data

MDPI

Article

Graph Theoretic and Pearson Correlation-Based Discovery of Network Biomarkers for Cancer

Raihanul Bari Tanvir, Tasmia Aqila, Mona Maharjan, Abdullah Al Mamun and Ananda Mohan Mondal *

School of Computing and Information Sciences, Florida International University, Miami, FL 33199, USA; rtanv003@fiu.edu (R.B.T.); taqil001@fiu.edu (T.A.); mmaha021@fiu.edu (M.M.); mmamu009@fiu.edu (A.A.M.)
* Correspondence: amondal@fiu.edu

Received: 16 April 2019; Accepted: 3 June 2019; Published: 5 June 2019

check for updates

Abstract: Two graph theoretic concepts—clique and bipartite graphs—are explored to identify the network biomarkers for cancer at the gene network level. The rationale is that a group of genes work together by forming a cluster or a clique-like structures to initiate a cancer. After initiation, the disease signal goes to the next group of genes related to the second stage of a cancer, which can be represented as a bipartite graph. In other words, bipartite graphs represent the cross-talk among the genes between two disease stages. To prove this hypothesis, gene expression values for three cancers—breast invasive carcinoma (BRCA), colorectal adenocarcinoma (COAD) and glioblastoma multiforme (GBM)—are used for analysis. First, a co-expression gene network is generated with highly correlated gene pairs with a Pearson correlation coefficient ≥ 0.9. Second, clique structures of all sizes are isolated from the co-expression network. Then combining these cliques, three different biomarker modules are developed—maximal clique-like modules, 2-clique-1-bipartite modules, and 3-clique-2-bipartite modules. The list of biomarker genes discovered from these network modules are validated as the essential genes for causing a cancer in terms of network properties and survival analysis. This list of biomarker genes will help biologists to design wet lab experiments for further elucidating the complex mechanism of cancer.

Keywords: bipartite graph; clique; network biomarker; Pearson correlation coefficient (PCC); gene co-expression network

1. Introduction

The present work is motivated by the prospective applications of protein-protein interaction (PPI) networks to diseases and other dynamic processes. Ideker and Sharan [1] enumerated four different applications of protein networks to diseases: i) identifying new disease genes, ii) studying the network properties of disease genes, iii) classifying diseases based on protein network, and iv) identifying disease-related subnetworks. Genome-wide PPI networks come with rich information about the dynamic processes such as the behavior of genetic networks in response to DNA damage [2] and exposure to arsenic [3], the prediction of protein function [4], genetic interaction [5], protein subcellular localization [6–11], the process of aging [12], and protein network biomarkers [13–15].

One of the widely used methods for elucidating biomarkers for diseases is through protein-protein interaction (PPI) or gene co-expression networks based on "guilt by association" concept. In a gene co-expression network, nodes represent the genes and edges represent the connection between genes due to significantly similar expression patterns over different samples. Several methods exist for inferring edges in gene networks. Pearson correlation is one of the most common co-expression measures employed in various studies [16,17]. Another common method, Mutual Information (MI) [18] is an information theoretic measure for measuring nonlinear relationship between genes or other variables. A threshold is applied after constructing the co-expression network to retain the most biologically significant correlations between genes.

The main purpose of analyzing gene co-expression networks is to identify the biologically significant modules consist of groups of genes with dense interactions. Usually, highly connected groups have a higher within-group homogeneity and can be considered as biologically significant modules performing a common task, such as shared regulatory inputs or functional pathways. Clustering is a popular method for finding relevant modules from gene co-expression networks. Weighted Gene Correlation Network Analysis (WGCNA) is the most widely used package for module finding [19] which applies hierarchical clustering to find modules. It applies a soft threshold during construction of a gene co-expression network. Several researchers have identified key differentially expressed genes associated with different cancers, such as breast, cervical, colon, esophageal, osteosarcoma and ovarian cancers [20–26], using WGCNA.

Lui et al. [27] used differential entropy technique to identify key genes in diabetes using rat's time-series gene expression data from case and control samples. Guan et al. [28], developed a prediction model using Bayes discriminant method to predict the prognosis of hepatocellular carcinoma based on gene co-expression network.

Graph theoretic methods are also applied for analysis of gene co-expression networks. Shi et.al. [29], proposed an algorithm named Iterative clique enumeration technique (ICE) to discover relatively independent maximal cliques for breast cancer on GEO dataset and found some highly correlated modules that may indicate the tumor grades. Similarly, Perkins et al. used spectral graph theory on Homo sapiens and *Saccharomyces cerevisiae* microarray data for clustering at various thresholds [30]. Zhang et al. [31], discovered the top five hub genes for bladder cancer using the centrality analysis method.

None of the previous studies used clique and bipartite combination to identify the biologically significant modules. The main goal of this paper is to explore the existence of clique-bipartite-like network modules in actual gene network for cancer. Mondal et al. [32] showed that clique-like structures and bipartite graphs could be the building blocks for disease progression, Figure 2 in [32]. The rationale is that a group of proteins or genes work together by forming a network (a clique-like structure) to accomplish a specific function, which could be related to a disease stage [32] and bipartite structure represents the cross-talk among genes between two disease stages.

In this study, gene co-expression network was constructed using highly correlated gene pairs with PCC ≥ 0.9. Three network modules—maximal clique-like graph, 2-clique-1-bipartite graph, and 3-clique-2-bipartite graph—are identified. Finally, the effectiveness of the key genes discovered from these network modules was validated using pathway and survival analyses.

2. Results

Three different types of cancers—breast invasive carcinoma (BRCA), colorectal adenocarcinoma (COAD), and glioblastoma multiforme (GBM)—are considered in the present study to identify network biomarkers. Gene correlation networks based on gene expression profiles of BRCA (20,155 genes for 1093 samples), COAD (19,828 genes for 379 samples), and GBM (19,660 genes for 153 samples) are developed with highly correlated gene pairs (PCC ≥ 0.9). From these networks, three types of gene network modules, considered as network biomarkers, are isolated: i) Single clique-like module based on maximal cliques named as "maximal clique-like" module, ii) clique-bipartite-like modules with two cliques and one bipartite graph named as "2-clique-1-bipartite" modules, which are discovered based on two cliques connected with maximum number of inter-clique connections, and iii) clique-bipartite-like modules with three cliques (A, B, C) and two bipartite graphs (A-B and B-C) named as "3-clique-2-bipartite" modules, which are discovered based on two bipartite graphs having relatively more edges compare to others.

This section is organized in following subsections: Section 2.1—results with the topology of gene co-expression networks; Section 2.2—results with cliques and maximal clique-like modules; Section 2.3—results with 2-clique-1-bipartite modules; and Section 2.4—results with 3-clique-2-bipartite modules.

2.1. Topology of Gene Co-Expression Networks

Table 1 shows the topology of gene co-expression networks for three cancers—BRCA, COAD, and GBM—generated using gene pairs with PCC ≥ 0.9. The network for COAD is the largest and densest composed of 607 genes and 3651 interactions with an average degree of 12. The network for BRCA is the smallest composed of 380 genes and 1034 interactions, which is a sparse network with an average degree of 5.4. The network for GBM is the sparsest with an average degree of 4.9.

Table 1. Topology of gene co-expression network with PCC > 0.9.

Cancer Name	# Of Genes	# Of Edges	Max Degree	Min Degree	Avg Degree
BRCA	380	1034	39	1	5.4
COAD	607	3651	75	1	12.0
GBM	506	1243	49	1	4.9

2.2. Cliques and Maximal Clique-Like Modules

NetworkX [33], a python package, was used to discover cliques of all possible sizes. The total number of cliques are 209, 1535, and 322 for BRCA, COAD, and GBM, respectively. The size of cliques and the corresponding number of cliques (frequency) for each cancer are presented in Supplementary Table S1. It is clear from this table that small-sized cliques (3-node, 4-node, etc.) appear more than the cliques of larger size, as expected. The gene co-expression networks for BRCA, COAD, and GBM have 3, 10, and 6 maximal cliques with 17, 19, and 11 genes, respectively, Supplementary Table S1.

For a particular cancer, most of the genes in maximal cliques are in common, Supplementary Table S2. Thus, it is better to combine the maximal cliques for a cancer to have a single maximal clique-like module for further analysis. The maximal clique-like modules for three cancers—BRCA, COAD, and GBM—are shown in Supplementary Figure S1. Finally, the maximal clique-like modules have 19, 30, and 14 genes for BRCA, COAD, and GBM, respectively, as shown in Table 2. Based on these modules, COAD and GBM cancers share six genes—CD4, HCK, ITGB2, LAIR1, LAPTM5, and SPI1. However, BRCA does not share any genes with the other two cancers. It can be concluded from the maximal clique-like modules that BRCA cancer has a unique behavior which is different from COAD and GBM, whereas COAD and GBM might have some common characteristics.

Table 2. List of genes in maximal clique-like modules for three Cancers—BRCA, COAD, and GBM.

Cancer	List of Genes in Maximal Clique-Like Modules
BRCA	CD2, CD247, CD3D, CD3E, CD5, CD96, CXCR3, IL2RG, LCK, LY9, PTPN7, SH2D1A, SIRPG, SIT1, SLA2, SLAMF1, SLAMF6, TBX21, UBASH3A
COAD	C1QB, C1QC, C3AR1, CD300A, CD4, CD53, CD86, CLEC7A, CSF1R, CYBB, CYTH4, DOK2, FCER1G, FPR3, HAVCR2, HCK, ITGB2, LAIR1, LAPTM5, LILRB1, LILRB4, LRRC25, MS4A4A, PDCD1LG2, SIGLEC7, SIGLEC9, SLAMF8, SPI1, TFEC, TYROBP
GBM	ALOX5, CD4, FERMT3, HCK, ITGB2, LAIR1, LAPTM5, NCKAP1L, PTPN6, SASH3, SPI1, STXBP2, VAV1, WAS

2.3. 2-Clique-1-Bipartite Modules

Figure 1 shows clique-bipartite-like modules composed of two cliques and one bipartite graph for BRCA, COAD, and GBM. The nodes in two cliques are represented by yellow (Clique-1) and gray (Clique-2) colors. Intra-clique connections are blue and inter-clique connections, forming a bipartite graph, are red. In identifying clique-to-clique connections, it is made sure that the two cliques do not have any gene in common. Finding the interconnected cliques is a combinatorial problem. Usually, cliques or cluster of genes representing different stages of a disease are more likely to have cross-talks or interconnections between two cliques. Bipartite graphs between genes of two stages represent the cross-talks. This study focuses on identifying cliques with maximal connections (cross-talks) only. There are 59, 145, and 44 edges that are connecting two cliques in Figure 1a–c, which are the highest in three respective cancers.

