Integrative network modeling highlights the crucial roles of Rho-GDI signaling pathway in the progression of non-small cell lung cancer

Saransh Gupta†, Haswanth Vundavilli†, Rodolfo S. Allendes Osorio, Mari N. Itoh, Attayeb Mohsen, Aniruddha Datta, Kenji Mizuguchi, Lokesh P. Tripathi*

***Abstract*— Non-small cell lung cancer (NSCLC) is the most prevalent form of lung cancer and a leading cause of cancer-related deaths worldwide. Using an integrative approach, we analyzed a publicly available merged NSCLC transcriptome dataset using machine learning, protein-protein interaction (PPI) networks and bayesian modeling to pinpoint key cellular factors and pathways likely to be involved with the onset and progression of NSCLC. First, we generated multiple prediction models using various machine learning classifiers to classify NSCLC and healthy cohorts. Our models achieved prediction accuracies ranging from 0.83 to 1.0, with XGBoost emerging as the best performer. Next, using functional enrichment analysis (and gene co-expression network analysis with WGCNA) of the machine learning feature-selected genes, we determined that genes involved in Rho GTPase signaling that modulate actin stability and cytoskeleton were likely to be crucial in NSCLC. We further assembled a PPI network for the feature-selected genes that was partitioned using Markov clustering to detect protein complexes functionally relevant to NSCLC. Finally, we modeled the perturbations in RhoGDI signaling using a bayesian network; our simulations suggest that aberrations in *ARHGEF19* and/or *RAC2* gene activities contributed to impaired MAPK signaling and disrupted actin and cytoskeleton organization and were arguably key contributors to the onset of tumorigenesis in NSCLC. We hypothesize that targeted measures to restore aberrant *ARHGEF19* and/or *RAC2* functions could conceivably rescue the cancerous phenotype in NSCLC. Our findings offer promising avenues for early predictive biomarker discovery, targeted therapeutic intervention and improved clinical outcomes in NSCLC.**

***Index Terms*— Systems Biology; Machine Learning; Lung Cancer; Bayesian Modeling; RhoGDI pathway; PPI networks**

†These authors contributed equally.

Saransh Gupta is with Indian Institute of Technology Kharagpur, West Bengal, India.

Haswanth Vundavilli and Aniruddha Datta are with the Department of Electrical and Computer Engineering & The Center for Bioinformatics and Genomic Systems Engineering, Texas A&M University, College Station, TX, USA.

Kenji Mizuguchi is with Laboratory of Bioinformatics, Artificial Intelligence Center for Health and Biomedical Research (ArCHER), National Institutes of Biomedical Innovation, Health and Nutrition, 7-6-8 Saito-Asagi, Ibaraki, Osaka, Japan and Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka, Japan.

Rodolfo S. Allendes Osorio, Mari N. Itoh and Attayeb Mohsen are with Laboratory of Bioinformatics, Artificial Intelligence Center for Health and Biomedical Research (ArCHER), National Institutes of Biomedical Innovation, Health and Nutrition, 7-6-8 Saito-Asagi, Ibaraki, Osaka, Japan.

Lokesh P. Tripathi is with Laboratory of Bioinformatics, Artificial Intelligence Center for Health and Biomedical Research (ArCHER), National Institutes of Biomedical Innovation, Health and Nutrition, 7-6-8 Saito-Asagi, Ibaraki, Osaka, Japan and Laboratory of Transcriptome Technology, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, Japan

*Correspondence: lokesh@nibiohn.go.jp; lokesh.tripathi@riken.jp

I. Introduction

Non-small cell lung cancer (NSCLC) represents around 85% of all lung cancers, with a 5-year overall survival rate of about 15%. Until recently, there were unclear guidelines and only a generalised approach for the management of NSCLC. There have been increasing attempts to develop strategies to combat NSCLC that take into account a wide range of aspects such as specific oncogenes, signaling pathways, stage of the disease, and other predictive factors [1]. Key to these strategies is the development of ad-hoc computational approaches that can pinpoint key drug targets and develop specific therapies for individual patients [2].

The rapid advancements in high-throughput omics technologies have led to a proliferation of biological data. Consequently, computational approaches in the domains of machine learning, statistics, and data science are increasingly being used to analyse large biological datasets and solve biological problems [3]. These techniques have paved way for cutting down on the low hit-to-lead ratio and the laborious, time-consuming and expensive experiments. Machine learning methods can help significantly to interrogate high-throughput biological data and generate new hypotheses without solely relying on experiments [4].

Biological pathways are an interconnected series of interactions between different biomolecules that modulate key cellular processes. When the functioning of a key pathway is significantly perturbed by aberrantly behaving oncogenes or tumor-suppressor genes, cellular physiology can be seriously compromised and may potentially lead to cancer [5]. Cancer genes have also been shown to clump together in a few, yet key cellular networks. Consequently, pathway and network analyses are useful tools to analyze and decipher the genes that are likely to drive the onset and progression of the cancerous phenotype. Bayesian networks is one such methodology that has been successful in probing the mechanistics of various cancers, including breast [6], pancreatic [7], and lung cancers [8].

In this study, we have used a merged lung cancer transcriptome dataset to unearth key cellular factors that are likely to strongly correlate with the onset of NSCLC and that can be targeted to potentially steer the out-of-control cancerous cell proliferation back to normalcy. The paper is designed as follows. In section 2, we discuss the methodology used using machine learning models, interpretative phenomenological analysis, and bayesian networks. In section 3, we discuss our results. In section 4, we discuss the significance of our results, offer concluding remarks, and lay out the probable course of action in investigating NSCLC and other diseases in the future.

An overview of the workflow used for this analysis is presented in Figure 1.

II. Materials and Methods

A. Datasets and Feature Selection for Machine Learning

We analyzed a publicly available merged lung transcriptome dataset comprising 1,118 patient-derived samples that included 925 primary NSCLC tumors paired against 193 normal lung tissues from ten independent Gene Expression Omnibus (GEO) microarray datasets [9]. Our analytical pipeline examined 10,077 genes within this dataset as features and disease status (NSCLC/Normal) as the target variables. To select the best contributing features to optimize model complexity, variance, and to avoid over-fitting, we employed Boruta feature selection [10]. Boruta algorithm employs the random forest classifier to estimate the variable importance feature. To minimize randomness and improve accuracy, Boruta generates additional shadow variables by shuffling the values of the original features and subsequently estimating feature importance. Boruta feature selection identifies the necessary features in predicting the target variable and decreases the model complexity at the same time. This procedure was repeated 100 times to impart robustness to the feature selection process.

