Ergün, S. Extracellular matrix in the tumor microenvironment and its impact on cancer therapy. *Front. Mol. Biosci.* **6**, 160 (2019).
21. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
22. Gustavsen, J.A. et al. Rcy3: Network biology using cytoscape from within r. F1000Research. https://doi.org/10.12688/f1000research.20887.3 (2019)
23. Chin, C.-H. *et al.* cytohubba: Identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* **8**(Suppl 4), S11 (2014).
24. Sun, C., Yuan, Q., Wu, D., Meng, X. & Wang, B. Identification of core genes and outcome in gastric cancer using bioinformatics analysis. *Oncotarget* **8**, 70271–70280 (2017).
25. Wang, L. *et al.* Identification and validation of key genes with prognostic value in non-small-cell lung cancer via integrated bioinformatics analysis. *Thorac. Cancer* **11**, 851–866 (2020).
26. Wang, Q. *et al.* Effect of AGER on the biological behavior of non-small cell lung cancer H1299 cells. *Mol. Med. Rep.* **22**, 810–818 (2020).
27. Wang, M., Wang, H. & Zheng, H. A mini review of node centrality metrics in biological networks. *Int. J. Netw. Dynam. Intell.* **1**(1), 99–110 (2022).
28. Lu, M. *et al.* Identification of significant genes as prognostic markers and potential tumor suppressors in lung adenocarcinoma via bioinformatical analysis. *BMC Cancer* **21**, 616 (2021).
29. Maharjan, M., Tanvir, R., Chowdhury, K. & Mondal, A. Determination of biomarkers for diagnosis of lung cancer using cytoscape-based GO and pathway analysis. in *Proceedings of the International Conference* (Athens, 2019).
30. Goldman, M. J. *et al.* Visualizing and interpreting cancer genomics data via the xena platform. *Nat. Biotechnol.* **38**, 675–678 (2020).
31. Xu, B. *et al. Carbonic anhydrase 4 serves as a novel prognostic biomarker and therapeutic target for non-small cell lung cancer: A study based on TCGA samples* (Comb. Chem, High Throughput Screen, 2023).

32. Chen, Y. *et al.* Rasip1 is a RUNX1 target gene and promotes migration of NSCLC cells. *Cancer Manag. Res.* **10**, 4537–4552 (2018).
33. Díaz, M. I. *et al.* Caveolin-1 suppresses tumor formation through the inhibition of the unfolded protein response. *Cell Death Dis.* **11**, 648 (2020).
34. Shi, Y.-B. *et al.* Multifaceted roles of caveolin-1 in lung cancer: A new investigation focused on tumor occurrence, development and therapy. *Cancers (Basel).* **12**, 291 (2020).
35. Sherman, B. T. *et al.* DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**(W1), W216–W221 (2022).
36. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
37. Walter, W., Sánchez-Cabo, F. & Ricote, M. GOplot: An R package for visually combining expression data with functional analysis. *Bioinformatics* **31**, 2912–2914 (2015).
38. Szklarczyk, D. *et al.* The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).
39. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
40. Robin, X. *et al.* pROC: An open-source package for R and s+ to analyze and compare ROC curves. *BMC Bioinform.* **12**, 77 (2011).
41. Tang, Z. *et al.* GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* **45**, W98–W102 (2017).
42. Sarica, A., Cerasa, A. & Quattrone, A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review. *Front. Aging Neurosci.* **9**, 329 (2017).

Acknowledgements
We would like to thank Siya Goel from Stanford University for many helpful discussions throughout this work.

Author contributions
I.W. conceived and conducted the experiment and analyzed the data. I.W. and X.W. wrote the paper. All authors reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-35165-w.

Correspondence and requests for materials should be addressed to I.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.