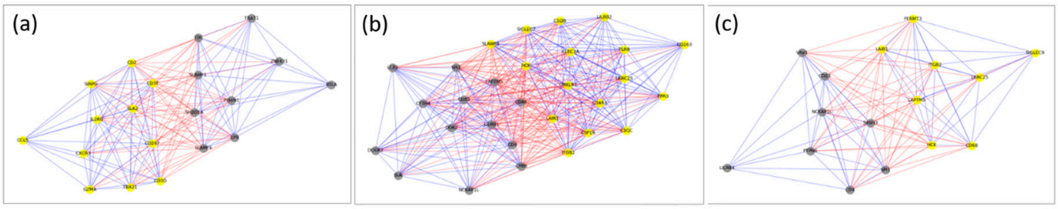

Figure 1. Clique-bipartite-like modules with maximal interconnections between two cliques. (**a**) BRCA; (**b**) COAD; and (**c**) GBM. Nodes in Clique-1 are yellow and nodes in Clique-2 are grey colored. Intra-clique connections are blue and inter-clique connections (a bipartite graph) are red.

Table 3 shows the list of genes discovered from these clique-bipartite-like modules. Based on these modules, COAD and GBM cancers share many genes in common. The common genes—in clique1 for both cancers are HCK, ITGB2, LAIR1, and LRRC25, and for clique2 are CD4, CD53, LILRB1, NCKAP1L, and SPI1. LAPTM5 is the only common gene between clique1 of GBM and clique2 of COAD. On the other hand, BRCA does not share any gene in common. It can be concluded from 2-clique-1-bipartite modules that BRCA cancer has unique behavior, which is different from COAD and GBM cancers, whereas COAD and GBM might have some common characteristics.

Table 3. List of genes in 2-clique-1-bipartite modules.

	BRCA	**COAD**	**GBM**
Clique1	CCL5, CD2, CD247, CD3D, CD3E, CXCR3, GZMA, IL2RG, SIRPG, SLA2, TBX21	C1QB, C1QC, C3AR1, CD163, CLEC7A, CMKLR1, CSF1R, FPR3, HCK, ITGB2, LAIR1, LILRB2, LRRC25, SIGLEC7, SLAMF8, TLR8	CD68, FERMT3, HCK, ITGB2, LAIR1, LAPTM5, LRRC25, SIGLEC9
Clique2	BTLA, ITK, LY9, PYHIN1, SH2D1A, SLAMF1, SLAMF6, TRAT1, ZNF831	CD4, CD53, CD86, CYBB, CYTH4, DOCK2, DOK2, LAPTM5, LCP2, LILRB1, NCKAP1L, SLA, SPI1	CD4, CD53, LILRB4, NCKAP1L, PTPN6, SASH3, SPI1, VAV1

2.4. 3-Clique-2-Bipartite Modules

The top three modules of 3-clique-2-bipartite from each cancer are considered for further analysis. Table 4 summarizes these modules in terms of clique size and the number of inter-clique connections. For example, BRCA-Module1 consists of three cliques of 13, seven, and four genes connected by two bipartite graphs of 56 and 13 connections.

Table 4. Summary statistics of 3-clique-2-bipartite modules.

	Clique-A	**Clique-B**	**Clique-C**	**Connections A-B**	**Connections B-C**
BRCA-Module1	13	7	4	56	13
BRCA-Module2	11	7	4	35	10
BRCA-Module3	8	6	6	8	18
COAD-Module1	16	14	7	85	53
COAD-Module2	16	14	6	111	51
COAD-Module3	16	12	7	69	40
GBM-Module1	9	9	5	30	19
GBM-Module2	9	7	6	22	23
GBM-Module3	9	7	4	36	14

Figure 2 shows the top three 3-clique-2Clique-2-bipartite modules for BRCA. Modules for COAD and GBM are shown in Figure S2. The nodes in three cliques are represented by yellow (clique-A), grey (clique-B) and orange (clique-C) colors. Intra-clique edges are colored blue and inter-clique edges are colored red.

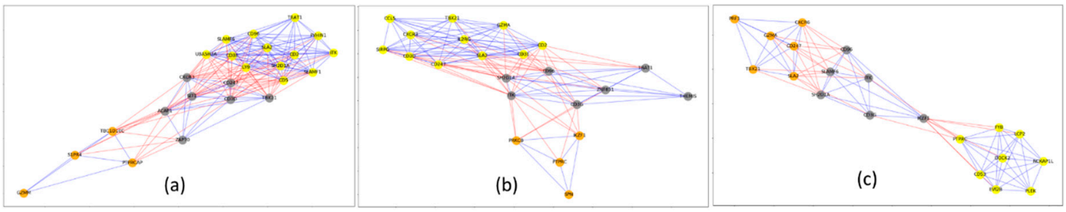

Figure 2. Top three 3-clique-2-bipartite modules for BRCA. Yellow nodes: Clique-A, gray nodes: Clique-B, Orange nodes: Clique-C. Blue: Intra-clique edges, Red: Inter-clique edges. (**a**) Cliques A, B, and C have 13, 7, and 4 nodes respectively. There are 56 connecting edges between cliques A and B and 13 connecting edges between cliques B and C.; (**b**) Cliques A, B, and C have 11, 7, and 4 nodes respectively. There are 35 connecting edges between cliques A and B and 10 connecting edges between cliques B and C.; (**c**) Cliques A, B, and C have 8, 6, and 6 nodes respectively. There are 8 connecting edges between cliques A and B and 18 connecting edges between cliques B and C.

The complete lists of genes that are present in each of the top three 3-clique-2-bipartite modules for BRCA, COAD, and GBM are presented in Supplementary Table S3. Observation of these list reveals that there are many genes in common in three modules of a particular cancer. Table 5 shows the combined list—44, 48, and 32 genes for BRCA, COAD, and GBM respectively. Three cancers share four genes—CD53, DOCK2, IKZF1, and NCKAP1L. Other than these four genes, BRCA and COAD share three more genes—ITK, PTPRC, and TBC1D10C; COAD and GBM share 10 more genes—ARHGAP30, CD4, CD86, CSF1R, HCK, ITGB2, LAIR1, LAPTM5, SASH3, and SPI; and BRCA and GBM do not share any more genes. Thus, BRCA and COAD share a total of seven genes; COAD and GBM share a total of 14 genes; and BRCA and GBM share only four genes. Again, based on 3-clique-bipapartite modules, COAD and GBM shares many genes, which means that they might have some common cause for cancer development. These lists of common genes might provide better insight from lab experiments.

Table 5. Combined list of genes from top three 3-clique-2 bipartite modules.

	List of Genes
BRCA-Modules	ACAP1, CCL5, CD2, CD247, CD3D, CD3E, CD3G, CD5, CD53, CD96, CXCR3, CXCR6, DOCK2, EVI2B, FYB, GZMA, GZMM, IKZF1, IL2RG, ITK, LCP2, LY9, NCKAP1L, PLEK, PRF1, PRKCB, PTPRC, PTPRCAP, PYHIN1, S1PR4, SH2D1A, SIRPG, SIT1, SLA2, SLAMF1, SLAMF6, SPN, TBC1D10C, TBX21, THEMIS, TRAT1, UBASH3A, ZAP70, ZNF831
COAD-Modules	APBB1IP, ARHGAP30, ARHGAP9, BTK, C3AR1, CD163, CD4, CD53, CD84, CD86, CLEC7A, CSF1R, CYBB, CYTH4, DOCK10, DOCK2, FPR3, HAVCR2, HCK, HCLS1, IKZF1, IL10RA, ITGAL, ITGB2, ITK, KLHL6, LAIR1, LAPTM5, LILRB1, LILRB4, LRRC25, MAP4K1, MNDA, MYO1G, NCKAP1L, PIK3R5, PTPRC, RASAL3, SASH3, SIGLEC7, SIGLEC9, SIRPB2, SLA, SLAMF8, SPI1, TBC1D10C, TRAF3IP3, WAS
GBM-Modules	ARHGAP30, ARL11, C1QA, C1QB, C1QC, CD33, CD4, CD53, CD68, CD86, CSF1R, DOCK2, DOCK8, FCER1G, FCGR3A, FERMT3, HCK, IKZF1, ITGB2, LAIR1, LAPTM5, MYO1F, NCF4, NCKAP1L, PLCG2, SASH3, SPI1, STXBP2, SYK, TYROBP, VAMP8, VAV1

3. Discussion

This section discusses the validation of key genes related to three cancers—BRCA, COAD, and GBM—discovered from three network modules—maximal clique-like modules, 2-clique-1-bipartite modules, and 3-clique-2-bipartite modules. First, since the key genes are discovered via network modules, this paper used a network-based app, CytoHubba [34] for validation. The app, CytoHubba, is capable of ranking genes in a network using 12 different graph-theoretic algorithms. The reason for using CytoHubba is that it produces successful results in predicting essential proteins from the yeast protein-protein interaction network [34]. Similarly, in a cancer gene co-expression network, the genes that cause cancer can be thought of as the essential genes for causing that cancer and most likely will have the similar network properties as essential proteins in PPI network. Second, a survival analysis is conducted to show the effectiveness of the key genes discovered using network modules. Finally, pathway and GO term enrichment analyses are conducted for the key genes.

3.1. Validation Using CytoHubba

Figure 3 shows the validation process using two validation metrics—Top 20 genes and Top 50 genes—developed using CytoHubba. The original or base gene network (network created with PCC ≥ 0.9) are analyzed using 12 scoring methods—betweenness, bottleneck, closeness, clustering coefficient (CC), degree, density of maximum neighborhood component (DMNC), eccentricity (EcC), edge percolated component (EPC), maximal clique centrality (MCC), maximum neighborhood component (MNC), radiality, and stress—of CytoHubba to create the list of genes as the benchmark for validation.

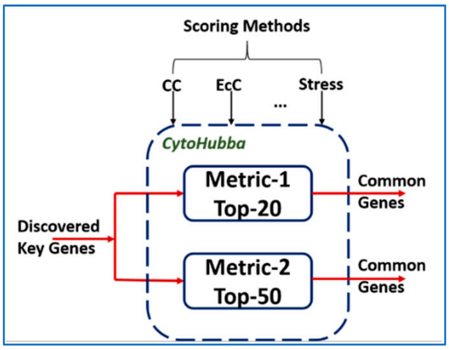

Figure 3. Validation process using two metrics. Metric-1: Top-20 genes from 12 scoring methods of CytoHubba; Metric-2: Top-50 genes from 12 scoring methods of CytoHubba.