Following Boruta feature selection, we retained only those features/genes that were also profiled in an independent study that performed next generation sequencing (RNAseq) of non-small cell lung cancer (GEO accession GSE81089) [11]; this RNAseq data was also used as an independent test set to assess model performances. The dataset was then separated into a training and validation set at an 80/20 split.

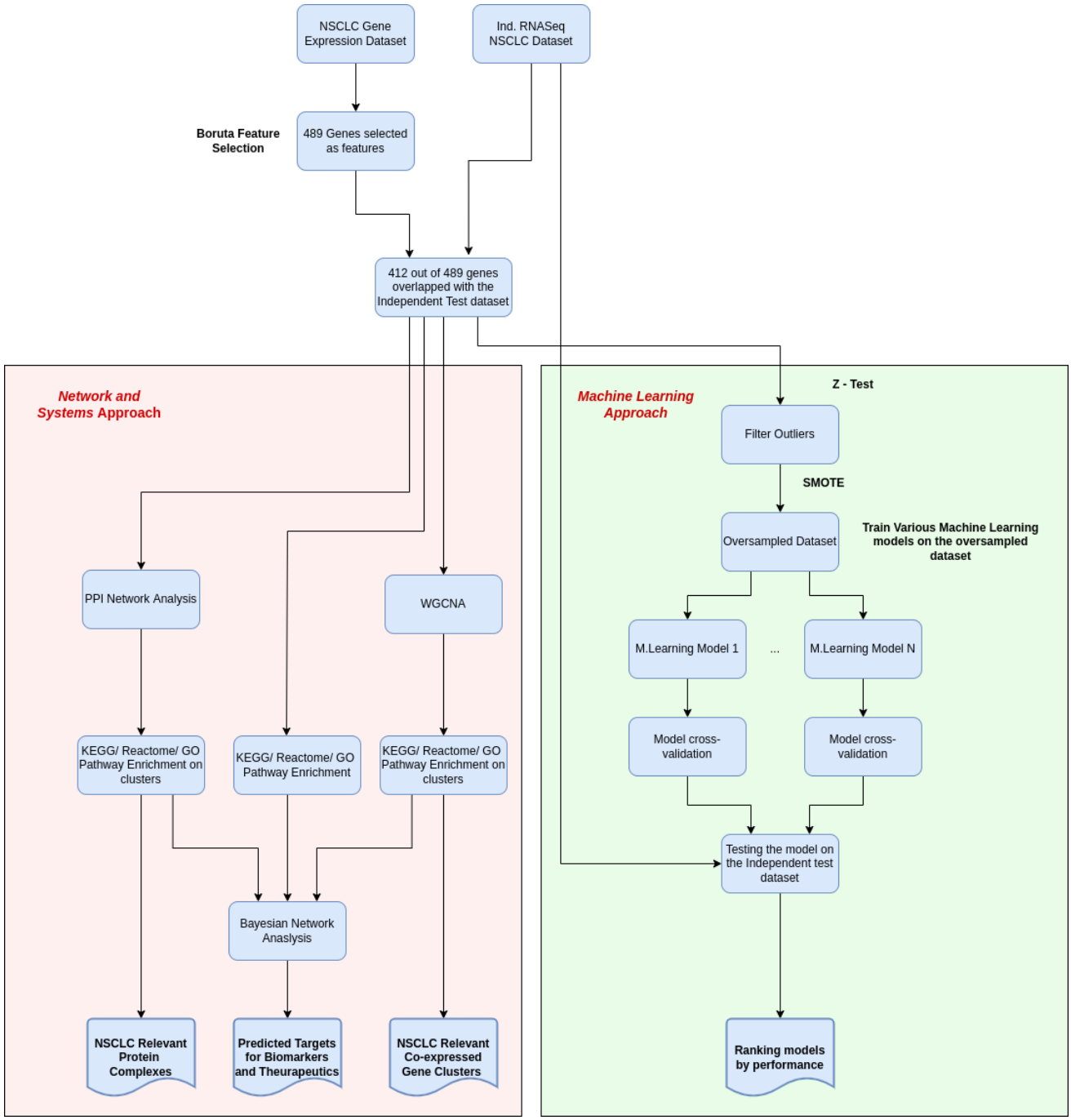


Fig. 1: An overview of the workflow to analyze the merged NSCLC transcriptome

B. Functional enrichment analysis

The selected feature genes were examined for the enrichment of specific biological themes using TargetMine[1] data analysis platform [12] and Ingenuity Pathway Analysis (IPA) platform. The enrichment of specific KEGG pathways [13], Reactome pathways [14], Gene Ontology (GO) associations [15] and IPA pathway associations was estimated using Fisher's exact test. The inferred p-values were further adjusted for multiple-test-correction to control the false-discovery-rate using the Benjamini-Hochberg procedure [16] and the annotations/pathways were judged to be significant if the adjusted p-values ≤ 0.05.

C. Co-expression gene module network analysis

We employed the weighted gene co-expression network analysis (WGCNA) package in R [17] to identify groups of co-expressed genes among the selected features. The blockwise-Module function was used with the default parameters and power = 14 (high correlation); deepSplit = 2 (medium-sensitivity) to allow for reasonably-sized modules of highly correlated genes. The selected genes were examined for functional enrichment analysis with TargetMine [12].

D. Outliers and Over-sampling

We used Z-score estimation to identify the outliers in the dataset [18]. Z-score for each data point was computed using the equation given below, where x is any data point and μ, σ are the mean and standard deviation of the dataset respectively. If $|Z| > 3$, the data point was judged to be an outlier and purged from the dataset.

