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

Motivation and Problems

Lung cancer is the deadliest cancer worldwide, accounting for about 25% - the highest - of all cancer deaths. The five year survival rate is only 18.6%. 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. Biomarkers provide insights into the molecular origins and behaviors of NSCLC, enabling early diagnosis and the identification of high-risk patients for personalized medicine.

In the search for biomarkers, topological network analysis is one of the most powerful and widespread tools in the field to analyze interactions between proteins and genes and identify essential biomarkers in the network. The focus of this research, topological scoring methods, score and rank nodes (genes) by different network features to output the highest scoring nodes. These high scoring nodes are the most essential genes in the network, and are further analyzed as potential biomarkers. However, existing studies suffer from low performance with the following limitations:

- Fail to fully consider the biological significance of the quantitative network methods used.
- Utilize popular scoring metrics without verification. The most popular metric, degree of freedom, was found to be one of the lowest performing metrics in this study.
- Limit to the use of a single metric, without being able to capture the intricate interactions among different genes and their impact on cancer prediction.

Purpose and Hypothesis

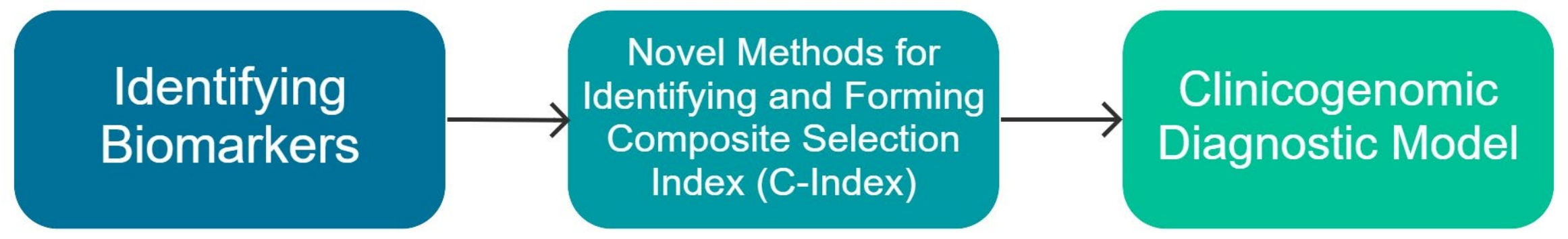
Objectives (2 main tasks):

- Fundamentally advance the topological network research with proposed methods to identify novel network scoring metrics with biological significance to guide the efficient search of critical biomarkers.
- Establish a learning-based clinicogenomic model that can comprehensively consider critical gene signatures and clinical covariates to accurately diagnose NSCLC.

Hypothesis:

- A good scoring metric can guide the efficient search of bio-significant biomarkers for more accurate diagnosis. The current most conventionally used metric, degree, doesn't necessarily find the best biomarkers.
- The use of multiple biomarkers simultaneously to predict disease can increase the performance, but too many biomarkers can result in decreased performance if the variance is too large. Thus, the ideal number of biomarkers and metrics to ensure high performance and low complexity must be found.

Methods



Identifying Biomarkers → Novel Methods for Identifying Biomarkers and Forming Composite Selection Index (C-Index) → Clinicogenomic Diagnostic Model

In this work, novel computational methods are developed to empower the topological network analysis for the efficient finding of critical biomarkers, and accurately diagnose NSCLC with a clinicogenomic model. My work is divided into three stages:

- Identify significant differentially-expressed-genes (DEGs), or biomarkers **(4-gene biomarker signature),** using functional enrichment analysis, complex network methods, and AUC to robustly validate the diagnostic performance of biomarkers. AUC evaluates the ability of the biomarkers to classify between disease and control. A diverse range of network topological score metrics were used to identify the most comprehensive set of biomarkers.
- Develop two novel techniques to advance network analysis:
  - Introduce a systematic method to identify the top topological metrics in protein networks that are essential for biomarker selection. Integrated AUC and aggregate gene expression were introduced and applied to guide the prediction of cancer with multiple biomarkers.
  - Proposing a novel composite selection index (***C-index***) to concurrently consider complementary factors in biomarker selection for higher diagnostic performance
- Explore a clinicogenomic machine learning model with top biomarkers selected and clinical covariates to accurately diagnose NSCLC