Metric-1 (Top-20 Genes): First, Top-20 genes are taken from each of the 12 scoring methods. Then, the genes that appear in two or more scoring methods are considered as the benchmark for validation. The benchmarks for BRCA, COAD, and GBM cancers consist of 41, 53, and 42 genes, respectively, see Supplementary Table S4.

Metric-2 (Top-50 Genes): Similarly, Top-50 genes are taken from each of the 12 scoring methods. Then, the genes that appear in two or more scoring methods are considered as the benchmark for validation. The benchmarks for BRCA, COAD, and GBM cancers consist of 92, 130, and 99 genes respectively, see Supplementary Table S4.

Table 6 shows the number of key genes obtained by combining the unique genes from three modules and the number of these key genes validated by metric-1 and metric-2. For example, 47 key genes were discovered from three network modules of BRCA. These 47 key genes were then compared with the benchmark genes in metric-1 and metric-2. Out of 47 key genes, 26 and 45 genes were found to be common in metric-1 and metric-2, respectively. This validation supports that the list of genes discovered using three modules—maximal clique-like modules, 2-clique-1-bipartite modules, and 3-clique-2-bipartite modules—are essential genes for causing a cancer. This also supports the proposed hypotheses that there exist clique-like and clique-bipartite-like structures, which can be considered as network biomarkers for cancers.

Table 6. Summary of validation.

Dataset.	Key Genes	Keys Genes Common with	
		Metric-1	Metric-2
BRCA	47	26	45
COAD	61	23	53
GBM	38	25	36

3.2. Survival Analysis

Cox proportional hazard regression [35], a semi-parametric method was used for calculating the Cox coefficients of the key genes (Supplemental Table S5). It can adjust survival rate estimation to quantify the effect to predictor variables, which are key genes in the present study. The clinical data of cancer patients (obtained from TCGA) were divided into two equal groups such that each group had the same ratio of dead and alive. One of the groups were used as training set for calculation of Cox coefficients of the key genes. Then, the prognostic risk of each patient in the test set was calculated based on the expression values of key genes using the gene expression grade index (GGI) [36]. The following equation calculates the risk:

$$\text{GGIRiskScore} = \sum x_i - \sum y_i$$

where, x_i and y_i are the expression level of genes with positive and negative cox coefficient.

According to GGI risk score, patients in the test were divided into two groups, as high and low risk groups. The patients with a top 50% GGI risk score are in the high-risk group and others are in the low-risk group. Then a log-rank test was performed to see if there are significant difference in the real survival risks between the two groups.

The survival analysis of key genes of three cancers is shown in Figure 4. It is clear from this figure that the key genes of BRCA, COAD, and GBM are capable of distinguishing between cancer patients in terms of survival in the respective cancers. The log rank p-values between high-risk and low-risk groups were 0.0411, 0.0100, and 0.0171. Log-rank p-values below 0.05 means there is a significant difference between the two groups in consideration. The hazard ratios between high-risk groups and low-risk groups are 1.6478, 2.1627, and 1.6569 for cancer patients of BRCA, COAD, and GBM. This means, for example, high-risk groups of COAD patients are 2.1627 more likely to die than low-risk patients.

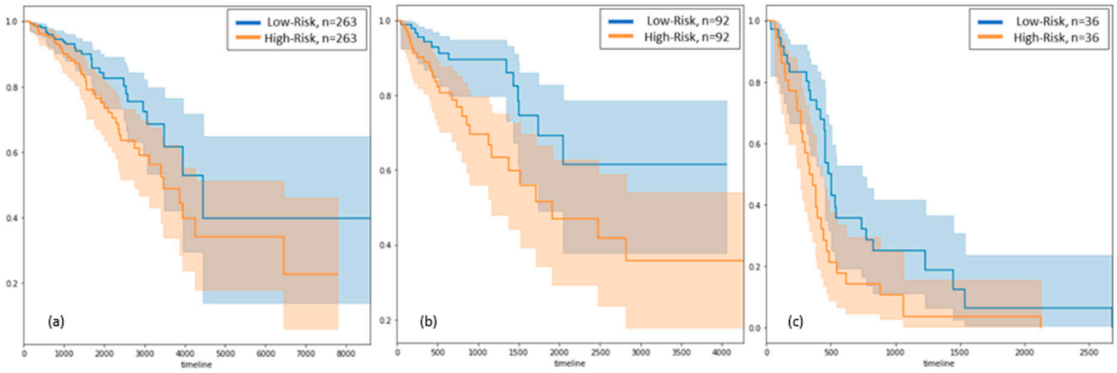

Figure 4. Survival Analysis in data sets of BRCA (**a**), COAD (**b**), and GBM (**c**) cancer patients, using their respective key genes as prognostic factors. The Kaplan–Meyer curve in blue is for the low-risk group and in orange for the high-risk group. The shaded blue and orange regions around their respective lines indicate the confidence interval. The y-axis is the probability of survival and the x-axis is the duration in days.

3.3. Pathway and Gene Ontology Enrichment of Key Genes

The pathway and Gene Ontology (GO) enrichment analyses are also performed for validation of key genes (List of key genes can be found in Supplementary Table S5). Pathway analysis was performed in ReactomeFIViz [37], a Cytoscape app. The false discovery rate (FDR) was calculated based on P-values using Benjamini–Hochberg method. The top ten pathways enriched in three cancers were compared. The pathways enriched in at least two cancers is listed in Supplementary Table S6. TCR signaling in -ve CD4+ T cells is enriched in all three cancers. There are five other pathways- Neutrophil degranulation, Osteoclast differentiation, Staphylococcus aureus infection, Natural killer cell mediated cytotoxicity and Fc gamma R-mediated phagocytosis are enriched in both COAD and GBM. This may be due to genes in common between COAD and GBM in maximal clique-like modules and 2-clique-1 bipartite modules. BRCA showed unique behavior in both modules. In 3-clique-2-bipartite module, BRCA had three genes in common with other two cancers and three more in common with COAD. The pathway T cell receptor signaling pathway is the only pathway enriched in both BRCA and GBM. In all three modules, COAD and GBM shared the same number genes. This is further observed in the enriched pathways they share in common.

A Cytoscape app, BiNGO [38] was used for GO enrichment analysis in three categories- biological process (BP), cellular component (CC), and molecular function (MF). BiNGO uses the Benjamini and Hochberg (false discovery rate) statistical method for multiple testing correction. The top ten enriched GO terms in three cancers were compared. GO terms common in at least two cancers are listed in Supplementary Table S7.

The biological processes enriched in all three cancers are Immune system process, regulation of immune system process, positive regulation of immune system, and T cell activation. Three more biological processes are enriched in both BRCA and GBM, and two more are in COAD and GBM. Most of the common enriched BPs are related to immune system. It is an accepted fact that immune cells have the ability to influence cancer [39]. This is another validation of the key genes discovered in the present study.

There are five cellular components enriched in all three cancers—plasma membrane, plasma membrane part, integral to plasma membrane, intrinsic to plasma membrane, and receptor complex. The dysregulation of the structural integrity of plasma membrane or its domain is known to promote oncogenic signaling [40]. Three other pathways—T cell receptor complex, membrane, and cell surface—are enriched in at least two of three cancers.

The three molecular functions enriched in all three cancers are molecular transducer activity, signal transducer activity, and protein binding. Two other molecular functions—GTPase regulatory activity and nucleoside-triphosphatase regulator activity—are enriched in COAD and GBM while receptor activity and non-membrane spanning protein tyrosine kinase are enriched in BRCA and COAD.

3.4. Future Direction

This study discovers key genes related to cancers from gene co-expression networks. There are three epigenetic factors that drive the cancer via gene expression of cancer genes, which are: i) DNA methylation, ii) histone modification, and iii) miRNA dysregulation. Future study will be conducted to determine how these three epigenetic factors are related to the genes discovered in this study. A study will be conducted for further analysis of the clique-like disease progression to identify the core clique, which could be a clique of three or more genes, for initiating a cancer utilizing the information from three epigenetic factors. Finally, we will explore how the core clique expands to a maximal clique-like structure in the final stage of a cancer.

4. Materials and Methods

4.1. Dataset Preparation

Gene expression data for BRCA, COAD, and GBM are obtained from LinkedOmics [41]. The datasets consist of gene expression values of 20155 genes for 1093 samples, 19,828 genes for

379 samples, and 19,660 genes for 153 samples, respectively, for BRCA, COAD, and GBM as mentioned in Table 7. In these datasets, all samples are cancer patients.

Table 7. Summary of gene expression data for BRCA, COAD, and GBM.

Cancer	No of Genes	No of Samples	Reduced no of Genes
Breast invasive carcinoma (BRCA)	20,155	1093	16,011
Colorectal adenocarcinoma (COAD)	19,828	379	15,769
Glioblastoma multiforme (GBM)	19,660	153	16,186

The missing values were imputed using the fancyimpute package in Python employing the k-nearest neighbors algorithm. The number of genes in the reduced datasets are 16,011, 15,769, and 16,186, respectively, for BRCA, COAD, and GBM. For the present study, highly correlated positive gene pairs, PCC ≥ 0.9 in each cancer are considered for creating the base networks for further analysis.

4.2. Method to Identify Clique and Clique-Bipartite-Like Modules

To discover the cluster of genes or cliques and how they are connected to each other by forming bipartite graphs, Python package NetworkX [33] is used. First, list of cliques with different sizes are discovered. Then, using the list of cliques and the original network (network created with PCC ≥ 0.9), three types of gene network modules, considered as network biomarkers, are discovered—i) maximal clique-like modules, ii) 2-clique-1Clique-1-bipartite modules, and iii) 3-clique-2Clique-2-bipartite modules.

Maximal clique-like module: The discovered cliques are organized in a list based on their size and frequency of occurrence. From the sorted list, the size and number of maximal (largest) cliques in each cancer are found and then combined together to get the maximal clique-like module. This process generates a single maximal clique-like module for each cancer.

2-Clique-1-bipartite module: These are clique-bipartite-like modules with two cliques and one bipartite graph, which are discovered based on two cliques connected with maximum number of inter-clique connections.