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

A heavily-skewed distribution in favour of the NSCLC samples in the merged dataset would have significantly impacted the efficacy of the ML algorithms and render classification a challenging task. We therefore, attempted to address this data imbalance by oversampling the minority class, i.e., the 'Normal' class using Synthetic Minority Oversampling Technique (SMOTE) [19]. SMOTE is a well known data augmentation approach that randomly selects a minority class instance p and finds its k nearest minority class neighbors. Next, it selects one of the k nearest neighbors q at random and connects p and q to infer a line segment in the feature space to create a synthetic data point. The synthetic instances are finally generated as a convex combination of the two chosen instances p and q. We used SMOTE to create as many synthetic examples for the minority class as were required to balance the two cohorts. This synthetically augmented dataset was used for the subsequent model building using different ML algorithms.

E. Machine Learning Models

After data processing and feature selection, we employed multiple machine learning classifiers, namely, Logistic Regression [20], Decision Trees Classifiers [21], Support Vector Classifiers [22], CATBOOST [23], Random Forest [24], XGBoost (Booster) [25], and Light Gradient Boosting Machine (LGBM) [26] to construct predictive models. We assessed the performances of the individual classifiers using the area under the Receiver Operating Characteristic curve (AUROC) scores and Cohen's kappa coefficient values.

F. Model evaluation

Model performances were evaluated by testing their prediction performances on the validation set as well as the independent test dataset. We estimated model accuracy and the area under the AUROC score using the scikit-learn package [27] from Python.

[1]https://targetmine.mizuguchilab.org

G. Bayesian network analysis

Interactions among genes are relatively sparse and therefore, bayesian modeling is an attractive approach to model gene regulatory networks [28], [29]. Bayesian networks have been previously deployed in modeling diverse biological pathways such as Peroxisome proliferator-activated receptor pathway, Alzheimer's disease, and diabetes [29]–[31].

A bayesian network is a collection of nodes which interact in a directed acyclic manner. Bayesian network may be represented as $\langle \mathcal{G}, \Phi \rangle$, where $\mathcal{G}$ is a Directed Acyclic Graph (DAG) and Φ is the Conditional Probability Distribution (CPD) of each node. Using the local Markov independence assumption, any Joint Probability Distribution (JPD) that satisfies the Markov condition can be described as a product of CPDs [32]. That is,

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)) \tag{2}$$

where X_i is a random variable and $Pa(X_i)$ is the set of parents of X_i.

1) Discretization: Discretization of gene expression values into a binary framework buffers the model against noise and reduces computational complexities [33]. We used the maximum likelihood estimator for the mean μ as the threshold and discretized the gene expression values into 0s and 1s. Values greater than the threshold were all normalized as '1' and values lesser than the threshold were all normalized as '0'.

2) Prior and Posterior distributions: Next, we applied bayesian modeling to estimate the network parameters. Let Φ_X be the probability that a node X takes the value '1'. For each node, we assumed that Φ_X was Beta distributed. That is,

$$\Phi_X \sim \text{Beta}(\alpha_X, \beta_X) \tag{3}$$

$$\text{Beta}(\Phi_X; \alpha_X, \beta_X) = \frac{\Phi_X^{\alpha_X - 1}(1 - \Phi_X)^{\beta_X - 1}}{B(\alpha_X, \beta_X)}$$

where $B(\alpha_X, \beta_X) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and Γ is the standard Gamma function.

The data likelihood for n observations is defined as:

$$P(X | Pa(X), \Phi_X) \sim B(n, \Phi_X) \tag{4}$$

$$B(i; n, \Phi_X) = \binom{n}{i} \Phi_X^i (1 - \Phi_X)^{n-i}$$

where i is the number of successes ('1s').

Since the binomial likelihood is a conjugate to the beta distribution, the posterior distributions of the variables were defined as:

$$P(\Phi_X | X) \sim Beta(\alpha'_X, \beta'_X) \tag{5}$$

where $\alpha'_X = (\alpha_X + i)$ and $\beta'_X = (\beta_X + n - i)$.

Using the above equation, the expected value is thus defined as:

$$E(\Phi_X | X) = \frac{\alpha'_X}{\alpha'_X + \beta'_X} \tag{6}$$

Similarly, if we have two nodes X and Y, connected such that Y is the parent of X, the conditional posterior probability is defined as:

$$P(\Phi_{X=1|Y=1}) \sim Beta(\alpha'_{X_1|Y_1}, \beta'_{X_1|Y_1})$$

and the expected value of $X_1|Y_1$ is defined as:

$$E(\Phi_X = 1 | Y = 1) = \frac{\alpha'_{X_1|Y_1}}{\alpha'_{X_1|Y_1} + \beta'_{X_1|Y_1}} \tag{7}$$

where $\alpha'_{X_1|Y_1} = (n_{11}+1)$, $\beta'_{X_1|Y_1} = (n_{01}+1)$, n_{11} is the number of observations where both X and Y are '1s', and n_{01} is the number of observations where X is '0' and Y is '1' simultaneously.

3) PPI network analysis: Lung-specific PPI networks for the selected features were constructed using TargetMine [12] as described previously [34]. We further employed Markov-Clustering (MCL) algorithm [35] to extract specific sub-clusters from the extended PPI network (library (markov-clustering) ver. "0.0.6.dev")

To ensure that the inferred clusters were biologically meaningful and minimally noisy, we retained only those clusters that included between five and twenty nodes. The clusters were examined for enriched functional themes as described above.

III. Results

A. Feature Selection and Functional Enrichment Analysis

We used Boruta feature selection protocol to identify an initial list of 489 genes out of the 10,077 genes originally sampled in the dataset. To ensure robustness of the selected features, we retained only 412 of these 489 genes (features) that overlapped in the training and the independent test datasets. Further, principle coordinate analysis followed by tSNE distribution analysis demonstrated that the selected features neatly resolved the NSCLC cohort from the Normal/Healthy cohort (Figure 2a).

We then examined the selected features for enriched biological themes to identify the key cellular processes that were dysregulated in NSCLC. Specifically, the signaling by Rho GTPases emerged one of the key enriched pathways- 13 of the selected feature genes were mapped to the enriched Reactome pathway "Rho GTPase cycle" (R-HSA-194840; p=0.024901); WGCNA algorithm further sequestered 44 genes (Blue module, Additional File 1) specifically perturbed in NSCLC cohort compared with the normal cohort; five of these genes were mapped to enriched Reactome pathway "Rho GTPases Activate Formins" (R-HSA-5663220; p=0.006487). RhoGDI pathway also emerged among the top five canonical pathways associated with the selected features by IPA analysis (Table I). Hence, we further investigated the probable roles of RhoGDI signaling in the context of NSCLC and as a potential target for therapeutic intervention.