GO term enrichment analysis of DEGs

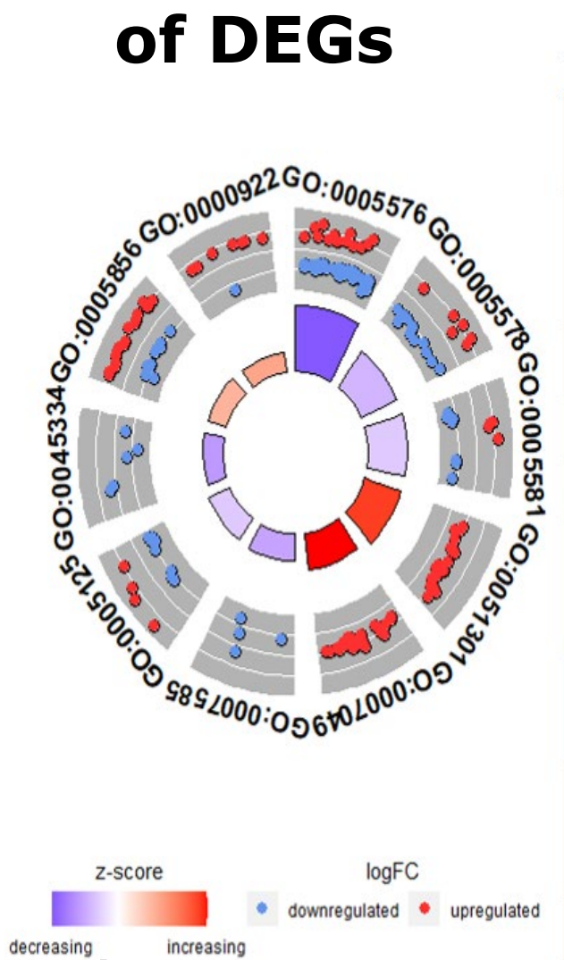


Fig. 1

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

🟩 BP 🟧 CC 🟦 MF

In the first stage, functional enrichment analysis reveals that the DEGs are significantly enriched in the cell cycle and extracellular receptor and signaling, which play critical roles in induced cancer progression and development, and disrupted cell progression (Fig 1). The top DEGs were then identified through topological network analysis and AUC performance evaluation.