3-Clique-2-bipartite module: With the list of cliques and the original network (network created with PCC ≥ 0.9), a list of three connected cliques A, B, and C is generated in a way such that clique A is connected to clique B and clique B is connected to clique C, but cliques A, B, and C do not have any common genes. This process takes longer than usual because of the high number of cliques and the problem is combinatorial in nature. Every combination of three cliques is being checked to see whether it fulfills the condition. These modules are identified by first sorting the list by number of edges connecting cliques A and B and then sorting by number of edges connecting cliques B and C. It is observed that if one of the edge-count (between cliques A and B) has the highest value then the other edge-count (between cliques B and C) has very low value. Finally, from the sorted list, structures having both the edge-counts higher than others are selected as the possible network modules for a cancer. The top three structures from each cancer are considered for further analysis.

5. Conclusions

This paper used two graph theoretic concepts—clique and bipartite graphs—to identify the network biomarkers for cancer from gene co-expression networks developed with highly correlated gene pairs. The gene expression profiles of three cancers—BRCA, COAD, and GBM—are considered for experiment. Results show that three types of network modules—maximal clique-like, 2-clique-1-bipartite, and 3-clique-2-bipartite graphs—derived using the simple graph theoretic concepts clique and bipartite graph are capable of representing cancer dynamics at the gene network level. The combined list of genes from three network modules for a particular cancer are validated with the benchmark developed from a network-based tools CytoHubba. The effectiveness of the key genes is also validated by survival and pathway analyses.

The discovered gene network modules provide a short list of genes related to cancer that can be used by the biologist to design wet lab experiment for further elucidation of the complex mechanism of cancer.

Supplementary Materials: The following are available online at http://www.mdpi.com/2306-5729/4/2/81/s1; Figure S1. Maximal clique-like modules; Figure S2. Top three 3-clique-2-bipartite modules for COAD and GBM; Table S1. Frequency of cliques according to sizes; Table S2. List of genes in maximal cliques; Table S3. List of genes for top three 3-clique-2-bipartite modules; Table S4. List of benchmark genes in top-20 and top-50 metrics; Table S5. List of genes combining three network modules; Table S6. Enriched pathways with key genes; Table S7. GO term enrichment analysis.

Author Contributions: Conceptualization: A.M.M. and R.B.T.; methodology: A.M.M., R.B.T., and T.A.; software: R.B.T. and T.A.; validation: A.A.M. and M.M.; formal analysis: R.B.T. and T.A.; investigation: R.B.T., T.A., M.M., and A.A.M.; resources: R.B.T., T.A., M.M., and A.A.M.; data curation: T.A. and R.B.T.; writing—original draft preparation: A.M.M.; writing—review and editing: A.M.M., T.A., R.B.T., M.M., and A.A.M.; visualization: A.M.M., R.B.T., M.M., and A.A.M.; supervision: A.M.M.; project administration: A.M.M.; funding acquisition: A.M.M.

Funding: This research is funded by NSF CAREER award #1651917 (transferred to #1901628) to AMM.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ideker, T.; Sharan, R. Protein networks in disease. *Genome Res.* **2008**, *18*, 644–652. [CrossRef] [PubMed]
2. Bandyopadhyay, S.; Mehta, M.; Kuo, D.; Sung, M.-K.; Chuang, R.; Jaehnig, E.J.; Bodenmiller, B.; Licon, K.; Copeland, W.; Shales, M.; et al. Rewiring of Genetic Networks in Response to DNA Damage. *Science* **2010**, *330*, 1385–1389. [CrossRef] [PubMed]
3. Haugen, A.C.; Kelley, R.; Collins, J.B.; Tucker, C.J.; Deng, C.; Afshari, C.A.; Brown, J.M.; Ideker, T.; Van Houten, B. Integrating phenotypic and expression profiles to map arsenic-response networks. *Genome Boil.* **2004**, *5*, R95. [CrossRef] [PubMed]
4. Lee, H.; Tu, Z.; Deng, M.; Sun, F.; Chen, T. Diffusion Kernel-Based Logistic Regression Models for Protein Function Prediction. *OMICS A J. Integr. Boil.* **2006**, *10*, 40–55. [CrossRef] [PubMed]
5. Qi, Y.; Suhail, Y.; Lin, Y.; Boeke, J.D.; Bader, J.S. Finding friends and enemies in an enemies-only network: A graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res.* **2008**, *18*, 1991–2004. [CrossRef] [PubMed]
6. Ananda, M.M.; Hu, J. NetLoc: Network based protein localization prediction using protein-protein interaction and co-expression networks. In Proceedings of the 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Hong Kong, China, 18–21 December 2010; pp. 142–148.
7. Mondal, A.; Lin, J.-R.; Hu, J. Network based subcellular localization prediction for multi-label proteins. In Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), Atlanta, GA, USA, 12–15 November 2011.
8. Mondal, A.M.; Hu, J. Protein Localization by Integrating Multiple Protein Correlation Networks. Proceedings of The 2012 International Conference on Bioinformatics & Computational Biology (BIOCOMP'12), Las Vegas, NV, USA, 16–19 July 2012; pp. 82–88.
9. Lin, J.-R.; Mondal, A.M.; Liu, R.; Hu, J. Minimalist ensemble algorithms for genome-wide protein localization prediction. *BMC Bioinform.* **2012**, *13*, 157. [CrossRef]
10. Mondal, A.; Hu, J. Scored Protein-Protein Interaction to Predict Subcellular Localizations for Yeast Using Diffusion Kernel. In *International Conference on Pattern Recognition and Machine Intelligence*; Springer: Berlin/Heidelberg, Germany, 2013.
11. Mondal, A.; Hu, J. Network based prediction of protein localisation using diffusion kernel. *Int. J. Data Min. Bioinform.* **2014**, *9*, 386–400. [CrossRef]
12. Faisal, F.E.; Milenkovic, T. Dynamic networks reveal key players in aging. *Bioinformatics* **2014**, *30*, 1721–1729. [CrossRef]

13. Kevin, C.; Andrews, A.; Ananda, M. Protein Subnetwork Biomarkers for Yeast Using Brute Force Method. In Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP), Las Vagas, NV, USA, 22–25 July 2013; pp. 218–223.
14. Timalsina, P.; Charles, K.; Mondal, A.M. STRING PPI Score to Characterize Protein Subnetwork Biomarkers for Human Diseases and Pathways. In Proceedings of the 2014 IEEE International Conference on Bioinformatics and Bioengineering, Boca Raton, FL, USA, 10–12 November 2014; pp. 251–256.
15. Maharjan, M.; Tanvir, R.B.; Chowdhury, K.; Mondal, A.M. Determination of Biomarkers for Diagnosis of Lung Cancer Using Cytoscape-based GO and Pathway Analysis. In Proceedings of the 20th International Conference on Bioinformatics & Computational Biology (BIOCOMP'19), Las Vegas, NV, USA, 29 July–01 Aug 2019. (Accepted).
16. Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 14863–14868. [CrossRef]
17. Wolfe, C.J.; Kohane, I.S.; Butte, A.J. Systematic survey reveals general applicability of 'guilt-by-association' within gene coexpression networks. *BMC Bioinform.* **2005**, *6*, 227. [CrossRef]
18. Butte, A.J.; Kohane, I.S. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.* **2000**, 418–429.
19. Zhang, B.; Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **2005**, *4*, 17. [CrossRef]
20. Tang, J.; Lu, M.; Cui, Q.; Zhang, D.; Kong, D.; Liao, X.; Ren, J.; Gong, Y.; Wu, G. Overexpression of ASPM, CDC20, and TTK Confer a Poorer Prognosis in Breast Cancer Identified by Gene Co-expression Network Analysis. *Front. Oncol.* **2019**, *9*, 310. [CrossRef]
21. Lalremmawia, H.; Tiwary, B.K. Identification of Molecular Biomarkers for Ovarian Cancer using Computational Approaches. *Carcinogenesis* **2019**. [CrossRef]
22. Maertens, A.M.; Tran, V.; Kleensang, A.; Hartung, T. Weighted Gene Correlation Network Analysis (WGCNA) Reveals Novel Transcription Factors Associated With Bisphenol A Dose-Response. *Front. Genet.* **2018**, *9*, 508. [CrossRef]
23. Shi, H.; Zhang, L.; Qu, Y.; Hou, L.; Wang, L.; Zheng, M. Prognostic genes of breast cancer revealed by gene co-expression network analysis. *Oncol. Lett.* **2017**, *14*, 4535–4542. [CrossRef]
24. Liu, X.; Hu, A.-X.; Zhao, J.-L.; Chen, F. Identification of Key Gene Modules in Human Osteosarcoma by Co-Expression Analysis Weighted Gene Co-Expression Network Analysis (WGCNA). *J. Cell. Biochem.* **2017**, *118*, 3953–3959. [CrossRef]
25. Zhang, C.; Sun, Q. Weighted gene co-expression network analysis of gene modules for the prognosis of esophageal cancer. *J. Huazhong Univ. Sci. Technol. [Med. Sci.]* **2017**, *37*, 319–325. [CrossRef]
26. Liu, R.; Zhang, W.; Liu, Z.; Zhou, H. Associating transcriptional modules with colon cancer survival through weighted gene co-expression network analysis. *BMC Genom.* **2017**, *18*, 361. [CrossRef]
27. Liu, Z.-P.; Gao, R. Detecting pathway biomarkers of diabetic progression with differential entropy. *J. Biomed. Inform.* **2018**, *82*, 143–153. [CrossRef]
28. Guan, L.; Luo, Q.; Liang, N.; Liu, H. A prognostic prediction system for hepatocellular carcinoma based on gene co-expression network. *Exp. Ther. Med.* **2019**, *17*, 4506–4516. [CrossRef]
29. Shi, Z.; Derow, C.K.; Zhang, B. Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. *BMC Syst. Biol.* **2010**, *4*, 74. [CrossRef]
30. Perkins, A.D.; Langston, M.A. Threshold selection in gene co-expression networks using spectral graph theory techniques. *BMC Bioinform.* **2009**, *10*, S4. [CrossRef]
31. Zhang, D.-Q.; Zhou, C.; Chen, S.-Z.; Yang, Y.; Shi, B. Identification of hub genes and pathways associated with bladder cancer based on co-expression network analysis. *Oncol. Lett.* **2017**, *14*, 1115–1122. [CrossRef]
32. Mondal, A.M.; Schultz, C.A.; Sheppard, M.; Carson, J.; Tanvir, R.B.; Aqila, T. Graph Theoretic Concepts as the Building Blocks for Disease Initiation and Progression at Protein Network Level: Identification and Challenges. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, Madrid, Spain, 3–6 December 2018.
33. Hagberg, A.A.; Schult, D.A.; Swart, P.J. Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference (SciPy), Pasadena, CA, USA, 19–24 August 2008; pp. 11–15.