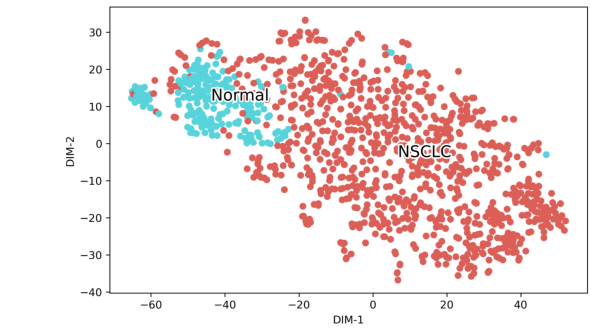


Fig. 2: a) Principle coordinate analysis followed by tSNE plots of the gene expression values of the selected features across Normal (Blue) and NSCLC (Red) cohorts. b) Enriched KEGG pathways associated with the selected features.

TABLE I: Top Canonical Pathways

Name	p-value	Overlap
Kinetochore Metaphase Signaling Pathway	8.61E-05	8.9% (9/101)
Proline Biosynthesis I	1.88E-03	50.0% (2/4)
Extrinsic Prothrombin Activation Pathway	2.70E-03	18.8% (3/16)
Coagulation System	3.45E-03	11.4% (4/35)
RhoGDI Signaling	5.36E-03	5.0% (9/180)

B. Machine Learning

We generated multiple prediction models using different machine learning classifiers and computed their area under curve-receiver operator characteristic (AUROC) score [36] and the Cohen's kappa coefficient [37] to assess their performances. AUROC score ranges from 0.5 and 1.0; where a score of 0.5 indicates that the model has no discriminatory ability to differentiate between classes, while a score of 1.0 indicates high measure of separability and therefore, an excellent model performance. Similarly, the Cohen's kappa coefficient ranges from 0 and 1, with 0 reflecting the worst performing model and 1 reflecting the best performing model.

Using the 412 selected features, we trained the classifiers both on the validation and the independent test data sets. Figures 3 and 4 show the AUROC and Cohen's kappa coefficient scores obtained by each method on the validation and test sets respectively. Based on scores and the strength of the classifiers, we further categorized them into ensemble classifiers (CatBoost, RandomForest, XGBoost, LGBM) and normal classifiers (LogisticRegression, DecisionTree, SVC).

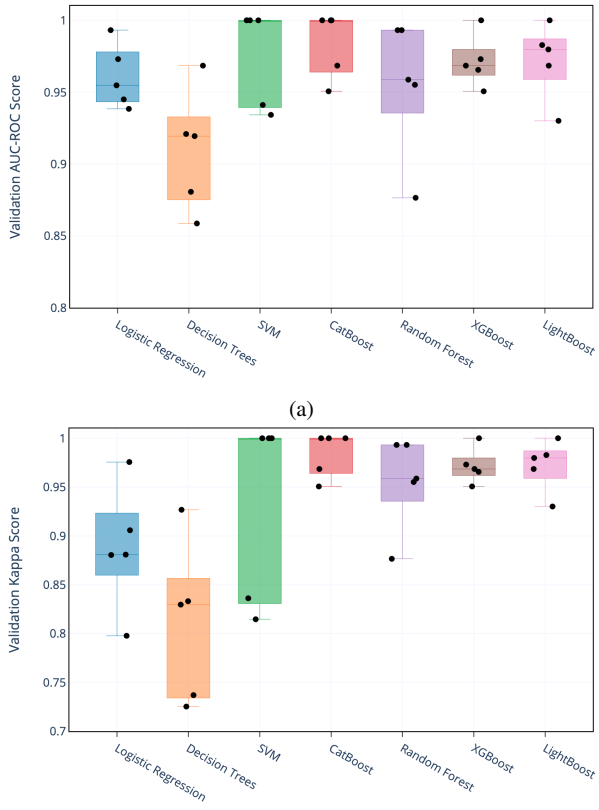
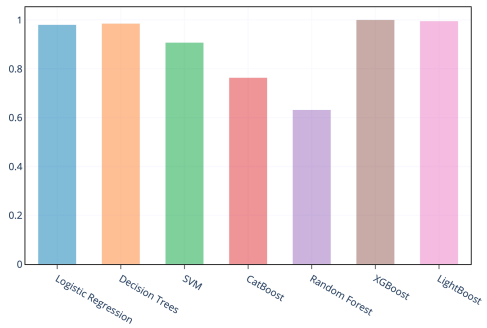
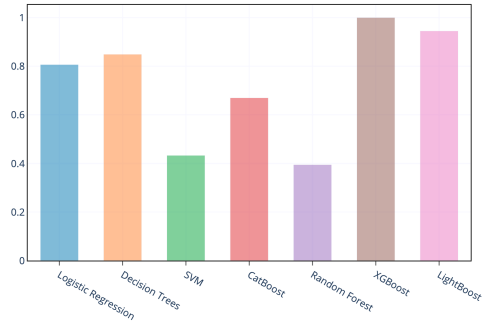


Fig. 3: a) AUROC values of different classifiers on the validation data set. b) Cohen's kappa coefficient values of different classifiers on the validation data set.