Complex Network of DEGs

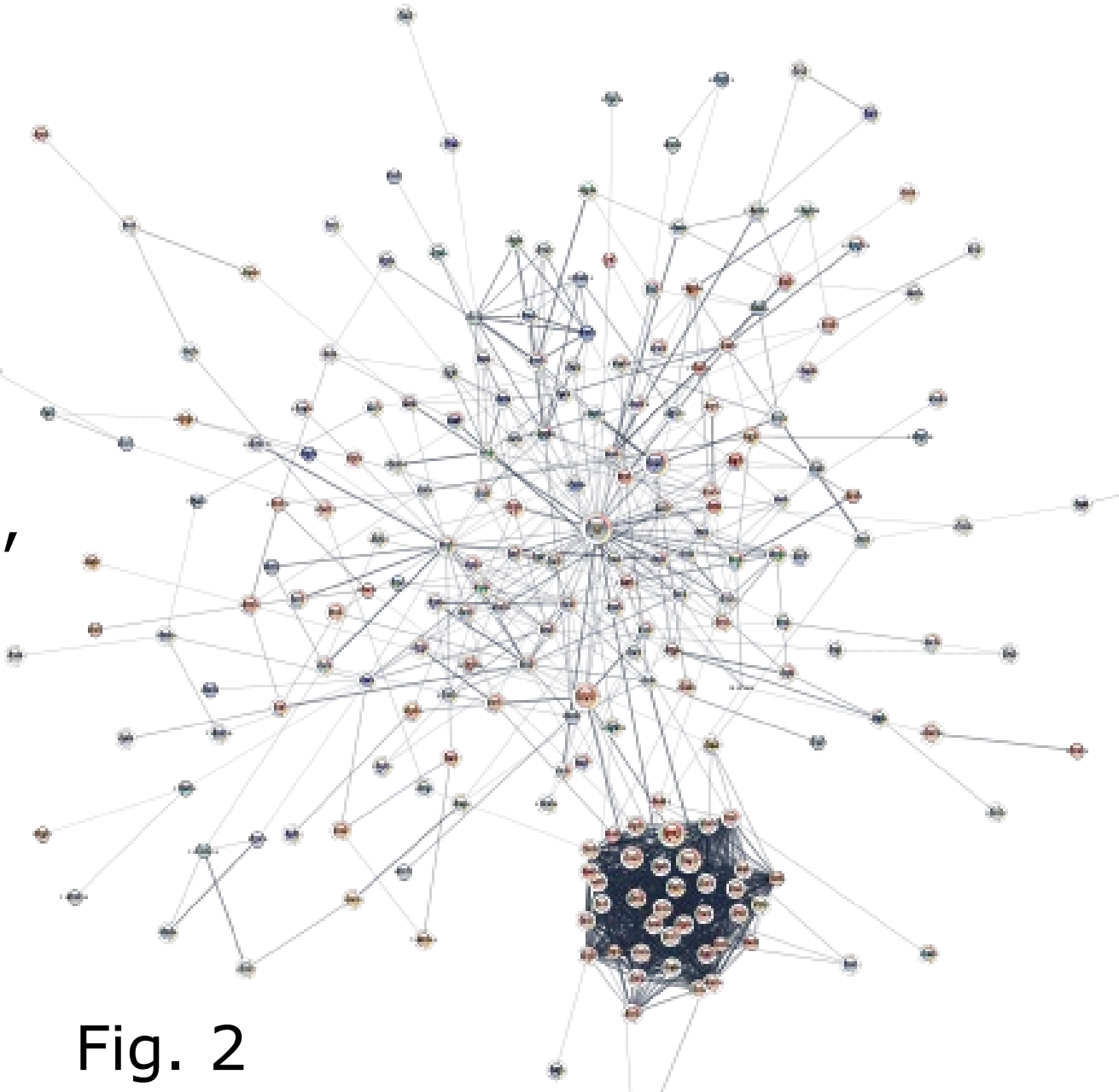


Fig. 2

In the second stage, topological scoring methods were evaluated in the gene network (Fig. 2) with a novel integrated AUC, which evaluates the performance of a group of genes used together for cancer prediction, rather than individually. Utilizing multiple metrics concurrently was explored, and the discovered metrics were compared with conventional metrics to evaluate improvement.

In the final stage, the most important NSCLC clinical attributes (age, sex, and smoking status), were incorporated with the top biomarker gene signature to create a clinicogenomic NSCLC diagnosis model with high accuracy. A Random Forest (RF) algorithm was used due to its robustness to overfitting and ability to handle non-linear data. The dataset was split into an 80% training set and 20% validation set, and was fitted to the training set data to predict whether a patient has NSCLC.

Results

4-gene Biomarker Signature

The 4-gene biomarker signature is the best performing combination, with a 92% AUC. It is composed of the top 4 biomarkers in the dataset: *AGER, CA4, RASIP1,* and *CAV1.* As shown in *Fig. 3* that compares gene expression between NSCLC and control patients, low expression of these four genes is associated with NSCLC. All four genes play a significant role in ECM receptor and signaling, and under regulation of them leads to defected protection against the spread of cancer.