34. Chin, C.-H.; Chen, S.-H.; Wu, H.-H.; Ho, C.-W.; Ko, M.-T.; Lin, C.-Y. cytoHubba: Identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* **2014**, *8* (Suppl. 4), S11. [CrossRef]
35. Mauger, E.A.; Wolfe, R.A.; Port, F.K. Transient effects in the cox proportional hazards regression model. *Stat. Med.* **1995**, *14*, 1553–1565. [CrossRef]
36. Sotiriou, C.; Wirapati, P.; Loi, S.; Harris, A.; Fox, S.; Smeds, J.; Nordgren, H.; Farmer, P.; Praz, V.; Haibe-Kains, B.; et al. Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade to Improve Prognosis. *J. Natl. Cancer Inst.* **2006**, *98*, 262–272. [CrossRef]
37. Wu, G.; Dawson, E.; Duong, A.; Haw, R.; Stein, L. ReactomeFIViz: The Reactome FI Cytoscape app for pathway and network-based data analysis. *F1000Research* **2014**, *3*, 146. [CrossRef]
38. Maere, S.; Heymans, K.; Kuiper, M. BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* **2005**, *21*, 3448–3449. [CrossRef]
39. Monette, A.; Bergeron, D.; Ben Amor, A.; Meunier, L.; Caron, C.; Mes-Masson, A.-M.; Kchir, N.; Hamzaoui, K.; Jurisica, I.; Lapointe, R. Immune-enrichment of non-small cell lung cancer baseline biopsies for multiplex profiling define prognostic immune checkpoint combinations for patient stratification. *J. Immunother. Cancer* **2019**, *7*, 86. [CrossRef]
40. Erazo-Oliveras, A.; Fuentes, N.R.; Wright, R.C.; Chapkin, R.S. Functional link between plasma membrane spatiotemporal dynamics, cancer biology, and dietary membrane-altering agents. *Cancer Metastasis Rev.* **2018**, *37*, 519–544. [CrossRef]
41. Vasaikar, S.V.; Straub, P.; Wang, J.; Zhang, B. LinkedOmics: Analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* **2017**, *46*, D956–D963. [CrossRef]

International Journal of
Molecular Sciences

MDPI

Article

Multi-Run Concrete Autoencoder to Identify Prognostic lncRNAs for 12 Cancers

Abdullah Al Mamun [1,†], Raihanul Bari Tanvir [1,†], Masrur Sobhan [1,†], Kalai Mathee [2,3], Giri Narasimhan [1,3], Gregory E. Holt [4,5] and Ananda Mohan Mondal [1,2,3,*]

[1] Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, FL 33199, USA; mmamu009@fiu.edu (A.A.M.); rtanv003@fiu.edu (R.B.T.); msobh002@fiu.edu (M.S.); giri@fiu.edu (G.N.)
[2] Department of Human and Molecular Genetics, Herbert Wertheim College of Medicine, Florida International University, Miami, FL 33199, USA; matheek@fiu.edu
[3] Biomolecular Sciences Institute, Florida International University, Miami, FL 33199, USA
[4] Department of Medicine, Miami VA Healthcare System, Miami, FL 33125, USA; gholt@med.miami.edu
[5] Department of Medicine, University of Miami, Miami, FL 33146, USA
[*] Correspondence: amondal@fiu.edu
[†] Equal contribution.

Abstract: Background: Long non-coding RNA plays a vital role in changing the expression profiles of various target genes that lead to cancer development. Thus, identifying prognostic lncRNAs related to different cancers might help in developing cancer therapy. Method: To discover the critical lncRNAs that can identify the origin of different cancers, we propose the use of the state-of-the-art deep learning algorithm concrete autoencoder (CAE) in an unsupervised setting, which efficiently identifies a subset of the most informative features. However, CAE does not identify reproducible features in different runs due to its stochastic nature. We thus propose a multi-run CAE (mrCAE) to identify a stable set of features to address this issue. The assumption is that a feature appearing in multiple runs carries more meaningful information about the data under consideration. The genome-wide lncRNA expression profiles of 12 different types of cancers, with a total of 4768 samples available in The Cancer Genome Atlas (TCGA), were analyzed to discover the key lncRNAs. The lncRNAs identified by multiple runs of CAE were added to a final list of key lncRNAs that are capable of identifying 12 different cancers. Results: Our results showed that mrCAE performs better in feature selection than single-run CAE, standard autoencoder (AE), and other state-of-the-art feature selection techniques. This study revealed a set of top-ranking 128 lncRNAs that could identify the origin of 12 different cancers with an accuracy of 95%. Survival analysis showed that 76 of 128 lncRNAs have the prognostic capability to differentiate high- and low-risk groups of patients with different cancers. Conclusion: The proposed mrCAE, which selects actual features, outperformed the AE even though it selects the latent or pseudo-features. By selecting actual features instead of pseudo-features, mrCAE can be valuable for precision medicine. The identified prognostic lncRNAs can be further studied to develop therapies for different cancers.

Keywords: autoencoder; concrete autoencoder; deep learning; feature selection; lncRNA; mrCAE

check for updates

Citation: Al Mamun, A.; Tanvir, R.B.; Sobhan, M.; Mathee, K.; Narasimhan, G.; Holt, G.E.; Mondal, A.M. Multi-Run Concrete Autoencoder to Identify Prognostic lncRNAs for 12 Cancers. *Int. J. Mol. Sci.* **2021**, *22*, 11919. https://doi.org/10.3390/ijms222111919

Academic Editor: Jung Hun Oh

Received: 1 October 2021
Accepted: 30 October 2021
Published: 3 November 2021

1. Introduction

Recent studies have shown that long non-coding RNAs (lncRNAs), which are longer than 200 nucleotides, play key roles in tumorigenesis [1–3]. lncRNAs also have key functions in transcriptional, post-transcriptional, and epigenetic gene regulation [4]. Schmitt and Chang discussed the impact of lncRNA in cancer pathways [5]. Hanahan and Weinberg described the involvement of lncRNAs in six hallmarks of cancer such as proliferation, growth suppression, motility, immortality, angiogenesis, and viability [6].

Hoadley et al. showed that cell-of-origin patterns dominate the molecular classification of tumors available in The Cancer Genome Atlas (TCGA) [7]. Their analysis used copy number, mutation, DNA methylation, RPPA protein, mRNA, and miRNA expression. However, they did not consider another important molecular signature of cancer: lncRNA expression. This work motivated us to investigate the importance of lncRNAs in identifying different types of cancer. We hypothesized that there should be a shortlist of salient features or important lncRNAs with prognostic capability that could dictate the origin of multiple cancers.

In general, feature selection is worthwhile when the whole set of features is difficult to collect or expensive to generate [8]. For example, in TCGA, the lncRNA expression profile dataset contains more than 12,000 features (lncRNAs) for 33 different cancers, and it is expensive to generate these data.

Standard dimension reduction methods, such as principal component analysis (PCA) [9] and autoencoders [10], can generate a greatly reduced set of latent features. However, these latent features are not the original features but functional combinations of the original features. Identifying original features increases the explainability of results and allows one to perform biological interpretation when diagnosing various deadly diseases, such as cancers. Recently, a few deep learning-based feature selection methods showed improvement in selecting original features in both supervised and unsupervised settings [8,11–13].

In our previous study [14], we showed that a deep learning-based unsupervised feature selection algorithm CAE [8] performed better in feature selection, especially in selecting a small number of features, compared to state-of-the-art supervised feature selection methods such as least absolute shrinkage and selection operator (LASSO) [15], random forest (RF) [16], and support vector machine with recursive feature elimination (SVM–RFE) [17]. However, the study was only based on the expression profiles of cancer patients. Questions remained unanswered regarding (a) whether the identified lncRNAs were cancer-specific or organ-specific, (b) which set to use as the final feature set given that CAE produces a different set of features in different runs, (c) whether the identified lncRNAs have prognostic capability, and (d) the validation method of the identified lncRNAs.

In this paper, to address question (a), we analyzed data from 12 cancers, with a normal to cancer sample ratio of at least 1:10. To address question (b), we ran CAE multiple times, with a fixed number of features to be selected in each run and the most frequently appearing features in multiple runs taken as the final set of features. To address question (c), survival analysis was performed to show that the identified features have prognostic capability. To address question (d), we checked the existence of identified lncRNAs in experimental works of literature, drug–lncRNA networks, and cancer hallmarks.

The contributions of this study are as follows: (1) the development of an optimal and stable feature selection framework, mrCAE; (2) the discovery of an optimal and stable set of 128 lncRNAs capable of identifying the origin of organs for 12 different cancers with an accuracy of 95%; (3) the demonstration that the lncRNAs identified using mrCAE from the expression profiles of cancer patients are truly cancer-specific, not organ-specific; (4) the survival or prognostic analysis of discovered lncRNAs; and (5) the validation of identified features, lncRNAs, with existing literature, drug–lncRNA networks, and hallmark lncRNAs.

2. Results

The lncRNA expression profiles of 12 cancers were analyzed with the goal of identifying the key lncRNAs using mrCAE. First, we showed that the features selected by CAE were truly cancer-specific, not organ-specific. Second, we showed the stochastic nature of CAE in selecting equally significant different sets of features in different runs. Third, we showed that mrCAE performed better than the single-run CAE and other state-of-the-art feature selection methods (LASSO, RF, SVM–RFE, MCFS, and UDFS). Fourth, we determined a stable set of lncRNAs that not only could stratify 12 different cancer types but also had the highest number of lncRNAs with prognostic behavior.