(a)

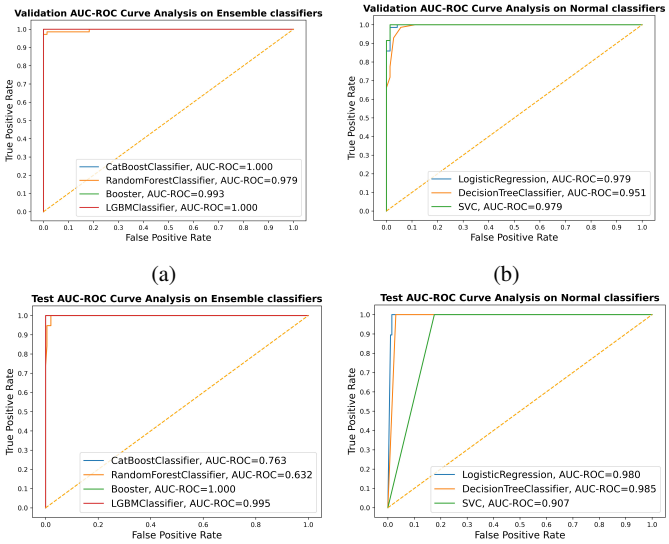


(b)

Fig. 4: a) AUROC values of different classifiers on the independent test data set. b) Cohen's kappa coefficient values of different classifiers on the independent test data set.

Next, we analyzed the ensemble and normal classifiers by plotting the respective ROC curves and computed their AUROC values as shown in Figure 5. The support vector classifier (SVC) and the decision tree classifier models seemingly under-performed, since these models lag behind when faced with a large dimensional dataset with only 412 features and 1,118 rows before over-sampling. This lag results in over-fitting of the normal classifiers, and hence, the ensemble classifiers are employed to overcome this limitation.

XGBoost (Booster) emerged from this analysis as the best-performing method (Figures 3 and 4). Based on a 5-fold cross-validation, XGBoost displayed the least deviation in the scores, and consequently, the smallest over-fitting model and offered the most robust predictions. XGBoost also achieved the highest AUROC score of 0.993 and the highest Cohen's kappa coefficient value of 1.



(a)

(b)

(c)

(d)

Fig. 5: a) ROC curve plotted for the ensemble classifiers using the validation data set. b) ROC curve plotted for the normal classifiers using the validation data set. c) ROC curve plotted for the ensemble classifiers using the independent test data set. d) ROC curve plotted for the normal classifiers using the independent test data set.

C. Bayesian Network Analysis

In parallel to the previously described machine learning approach, we constructed a bayesian network equivalent of the RhoGDI signaling pathway from literature [38], and mapped the key regulatory genes that could potentially be involved in the progression of and cellular proliferation in NSCLC.

Starting from our original list of 412 genes, we selected those that matched the RhoGDI signaling pathway or that are known to be related to said pathway as according to the definitions of the KEGG, Reactome and IPA pathway maps, and used them as input for the definition of the network. Then, we inferred the expected values for each node (using equations 6 and 7) and overlaid these to the nodes of the RhoGDI pathway in order to create both healthy (Figure 6) and tumor (Figure 7) versions of it.

The expected values reflect the probability of a given gene to be over-expressed given the over-expression of its upstream genes, with lower values indicating a relative lower probability of the gene being over-expressed, whereas a higher value indicates a relatively higher

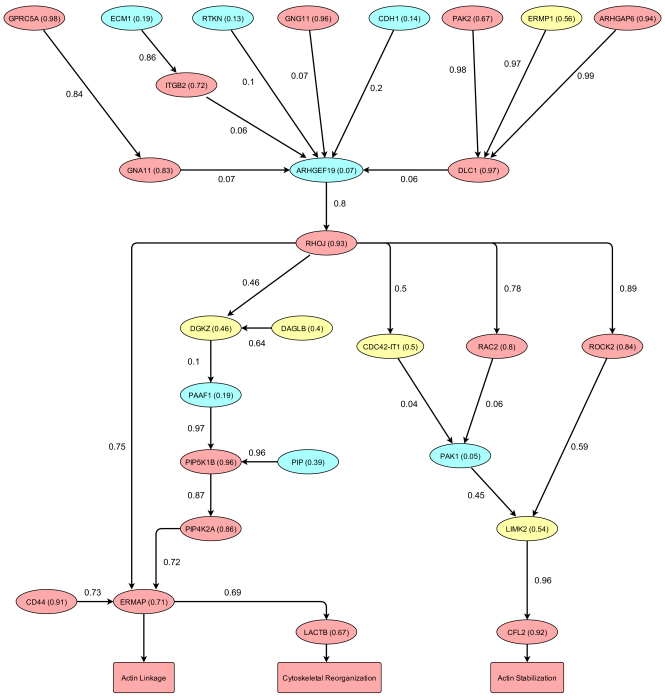


Fig. 6: Bayesian network of the RhoGDI pathway for Healthy samples. Genes are represented as nodes and the arrows represent interactions between them. The values within the nodes correspond to the expected values of healthy samples computed using equation 6. Nodes with scores less than 0.4 are shaded in blue, those with scores between 0.4 and 0.6 in yellow, and the ones with scores greater than 0.6 in red. The values next to the arrows indicate the expected conditional probabilities inferred using equation 7.

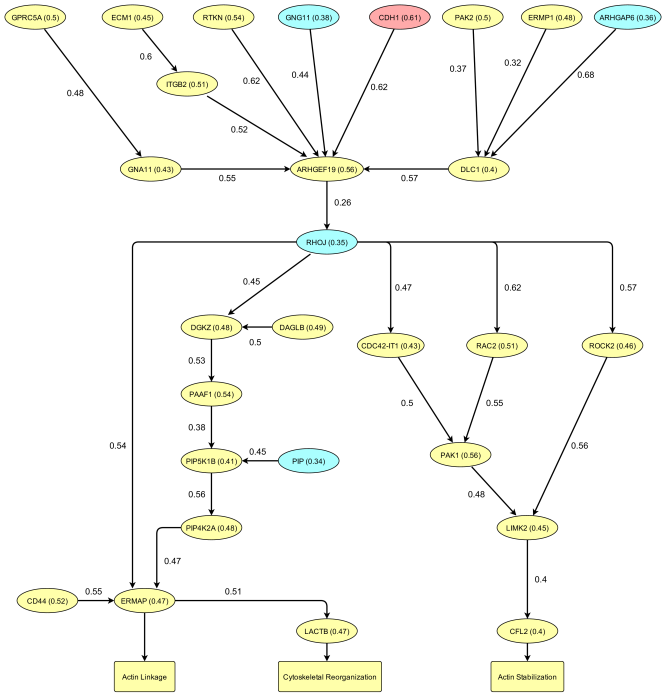


Fig. 7: Bayesian network of the RhoGDI pathway for NSCLC samples. Genes are represented as nodes and the arrows represent interactions among them. The values within the nodes correspond to the expected values of NSCLC samples computed using equation 6. Nodes with scores less than 0.4 were shaded in blue, those with scores between 0.4 and 0.6 in yellow, and the ones with scores greater than 0.6 were shaded in red. The values next to the arrows indicate the expected conditional probabilities inferred using equation 7.

probability. To help quickly identify high- (and low-) likely over-expressed genes, we used color blue to display nodes with value less than 0.4; yellow for nodes with values between 0.4 and 0.6; and red for nodes with values higher than 0.6.