Comparison of gene expression between NSCLC and control patients

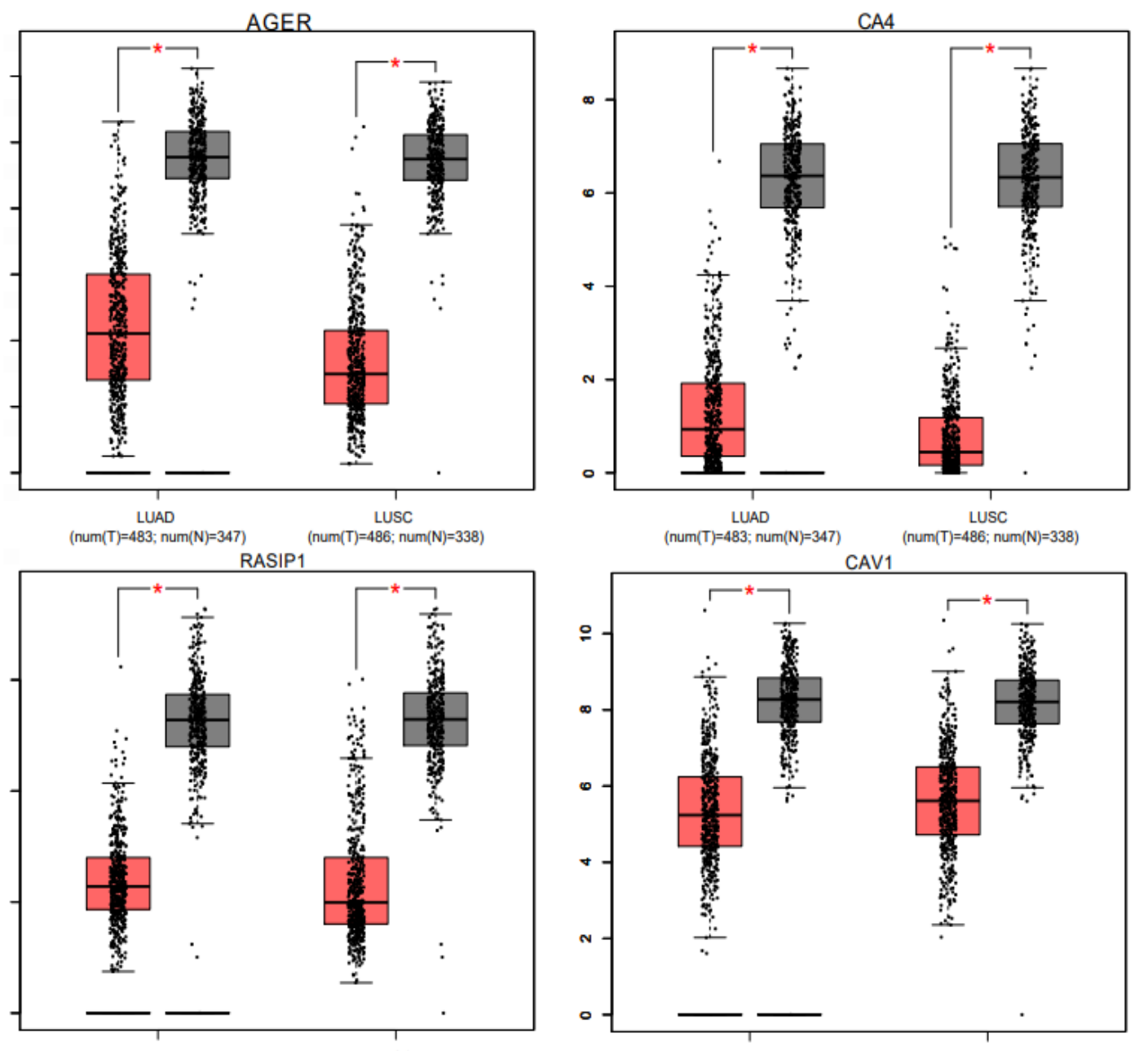


Fig. 3 **(red = tumor, gray = control)**

C-Index

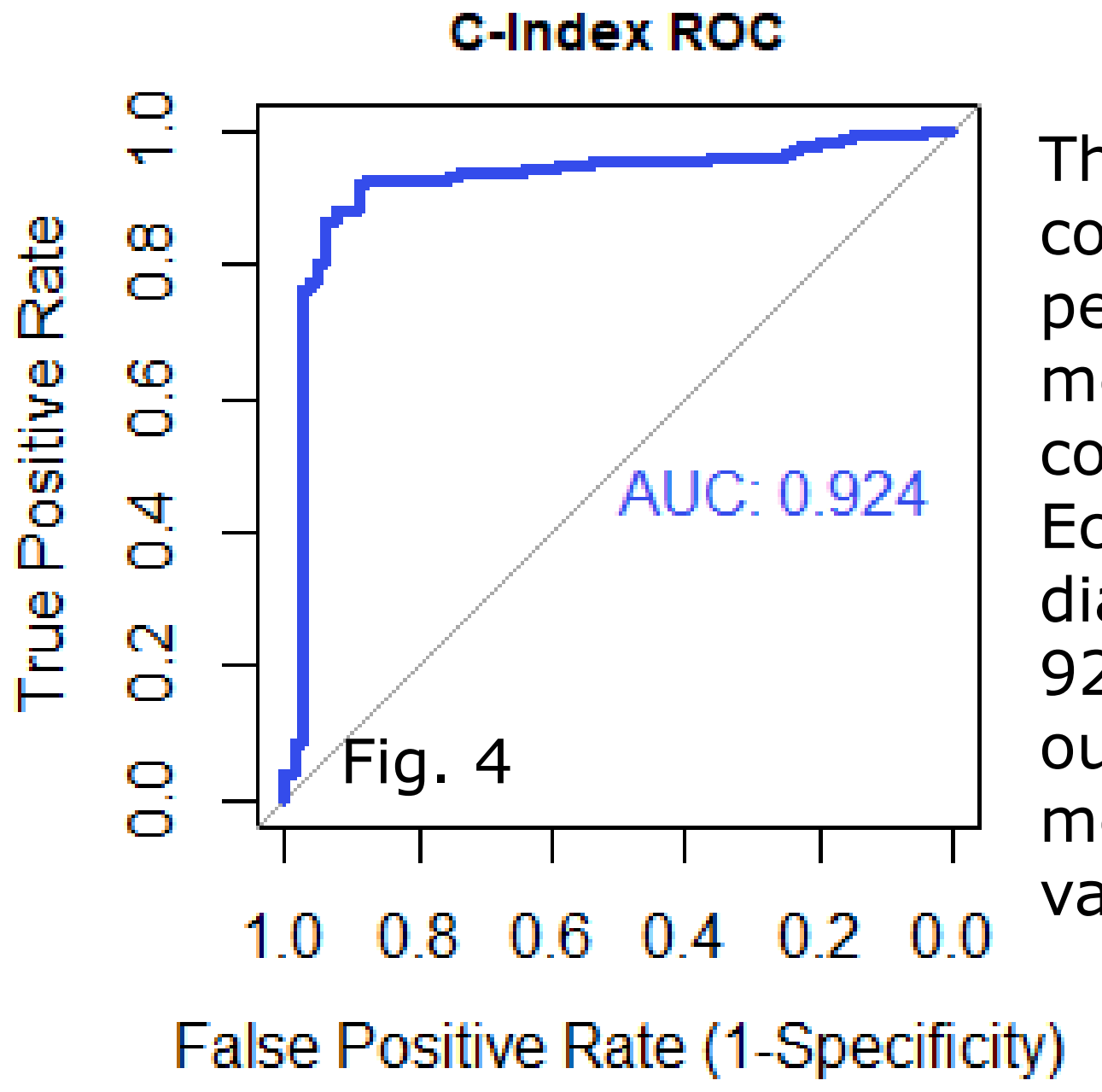


Fig. 4

The proposed C-Index is composed of the top performing topological metrics: Clustering coefficient, Bottleneck, and Eccentricity. With a diagnostic performance of 92% (Fig. 4), it greatly outperformed conventional metrics (degree) by 20%, a vast improvement.

Clinicogenomic Diagnostic Model

The clinicogenomic diagnostic model had an incredibly high performance: the 95% accuracy and confusion matrix (Fig. 5) and 96% AUC (Fig. 6) show that the model performs at high rates and is accurate in all aspects when diagnosing NSCLC. The high sensitivity (True Positives, 0.96) and specificity (False Negatives, 0.8) demonstrate that the model rarely misdiagnoses patients.