2.1. Features Selected from Tumor Tissues Are Cancer-Specific Not Organ-Specific

To check that the features selected by CAE from the lncRNA expression profiles of cancer samples were truly cancer-specific, not organ-specific, we separately ran CAE on tumor and normal samples to identify two sets of 80 features (lncRNAs). Figure 1a shows only five commons between 80 tumor and 80 normal features, which evidenced that 75 out of 80 features were unique to both tumor and normal tissues. It is clear from the t-SNE plots of Figure 1b,c that the tumor and normal features could distinctively cluster 12 tumor tissues and corresponding normal tissues, respectively. However, when we cross-validated the t-SNE plot of tumor tissues using normal features (Figure 1d) and the t-SNE plot of normal tissues using tumor features Figure 1e, we found no distinct clusters for 12 tumor and corresponding normal tissues. *Supplementary 2* shows similar results for the 40-feature and 60-feature scenarios. These experiments proved that the features derived from tumor samples were truly cancer-specific, not organ-specific.

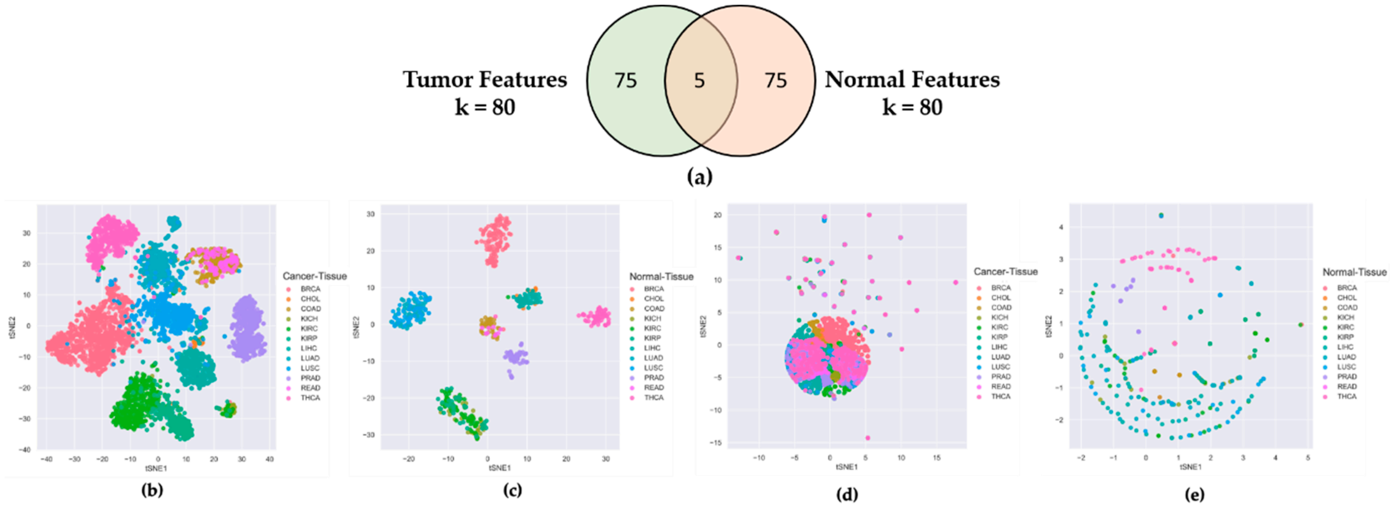

Figure 1. Comparing Tumor Features with Normal Features. (**a**) Venn diagram of 80 tumor features and 80 normal features derived from CAE; (**b**) t-SNE plot of tumor samples using tumor features; (**c**) t-SNE plot of normal samples using normal features; (**d**) t-SNE plot of tumor samples using normal features; (**e**) t-SNE plot of normal samples using tumor features.

2.2. CAE Produces Different Sets of Significant Features in Different Runs

Though CAE selects a subset of the most significant features from a given dataset, it produces different sets of significant features in different runs due to its stochastic nature [8]. To show the stochastic nature of CAE, three sets of 60 features were selected for the experiment. Figure 2 shows the (a) Venn diagram, (b) classification accuracy of 12 cancer types, (c) mean squared error (MSE) of reconstructing original feature space, and (d) t-SNE plots of clustering 12 different types of cancer samples using three sets of features.

Figure 2 presents evidence that the CAE selected different sets of most informative features in different runs. This observation motivated us to hypothesize that a feature appearing in multiple runs of CAE (mrCAE) carries the most meaningful information for a given dataset.

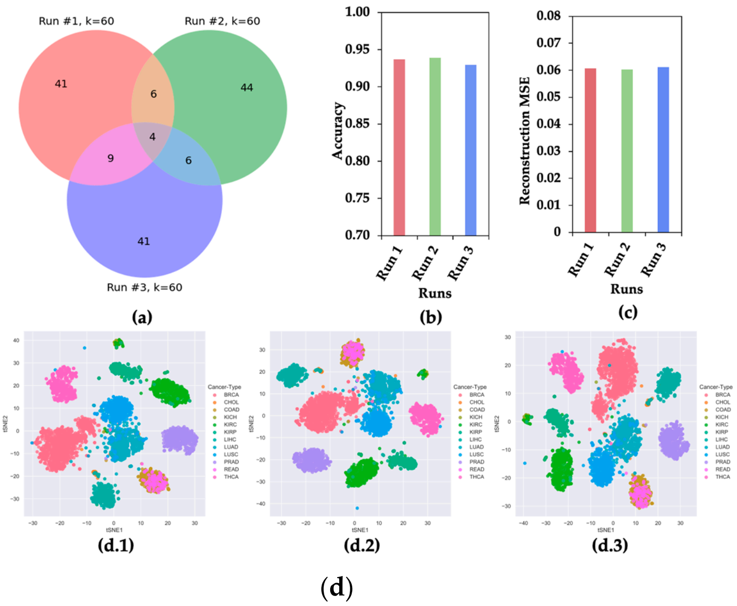

Figure 2. CAE Property of Selecting Different Sets of Features in Different Runs. (**a**) Venn Diagram, (**b**) accuracy of classifying 12 cancer types, (**c**) reconstruction mean squared error (MSE), and (**d**) t-SNE plots for 12 cancer samples using three sets of 60 features selected in three runs.

2.3. Comparison of mrCAE with Existing Feature Selection Approaches

Before comparing mrCAE with the existing feature selection approaches, we evaluated the performance of a single-run CAE with a different number of selected features, which guided us regarding how many features we had to select for comparison. In Figure 3a, it is noticeable that even with a smaller number of only ten features, the average accuracy of CAE was close to 85%. There was a sharp increase in average accuracy (91%) with 20 features, followed by a slight increase (92% accuracy) with up to 60 features. Then, the curve reached a plateau. This figure suggests that selecting 40 features (before starting plateau) while using different algorithms is a good choice for comparison.

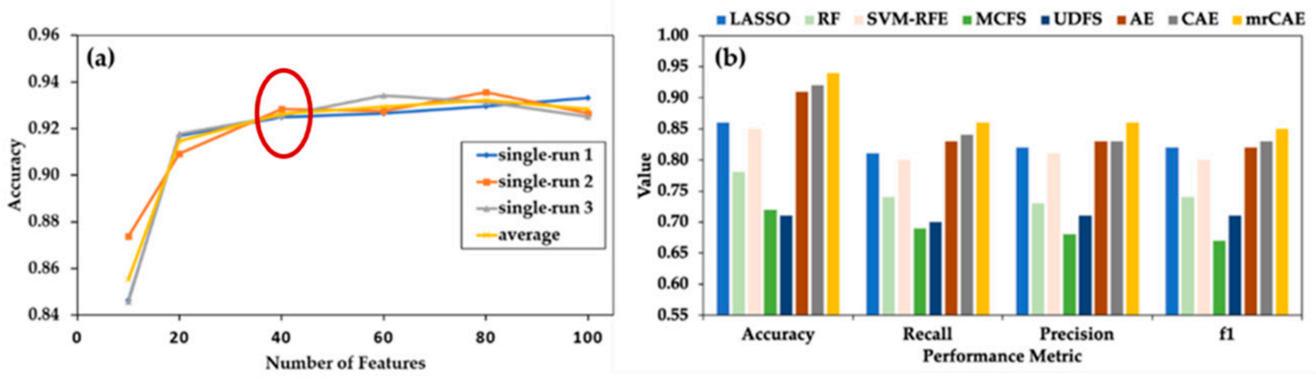

Figure 3. Comparing mrCAE with other feature selection approaches. (**a**) Behavior of single-run CAE to decide the number of features to be selected for comparison. CAE was run three times to select six sets of 10, 20, 40, 60, 80, and 100 features. "single avg" represents the average accuracy of three runs. (**b**) Classification performance using 40 features selected by LASSO, RF, SVM-RFE, MCFS, UDFS, AE, CAE and mrCAE. Note that, each approach selects 40 actual features except AE, which selectes 40 latent features.

Selection of 40 Features from mrCAE: CAE was run 100 times to select 100 features in each run. Over 100 runs, it selected a total of 534 unique features. The frequency of appearing these features in 100 runs ranged between 1 and 98. The 40 most frequent features, the top 40 features from the sorted list in descending order based on frequency, were used to measure the performance of mrCAE.

Figure 3b shows the classification performance when using the sets of 40 lncRNAs selected from the LASSO, RF, SVM–RFE, MCFS, UDFS, AE, CAE, and mrCAE feature selection algorithms. It is clear that mrCAE performed better than any other feature selection approaches in accuracy, recall, precision, and F1 score.

2.4. mrCAE to Select a Stable Set of Features

mrCAE Systems: To identify a unique and stable set of lncRNAs that not only can distinguish between 12 different cancer types but also have the highest number of features with prognostic behavior, we designed mrCAE systems with 10, 20, 40, 60, 80, 100, and 120 runs. In each of the single runs of an mrCAE system, 100 lncRNAs were selected. Table 1 shows the summary statistics of mrCAE systems, including the total number of unique lncRNAs selected and the maximum frequency of an lncRNA appearing in each mrCAE system. The minimum frequency was 1 for all the different mrCAE systems. As shown in Table 1, a total of 223 unique lncRNAs (combined list of 10 sets of 100 lncRNAs) were selected by the 10-run mrCAE system, and the frequency of an lncRNA appearing in multiple runs ranged between 1 and 10. Similarly, a total of 575 unique lncRNAs were selected by the 120-run mrCAE system, and the frequency of an lncRNA appearing in multiple runs ranged between 1 and 117.

Table 1. Summary statistics of mrCAE systems in selecting lncRNAs.

mrCAE	Total LncRNAs	Min Frequency	Max Frequency
10-run mrCAE	223	1	10
20-run mrCAE	313	1	20
40-run mrCAE	400	1	40
60-run mrCAE	464	1	60
80-run mrCAE	499	1	80
100-run mrCAE	534	1	98
120-run mrCAE	575	1	117

Frequent and Stable Features: Features appearing more than once in mrCAE system were considered frequent features. Features with higher frequencies were considered stable features.