When comparing the over-expression probability values, we observed that genes at the bottom of the RhoGDI pathway, namely *LACTB* (lactamase beta; 0.67), *ERMAP* (erythroblast membrane associated protein [Scianna blood group]; 0.71), and *CFL2* (cofilin 2; 0.92), were less likely to be over-expressed in NSCLC as compared to Normal (Figure 6). Since these genes directly influenced the output processes of Cytoskeletal Reorganization, Actin Linkage, and Actin Stabilization (Figure 6), we proceeded to examine these three processes, in order to pinpoint their key modulators and potential targets for gene intervention.

1) Gene Intervention: The activity of output processes Cytoskeletal Reorganization, Actin Linkage, and Actin Stabilization appeared to be significantly diminished in NSCLC compared with Normal, hence, we attempted to pinpoint the key factors that could be potentially modulated to restore these processes to normalcy.

For this, we computed the conditional probabilities for an output gene (LACTB, ERMAP and CFL2) to be over-expressed given an upstream gene was under-expressed simultaneously. That is, we computed P (an output gene is $\uparrow$ | an upstream gene is $\downarrow$) for each combination of an output gene and an upstream gene in the RhoGDI pathway (Additional File 2). Then, to identify the genes that were significantly involved across all three output processes, we computed their average probability scores. Our calculations suggest that the genes with the highest scores, *ARHGEF19* (Rho guanine nucleotide exchange factor 19; 0.77) and *RAC2* (Rac family small GTPase 2; 0.77) are key modulators of the three output processes.

2) PPI network analysis of feature selected genes: Starting from the original 412 feature-selected genes, we constructed an ensemble PPI network that also incorporated their additional, interacting genes. This NSCLC PPI network consisted of 726 genes (nodes) with 2,407 interactions (edges) between them (Additional File 3). To examine the individual PPIs and cliques that may be potentially relevant to cancer progression, we extracted the largest connected component of the network (688 nodes and 2,380 edges) and employed MCL clustering to partition it into constituent sub-clusters. A total of 134 sub-clusters were inferred in this manner; 33 sub-clusters that ranged from from five to 20 nodes were retained for subsequent analysis (Additional File 4). We retrieved several sub-clusters that were associated with KEGG and/or Reactome pathways relevant to NSCLC (Additional File 5). Sub-cluster 0 included genes that were mapped to the enriched Reactome pathway "Rho GTPases Activate WASPs and WAVEs" (R-HSA-5663213; p=0.018938); sub-cluster 8 included genes that were mapped to "Rho GTPases Activate Formins" (R-HSA-5663220; p=6.2507e-08), "Rho GTPase Effectors" (R-HSA-195258; p=3.36167e-06), "Signaling by Rho GTPases" (R-HSA-194315; p=0.0001319) and "Signaling by Rho GTPases, Miro GTPases and RhoBTB3" (R-HSA-9716542; 0.00013974) pathways; sub-cluster 61 included genes that were mapped to "Rho GTPases Activate Formins" (R-HSA-5663220; p=1.925936e-10); "Rho GTPase Effectors" (R-HSA-195258; 2.913958e-07); "Signaling by Rho GTPases" (R-HSA-194315; p=0.000121); and "Signaling by Rho GTPases, Miro GTPases and RHOBTB3" (R-HSA-9716542; p=0.000136) pathways and; sub-cluster 67 included genes that were mapped to "Rho GTPase cycle" (R-HSA-9012999; p=0.000285), "Signaling by Rho GTPases, Miro GTPases and RhoBTB3" (R-HSA-9716542; p=0.000931); "Signaling by Rho GTPases" (R-HSA-194315; p=0.001253) and "RhoV GTPase cycle" (R-HSA-9013424; p=0.039605) pathways (Additional Files 6 and 7). Notably, sub-cluster 61 included KNL1, which is a central molecule in the Kinetochore Metaphase Signaling Pathway that emerged among the top five canonical pathways in IPA analysis (Table I).

IV. Discussion

NSCLC is the most widespread of all lung cancers and a leading cause of cancer-related mortality worldwide. Furthermore, NSCLC patients remain highly susceptible to a relapse post-operatively. The availability of high-throughput data from NSCLC clinical cohorts has offered unprecedented opportunities to investigate NSCLC biology and evolve new and more effective strategies for NSCLC treatment.

In this study, we have employed an integrative computational approach involving machine learning, functional enrichment analysis, bayesian modeling and PPI network analysis to analyze NSCLC transcriptome dataset. First, using Boruta feature selection, we prioritized the genes (features) that could effectively differentiate NSCLC cohort from the Normal/Healthy cohort. Next, we generated multiple machine learning models using the selected features and evaluated their effectiveness in delineating the NSCLC cohort from the Normal. XGBoost displayed the best performance in these trials, displaying high AUROC scores and Cohen's Kappa coefficient values. Our analysis therefore, demonstrated the suitability of the feature-selected genes to build predictive diagnostic models for NSCLC.

Functional enrichment analysis (and co-expression network analysis using WGCNA) of the feature-selected genes highlighted RhoGDI signaling as one of the key pathways modulating tumorigenesis of NSCLC. We generated a PPI network with the feature selected genes, which was subsequently partitioned into sub-clusters using

the Markov Cluster Algorithm. Four of these sub-clusters were associated with the RhoGDI Signaling pathway. Moreover, one sub-cluster included KNL1, a key gene involved in the Kinetochore Metapahase Signaling Pathway, which had previously emerged as one of the key canonical pathways from the functional enrichment analysis. We believe that these four sub-clusters indeed represent protein complexes highly relevant to NSCLC.