Model Confusion Matrix

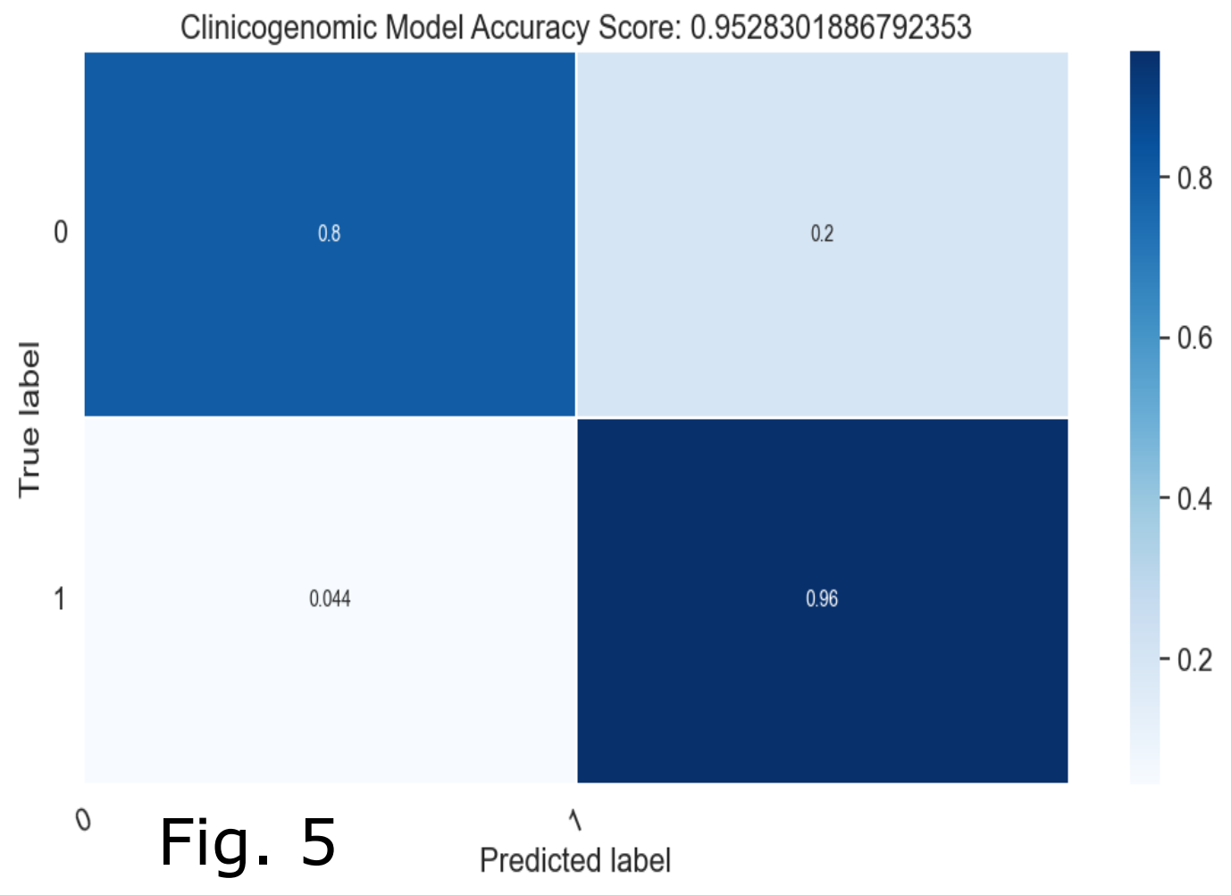


Fig. 5

Model ROC

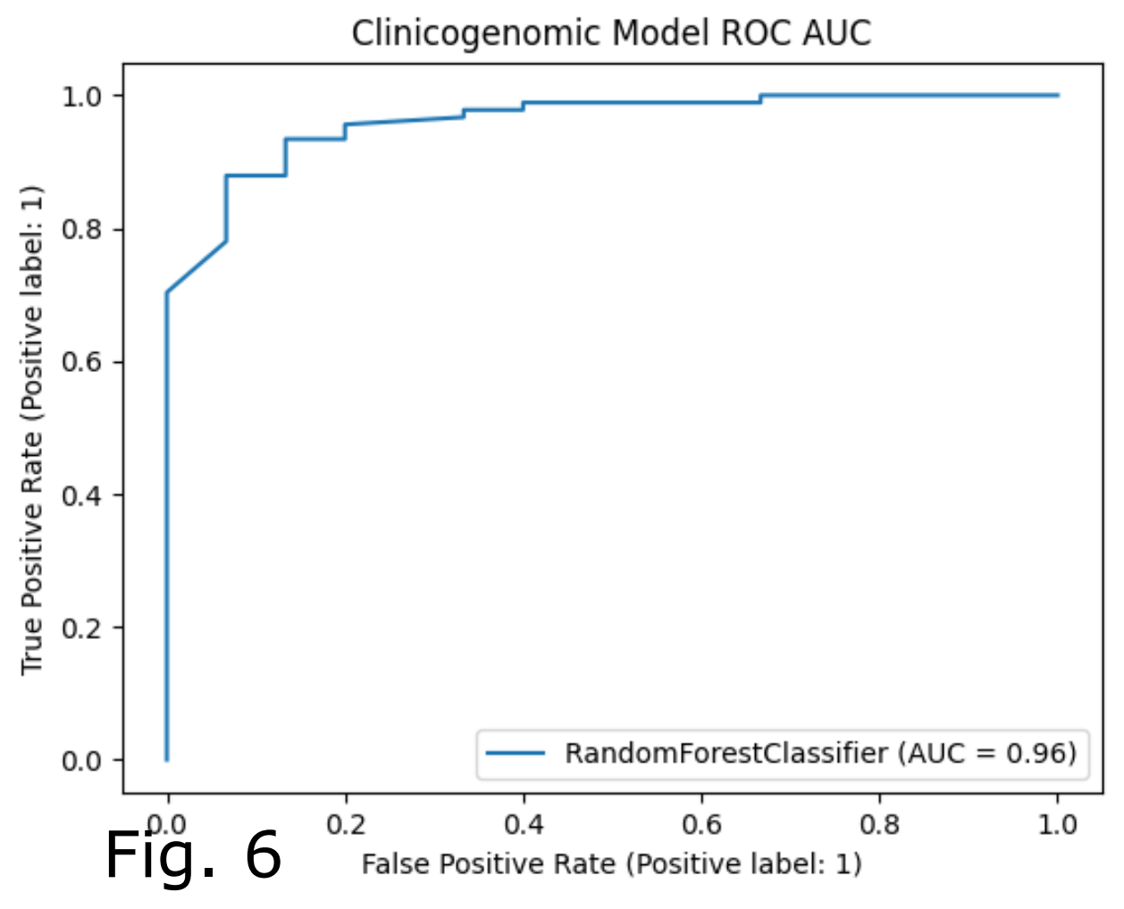


Fig. 6

Discussion

We have the following observations:

- A **4-gene biomarker signature** consisting of AGER, CA4, RASIP1, and CAV1 achieves the highest performance. These four genes can be further explored as possible therapeutical targets for future drug treatment.
- A novel composite index (**C-index**) that combines clustering coefficient, bottleneck, and eccentricity achieves the strongest performance. In protein networks, the clustering coefficient identifies significant communities that work together to induce cancer, bottlenecks are key connectors that participate in several regulatory pathways to control the interactions of a large number of proteins, and the eccentricity evaluates the ease of a gene to be functionally reached by all other genes in a network.
- 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. A 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.

Conclusion

A set of studies have been conducted to advance the topological network research for the early diagnosis of non-small cell lung cancer, fully meeting and exceeding the objectives and hypothesis:

- Performing comprehensive analysis of a merged dataset to identify a 4-gene biomarker signature that achieves a 92% AUC diagnostic performance
- Introducing a novel and systematic method to identify the top topological metric in protein networks that is essential for biomarker selection
- Proposing a novel composite selection index (C-index) that can concurrently consider complementary factors in biomarker selection to greatly increase accuracy of NSCLC diagnosis, with the AUC increased from 75% to 92% under low complexity biomarker search.
- Exploiting the use of a clinicogenomic machine learning model with top biomarkers selected and clinical covariates to greatly increase the accuracy of NSCLC diagnosis, with 95% accuracy and 96% AUC.

The results demonstrate the effectiveness of the proposed method in finding critical biomarkers for accurate yet low-complexity diagnosis of NSCLC.

Future Work

The proposed methodology of finding top score metrics is effective and general. It is not restricted to the use in NSCLC diagnosis, but can be extended to early diagnose other types of cancers. The methodology along with the novel composition of complementary score metrics help to fundamentally advance the topological network research, and can enable highly accurate yet efficient and low cost search of biomarkers that are critical for the early cancer diagnosis and prevention. As part of the future work, the proposed methods and score metrics identified will be tested over more datasets and applied to the diagnoses of a variety of cancers.