The Top Frequent Features: The top frequent features, for example, Top-10 features in any mrCAE system, were the first ten features from the combined list sorted in descending order based on frequency. To identify a stable set of lncRNAs, we selected the top features from each of the seven mrCAE systems in six different categories: Top-10, Top-20, Top-40, Top-60, Top-80, and Top-100. Table 2 shows the ranges of frequency for the top features in six different categories. It is noticeable that the most frequent feature appeared in 10, 20, 40, 60, and 80 runs in the cases of 10-, 20-, 40-, 60- and 80-run mrCAE systems, respectively, but the trend was not maintained for 100- and 120-run systems. In other words, the most frequent feature appeared in each run of each mrCAE system except for the 100-run and 120-run systems, for which it (most frequent feature) appeared in 98 and 117 runs, respectively. It can be concluded that for the given lncRNA expression profile dataset of 12 cancers, the mrCAE system with 100 or more runs could not produce the most frequent features in each run. Thus, a 100-run mrCAE can be considered to be the optimal configuration for this dataset, and the results from 120-run mrCAE were not considered for subsequent analyses.

Table 2. Ranges of frequency for the top features in six categories.

	Ranges of Frequency					
mrCAE	**Top-10**	**Top-20**	**Top-40**	**Top-60**	**Top-80**	**Top-100**
10-run mrCAE	(10–10)	(9–10)	(6–10)	(4–10)	(3–10)	(2–10)
20-run mrCAE	(19–20)	(15–20)	(11–20)	(8–20)	(5–20)	(4–20)
40-run mrCAE	(36–40)	(29–40)	(22–40)	(15–40)	(11–40)	(8—40)
60-run mrCAE	(53–60)	(44–60)	(31–60)	(21–60)	(16–60)	(13–60)
80-run mrCAE	(69–80)	(60–80)	(42–80)	(28–80)	(22—80)	(17–80)
100-run mrCAE	(84–98)	(74–98)	(53–98)	(35–98)	(27–98)	(21–98)
120-run mrCAE	(99–117)	(85–117)	(62–117)	(44–117)	(34–117)	(25–117)

Finally, this experiment resulted in six unique sets of features corresponding to Top-10, Top-20, Top-40, Top-60, Top-80, and Top-100 features, as shown in the Venn diagram of Figure 4. For example, combining six sets of top-10 features from 10-, 20-, 40-, 60-, 80-, and 100-run mrCAE systems produced a unique list of 14 lncRNAs.

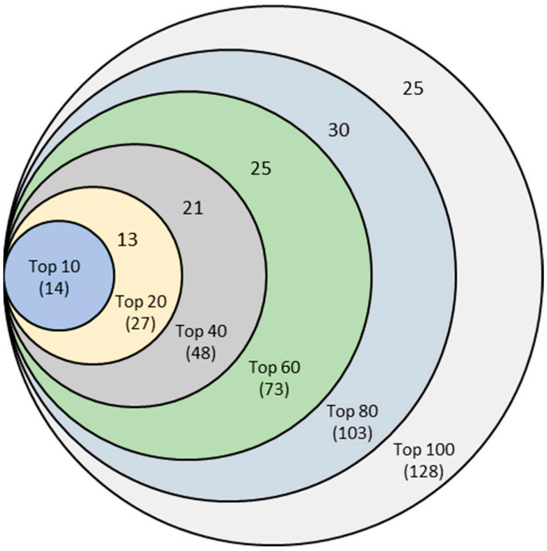

Figure 4. Venn diagram of six sets of unique features identified from six mrCAE systems. The mrCAE consisted of 10, 20, 40, 60, 80, and 100 runs. Each of these runs was conducted to select 100 features. The smallest set (light blue), containing 14 features, represents the unique features coming from six sets of 10 most frequent features from 10-, 20-, 40-, 60-, 80-, and 100-run mrCAE systems. Similarly and so on, the 2nd smallest set contains 27 (14 + 13) unique features from six sets of Top-20 features selected.

The Venn diagram shows that each set of unique features was a subset of the following more extensive unique feature set. Finally, we can conclude that the 128 unique features (*Supplementary 3*)—produced from the union of six sets of Top-100 features coming from 10-, 20-, 40-, 60-, 80-, and 100-run mrCAE systems—represented the stable and optimal feature set. We used this set of lncRNAs to conduct the downstream study, including survival and prognostic analyses and validation.

2.5. Prognostic Capability of Significant lncRNAs

To evaluate the prognostic capabilities of the selected 128 stable lncRNAs, survival analyses of patients with different cancer types were performed. Any lncRNAs with zero expression values for most of the cancer samples were excluded from the survival analysis of that cancer. The patients with values less than or equal to the median were labeled group A. Those with values greater than the median were labeled group B. After dividing into two groups, a log-rank test was conducted, and the hazard ratio was calculated as the hazard rate of group A vs. hazard rate of group B to check the prognostic capability of an lncRNA. The criteria for an lncRNA to be prognostic are log-rank test p-value ≤ 0.05 and Hazard Ratio (HR) $\neq 1.0$. Kaplan–Meier curves were plotted to show the prognostic behavior of lncRNAs.

Figure 5a shows the Kaplan–Meier plot for GATA3-AS1, one of the 11 prognostic lncRNAs for breast cancer, and Figure 5b shows the forest plot of survival analyses for 11 prognostic lncRNAs. It can be observed from Figure 5a that group B (red) had a higher rate of survival than group A (blue), meaning that lncRNA GATA3-AS1 could successfully distinguish the high-risk group (Group A) of BRCA patients from the low-risk group (Group B). In other words, the cohort with a low expression (blue) of GATA3-AS1 had a 1.53-times higher rate of death than the high-expression cohort (red). Thus, the cohorts with low-expression values for seven lncRNAs (HR > 1.0) showed higher chances of death compared to the high-expression cohorts (Figure 5b). On the other hand, the cohorts with low-expression values for four lncRNAs (HR < 1.0) showed lower chances of death compared to the high-expression cohorts (Figure 5b). *Supplementary 4* shows the forest plots for other cancer types.

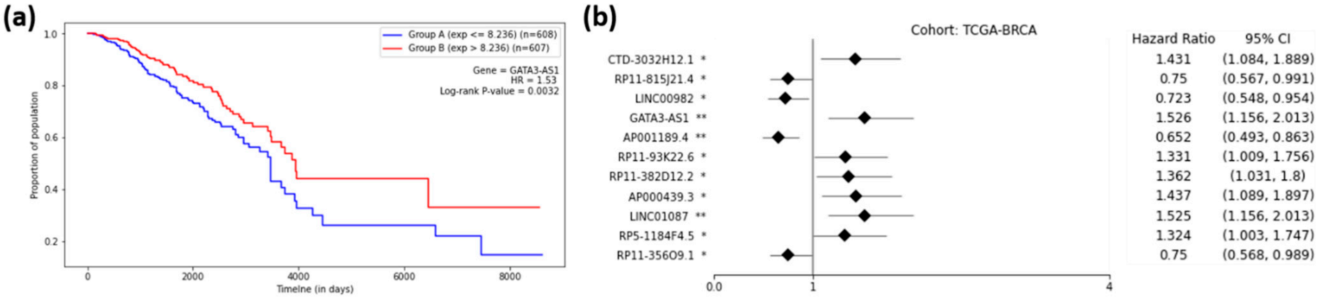

Figure 5. Survival Analysis of TCGA-BRCA. (**a**) Kaplan–Meier Curve for the GATA3-AS1 lncRNA on the TCGA-BRCA cohort. Group A (blue) is the group with an expression less than or equal to the median, and Group B (red) is the group with an expression greater than the median. (**b**) Forest plot of survival analysis for 11 prognostic lncRNAs on the BRCA cohort. The asterisks represent the log-rank p-values (*—$p \leq 0.05$, **—$p \leq 0.01$).

The number of prognostically significant lncRNAs for each type of cancer is given in Table 3. The highest number of prognostic lncRNAs were discovered for KIRC (31 lncRNAs), followed by LUAD (22 lncRNAs) and LUSC (18 lncRNAs). The proposed approach failed to discover any prognostic lncRNA for CHOL, potentially because the cohort consisted of only 36 patients (Table 1). Some of the lncRNAs were found to be prognostic for more than one cancer. Of 128 stable set of lncRNAs, 76 were found to be prognostic.

Table 3. Summary of survival analysis regarding the number of prognostic lncRNAs for each of the 12 TCGA cancer types.

BRCA	CHOL	COAD	KICH	KIRC	KIRP	LIHC	LUAD	LUSC	PRAD	READ	THCA	Total
11	0	3	3	31	15	1	22	18	4	4	10	**76**

2.6. Validations

The stable set of 128 lncRNAs derived from mrCAE was validated with the existing literature [18]. Of 128 lncRNAs, 103 were found to be known lncRNAs (*Supplementary 5*) associated with different cancer types; see Figure 6a. For example, 98 lncRNAs are associated with BRCA, 52 lncRNAs are related to LUAD, and 37 lncRNAs are related to KIRP. Some lncRNAs were also found in four different cancer hallmarks (Figure 6b and *Supplementary 6*); for example, six lncRNAs were found to be related to cancer prognosis. We also validated the top 128 lncRNAs with existing drug–lncRNA networks (*Supplementary 7*). We found that 113 out of 128 lncRNAs are associated with 24 different drugs primarily used in cancer-related treatments, as shown in Figure 6c,d. For example, the drug nilotinib is mainly used to treat a specific type of blood cancer associated with 18 different lncRNAs (Figure 6e); a drug–lncRNA network was formed based on the Spearman correlation coefficient between lncRNA expression levels and the IC50 values of the drug [19].