In parallel, we implemented a bayesian framework to pinpoint specific genes and processes that were linked with aberrations in RhoGDI signaling and were likely to be responsible for the proliferation of the NSCLC phenotype. Bayesian modeling analysis suggested that the dysregulation of Actin stabilization induced by aberrations in the expressions of *LACTB*, *ERMAP* and *CFL2* genes were likely to play crucial roles for NSCLC. Taken together, we hypothesized that these cellular processes are significantly under-expressed and therefore more likely to be destabilized in NSCLC as compared to Normal.

Next, we examined the activities of *ARHGEF19* and *RAC2* gene products in the context of RhoGDI signaling and we hypothesized that they were primal in modulating RhoGDI signaling and influencing Actin stabilization and cytoskeleton reorganization further downstream. Indeed, it has been demonstrated previously that the stabilization of the Actin cytoskeleton structures and inhibiting Focal adhesion turnover was able to impede the progression and invasion of NSCLC [39]. Moreover, the over-expression of *ARHGEF19* was previously hypothesized to play a crucial role in NSCLC tumorigenesis by activating MAPK signaling [40]. Our analysis has now suggested that *ARHGEF19* may play a similarly crucial role in modulating the stability of the Actin and cytoskeleton reorganization and therefore, would be an extremely attractive candidate for therapeutic intervention to normalize MAPK signaling and Actin stabilization in the context of NSCLC. Also consistent with our observations, *RAC2* was previously reported to be upregulated in NSCLC tissues and linked with poor prognoses of NSCLC patients [41] and therefore is an attractive candidate for anti-NSCLC therapy. As mentioned before, our study is strongly focused on the expression of genes associated with NSCLC and the selection of the initial datasets clearly reflects this focus. While the treatment of NSCLC has been, to the best of our knowledge, focused on mutations in tumor driver genes such as EGFR, the emergence of resistance against EGFR targeted treatments has necessitated exploring additional parameters such as signaling pathway perturbations that can be examined in the context of gene expression. By using the Boruta selection process to determine the relevant genes in the original dataset, we highlighted the expression of genes that were strongly correlated with our NSCLC cohort, leaving out others that are known to be drivers in lung cancer when mutated.

One study by Liang and colleagues directly links the expression of protein EGFR (among others) to NSCLC, however, the data used by these authors include information from the patients medical records (such as smoking history) together with tumor histology [42]. Here, the authors were able to correlate the expression of EGFR to tumor stage and lymph node metastasis; however, given the limitations of the chosen dataset (which does not include such information), this was a path impossible for us to follow. Nevertheless, it indeed provides an interesting and challenging line of research to continue applying the methods first introduced here, and we look forward at addressing it in the future.

In conclusion, our study has highlighted gene expression signatures and protein complexes strongly correlated with NSCLC. More specifically, we have provided roles for RhoGDI signaling and the *ARHGEF19* and *RAC2* genes in the proliferation of NSCLC phenotype. Collectively, our findings have offered a deeper mechanistic understanding of the pathophysiology of NSCLC and showcased newer avenues for effective anti-NSCLC strategies. In particular, therapeutic targeting of RhoGDI signaling appears to be a promising approach for the treatment of NSCLC. As part of our future work, we expect applying our methodology in the context of collecting more data, could be helpful in further understanding the role the RhoGDI pathway and/or other well know tumor driver pathways/genes might have, not only in the identification, but also in the prognosis of NSCLC patients.

ACKNOWLEDGMENT

This work was supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (Grant Number 17K07268) to K.M. and from the Ministry of Health, Labour and Welfare (Grant Number 19AC5001) to K.M. and in part by the National Science Foundation under Grants ECCS-1609236 and ECCS-1917166 to A.D. The statements made herein are solely the responsibility of the authors.

DATA AVAILABILITY

The custom scripts, data sets and supplementary material associated with this study are available at https://github.com/lokeshtr/NSCLC.

REFERENCES

[1] C. Yi, Y. He, H. Xia, H. Zhang, and P. Zhang, "Review and perspective on adjuvant and neoadjuvant immunotherapies in nsclc," *OncoTargets and therapy*, vol. 12, p. 7329, 2019.

[2] D. H. Schanne, J. Heitmann, M. Guckenberger, and N. H. Andratschke, "Evolution of treatment strategies for oligometastatic nsclc patients–a systematic review of the literature," *Cancer treatment reviews*, vol. 80, p. 101892, 2019.

[3] N. Wani and K. Raza, "Integrative approaches to reconstruct regulatory networks from multi-omics data: A review of state-of-the-art methods," *Computational biology and chemistry*, vol. 83, p. 107120, 2019.

[4] G. Nicora, F. Vitali, A. Dagliati, N. Geifman, and R. Bellazzi, "Integrated multi-omics analyses in oncology: A review of machine learning methods and tools," *Frontiers in Oncology*, vol. 10, p. 1030, 2020.

[5] M. Kunz, J. Jeromin, M. Fuchs, J. Christoph, G. Veronesi, M. Flentje, S. Nietzer, G. Dandekar, and T. Dandekar, "In silico signaling modeling to understand cancer pathways and treatment responses," *Briefings in bioinformatics*, vol. 21, no. 3, pp. 1115–1117, 2020.

[6] H. Vundavilli, A. Datta, C. Sima, J. Hua, R. Lopes, and M. Bittner, "Bayesian inference identifies combination therapeutic targets in breast cancer," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 9, pp. 2684–2692, 2019.

[7] A. Bradley, R. Van der Meer, and C. J. McKay, "A prognostic bayesian network that makes personalized predictions of poor prognostic outcome post resection of pancreatic ductal adenocarcinoma," *PloS one*, vol. 14, no. 9, p. e0222270, 2019.

[8] B. Zhao, Y. Wang, Y. Wang, W. Chen, L. Zhou, P. H. Liu, Z. Kong, C. Dai, Y. Wang, and W. Ma, "Efficacy and safety of therapies for egfr-mutant non-small cell lung cancer with brain metastasis: an evidence-based bayesian network pooled study of multivariable survival analyses," *Aging (Albany NY)*, vol. 12, no. 14, p. 14244, 2020.

[9] S. B. Lim, S. J. Tan, W.-T. Lim, and C. T. Lim, "A merged lung cancer transcriptome dataset for clinical predictive modeling," *Scientific data*, vol. 5, p. 180136, 2018.

[10] M. B. Kursa, W. R. Rudnicki *et al.*, "Feature selection with the boruta package," *J Stat Softw*, vol. 36, no. 11, pp. 1–13, 2010.

[11] A. Mezheyeuski, C. H. Bergsland, M. Backman, D. Djureinovic, T. Sjöblom, J. Bruun, and P. Micke, "Multispectral imaging for quantitative and compartment-specific immune infiltrates reveals distinct immune profiles that classify lung cancer patients," *The Journal of pathology*, vol. 244, no. 4, pp. 421–431, 2018.

[12] Y.-A. Chen, L. P. Tripathi, T. Fujiwara, T. Kameyama, M. N. Itoh, and K. Mizuguchi, "The targetmine data warehouse: enhancement and updates," *Frontiers in genetics*, vol. 10, p. 934, 2019.

[13] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, "Kegg: integrating viruses and cellular organisms," *Nucleic Acids Research*, 2020.

[14] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 46, no. D1, pp. D649–D655, 2018.

[15] G. O. Consortium, "The gene ontology resource: 20 years and still going strong," *Nucleic acids research*, vol. 47, no. D1, pp. D330–D338, 2019.

[16] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi, and I. Golani, "Controlling the false discovery rate in behavior genetics research," *Behavioural brain research*, vol. 125, no. 1-2, pp. 279–284, 2001.

[17] P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[18] D. Cousineau and S. Chartier, "Outliers detection and treatment: a review." *International Journal of Psychological Research*, vol. 3, no. 1, pp. 58–67, 2010.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[20] C. Sammut and G. Webb, "Logistic regression," 2010.

[21] A. M. Ahmed, A. Rizaner, and A. H. Ulusoy, "A novel decision tree classification based on post-pruning with bayes minimum risk," *Plos one*, vol. 13, no. 4, p. e0194168, 2018.

[22] N. Cristianini and E. Ricci, "Support vector machines." 2008.

[23] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," in *Advances in neural information processing systems*, 2018, pp. 6638–6648.

[24] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[25] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[26] F. Alzamzami, M. Hoda, and A. El Saddik, "Light gradient boosting machine for general sentiment classification on short texts: A comparative evaluation," *IEEE Access*, 2020.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[28] Z. S. Chan, L. Collins, and N. Kasabov, "Bayesian learning of sparse gene regulatory networks," *Biosystems*, vol. 87, no. 2-3, pp. 299–306, 2007.

[29] H. Vundavilli, A. Datta, L. Tripathi, and K. Mizuguchi, "Network modeling and inference of peroxisome proliferator-activated receptor pathway in high fat diet-linked obesity," *bioRxiv*, 2020.

[30] R. Li, J. Yu, S. Zhang, F. Bao, P. Wang, X. Huang, and J. Li, "Bayesian network analysis reveals alterations to default mode network connectivity in individuals at risk for alzheimer's disease," *PLoS One*, vol. 8, no. 12, p. e82104, 2013.

[31] S. Marini, E. Trifoglio, N. Barbarini, F. Sambo, B. Di Camillo, A. Malovini, M. Manfrini, C. Cobelli, and R. Bellazzi, "A dynamic bayesian network model for long-term simulation of clinical complications in type 1 diabetes," *Journal of biomedical informatics*, vol. 57, pp. 369–376, 2015.

[32] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using bayesian networks to analyze expression data," *Journal of computational biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[33] R. E. Neapolitan *et al.*, *Learning bayesian networks*. Pearson Prentice Hall Upper Saddle River, NJ, 2004, vol. 38.

[34] L. P. Tripathi, M. N. Itoh, Y. Takeda, K. Tsujino, Y. Kondo, A. Kumanogoh, and K. Mizuguchi, "Integrative analysis reveals common and unique roles of tetraspanins in fibrosis and emphysema," *Frontiers in genetics*, vol. 11, 2020.

[35] S. Van Dongen and C. Abreu-Goodger, "Using mcl to extract clusters from networks," in *Bacterial Molecular Networks*. Springer, 2012, pp. 281–295.

[36] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[37] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica: Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[38] D. Vergara, P. Simeone, S. De Matteis, S. Carloni, P. Lanuti, M. Marchisio, S. Miscia, A. Rizzello, R. Napolitano, C. Agostinelli *et al.*, "Comparative proteomic profiling of hodgkin lymphoma cell lines," *Molecular BioSystems*, vol. 12, no. 1, pp. 219–232, 2016.

[39] B. Wu, S. Yang, H. Sun, T. Sun, F. Ji, Y. Wang, L. Xu, and D. Zhou, "Keap1 inhibits metastatic properties of nsclc cells by stabilizing architectures of f-actin and focal adhesions," *Molecular Cancer Research*, vol. 16, no. 3, pp. 508–516, 2018.

[40] Y. Li, Z. Ye, S. Chen, Z. Pan, Q. Zhou, Y.-Z. Li, W.-d. Shuai, C.-M. Kuang, Q.-H. Peng, W. Shi *et al.*, "A rhgef 19 interacts with braf to activate mapk signaling during the tumorigenesis of non-small cell lung cancer," *International Journal of Cancer*, vol. 142, no. 7, pp. 1379–1391, 2018.

[41] H. Pei, Z. Guo, Z. Wang, Y. Dai, L. Zheng, L. Zhu, J. Zhang, W. Hu, J. Nie, W. Mao *et al.*, "Rac2 promotes abnormal proliferation of quiescent cells by enhanced junb expression via the mal-srf pathway," *Cell Cycle*, vol. 17, no. 9, pp. 1115–1123, 2018.

[42] H. Liang, J. Zhang, C. Shao, L. Zhao, W. Xu, L. C. Sutherland, and K. Wang, "Differential expression of rbm5, egfr and kras mrna and protein in non-small cell lung cancer tissues," *Journal of Experimental & Clinical Cancer Research*, vol. 31, no. 36, 2012. [Online]. Available: https://doi.org/10.1186/1756-9966-31-36