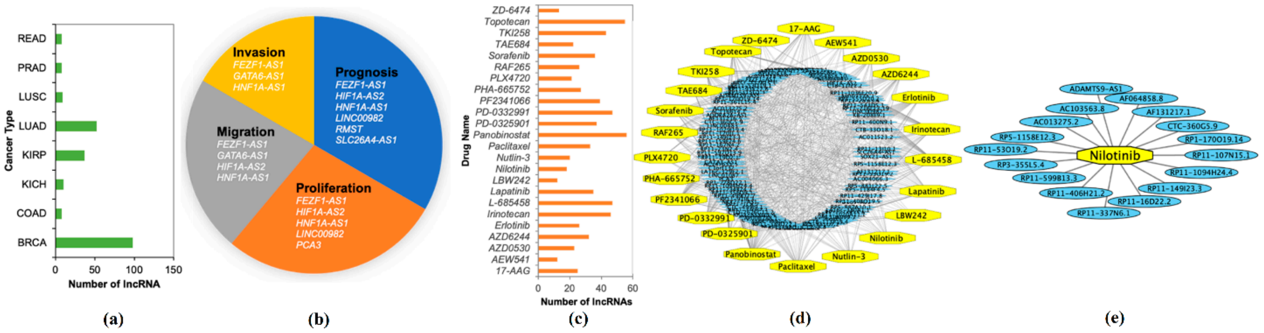

Figure 6. Validation of Identified lncRNAs. (**a**) Number of known lncRNAs derived by mrCAE related to different cancer types found in [20–23]; (**b**) mrCAE derived lncRNAs related to different cancer hallmarks [24]; (**c**) number of lncRNAs, related to different cancer drugs [19]; (**d**) drug–lncRNA networks for all 24 drugs; (**e**) an example lncRNA–drug network for nilotinib, which is used to treat certain blood cancers associated with 18 different lncRNAs. (**d**,**e**) were generated using Cytoscape.

3. Discussion

The objective of the present study was to identify significant lncRNAs that carry meaningful information on (a) identifying the origins of multiple cancer, (b) evaluating the prognostic capability of differentiating high-risk and low-risk groups of patients of particular cancers, and (c) having potential for targeted therapy. The original CAE algorithm is capable of identifying subsets of important features. However, due to the stochastic nature of the algorithm, it produces different subsets in different runs [8]. Thus, our hypothesis was that the most frequently appearing lncRNAs in multiple runs of CAE (mrCAE) would produce a biologically meaningful set of features.

Our investigation showed that the lncRNAs selected by the proposed mrCAE carry meaningful information on the prognostic capability of differentiating high- and low-risk groups of patients of particular cancers, as explained in Section 2.5. We also showed the biological relevance of the selected lncRNAs by comparing them with existing literature, drug–lncRNA networks, and hallmark lncRNAs (Figure 6).

Figure 5 shows that the lncRNAs selected by the proposed mrCAE outperformed both the single-run CAE and the standard autoencoder, along with other feature selection approaches. Thus, the current results confirmed that the proposed mrCAE could be utilized as a tool for identifying a stable set of meaningful features. It should be noted that the proposed mrCAE approach is very similar to a common bioinformatic approach of bootstrapping analysis used to evaluate the stability of results. A shortcoming of CAE is that it produces different sets of the most informative features in different runs, which makes it difficult to use in precision medicine. We propose using a multi-run CAE approach to reduce the stochasticity in CAE outcomes, i.e., to select a stable set of features. The frequent features that appear in multiple runs are considered to be the stable set of features. The bootstrapping effect could be the reason that mrCAE performs better than the CAE and standard AE.

4. Materials and Methods

4.1. Data Preparation

To characterize the cancer-associated lncRNA, expression profiles and clinical data for 33 different cancers were downloaded from the UCSC Xena database [25]. Each lncRNA expression was processed using a min–max normalization method to achieve good training performance. For this study, we considered the cancer types for which the number of normal samples was at least 10% of cancer samples, and 12 cancer types met this criterion. The distributions of cancer and normal samples for 12 cancers are shown in Table 4.

Table 4. Sample distributions of 12 cancers considered in this experiment.

	BRCA	CHOL	COAD	KICH	KIRC	KIRP	LIHC	LUAD	LUSC	PRAD	READ	THCA
Normal	113	9	41	23	72	32	50	57	49	52	9	58
Cancer	1088	36	301	65	527	286	369	510	498	493	94	501

This dataset contained about 60 thousand RNAs expression profiles, including coding genes (mRNAs) and non-coding genes (lncRNAs and miRNAs). In this study, only the expression profiles of lncRNA (n = 12,309) were considered for analysis and model evaluation. The final dataset contained 4768 cancer patients and 565 normal patients.

4.2. Features Selection Using Multi-Run Concrete Autoencoder

To select important features (lncRNAs), a state-of-the-art deep learning-based unsupervised algorithm, concrete autoencoder (CAE) [8], was iteratively run multiple times. We named this approach multi-run CAE (mrCAE). The reason for using mrCAE is that CAE selects the most informative features in a stochastic manner, meaning that different sets of informative features are selected in different runs. The assumption we made while running CAE multiple times was that if a feature appeared in more than one run, it can be considered a stable feature.

4.2.1. Architecture and Working Principle of CAE

The architecture of the CAE shown in Figure 7 consisted of a single encoding layer, also known as the feature selection layer shown in yellow, and arbitrary decoding layers (e.g., a deep feedforward neural network), shown in the box on the right. The detailed algorithm is available in [8]. The function of the encoder is to select a given number of k actual features (not latent features in the case of a traditional Autoencoder) in a stochastic manner from the original large input feature space **X** of size n. The function of the decoder is to reconstruct the original features ($\mathbf{X}'$ is the reconstructed feature vector) using the k features selected by the encoder.

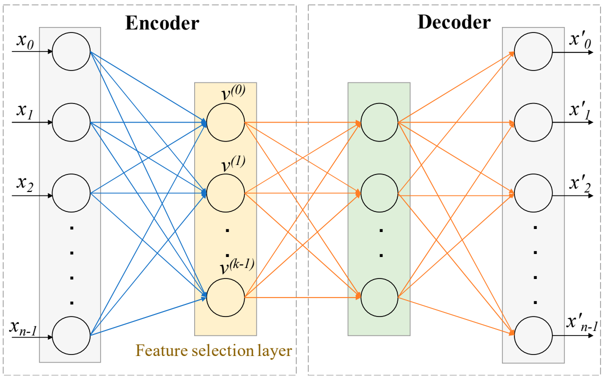

Figure 7. Architecture of Concrete Autoencoder. CAE architecture consists of an encoder and a decoder. The layer after the input layer of the encoder is called the concrete feature selection layer, as shown in yellow. This layer has k number of nodes, where each node is used for each feature to be selected. The decoder is used to check how well the input features can be reconstructed using the selected *k* features. The output layer has the same number of nodes as the input layer. $X = [x_0, x_1, \ldots ., x_{n-1}]$ = input features. $X' = [x'_0, x'_1, \ldots ., x'_{n-1}]$ = reconstructed features.

How input features are selected depends on the temperature of the selection layer, which is modulated from a high value to a small value using a simple annealing schedule [8]. As the temperature of the selection layer approaches zero, the layer selects *k* individual input features. The decoder of a concrete autoencoder serves as the reconstruction function.

It is the same as that of a standard autoencoder. Thus, the concrete autoencoder can be used to select a discrete set of k features that are optimized for an arbitrarily complex reconstruction function.

Training and Testing/Validation of CAE: The samples in a cohort were divided into 80/20 split in a stratified manner for training and testing. In the training phase, 80% of samples were used to select the k informative features. In the testing/validation phase, 20% of samples were used to reconstruct their original features using the selected k features.

4.2.2. Hyperparameter Tuning for CAE

The hyperparameters of CAE were tuned for the lncRNA expression data of 12 TCGA cancer types. We kept two of the parameters the same as those used in the original CAE developed by Abid et al. [8]. These two parameters were leaky ReLU with a threshold value of 0.1 and a 10% dropout rate. To tune the number of nodes in two hidden layers of the decoder, the model was tested by varying the number of nodes from 240 to 340 with a step size of 10. It was found that a decoder with 300 nodes in both layers yielded the highest accuracy. Thus, the number of nodes in two hidden layers of the decoder was selected to be 300.

To tune the number of epochs and learning rate, the random search [26] approach was used. For the number of epochs, we used values if 200, 300, 500, 1000, 1500, 2000, 2500, and 3000. Similarly, for the learning rate, the values were 0.001, 0.002, 0.005, 0.0005, 0.01, and 0.05. In every run of CAE, the values of the two hyperparameters were randomly selected. With 300 epochs and 0.002 learning rate, the 100 features selected by the CAE produced the highest accuracy in classifying 12 cancer types using SVM. So, these parameter values were chosen for further analysis. Details of hyperparameter tuning are available in *Supplementary 1*.

For every iteration of a single run in the hyperparameter tuning phase, the temperature, mean–max probability (mean of maximum probabilities of the selected features), training loss, and validation loss were observed and plotted. The plot painted a clear picture of the learning process in the CAE at every epoch, so we named it the characteristic plot of CAE and present it in Figure 8.

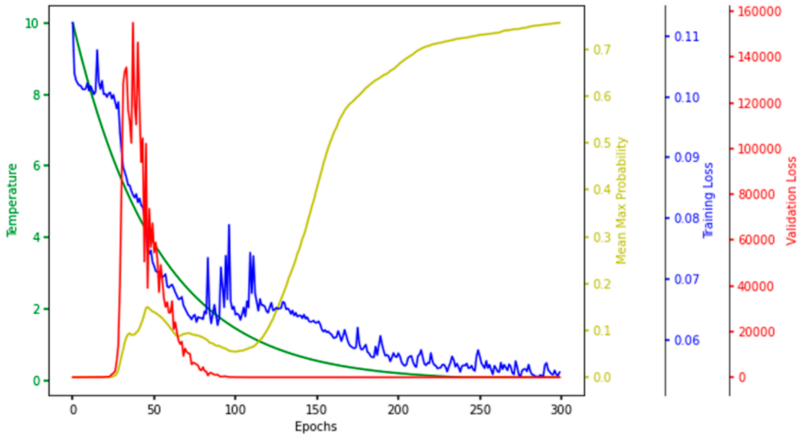

Figure 8. Characteristic Plot of Concrete Autoencoder. Temperature (green), mean–max probability (yellow), training loss (blue), and validation loss (red) are plotted at different scales.

One of the main objectives of making this plot was to see if the model converged in terms of loss, which is evident in Figure 8, which shows that the training and validation loss converged to a lower value. Each node in the concrete selection layer learned a probability value for every feature, and the node selected the one with the highest probability. The higher the mean–max probability was, the more each node in the concrete selector was confident of one of the features. So, the mean–max probability should be as high as possible.

4.3. Comparing mrCAE with Other Feature Selection Approaches

The feature selection capability of the mrCAE was compared with the standard autoencoder (AE), three frequently used embedded feature selection models (LASSO [15], random forest (RF) [16], and support vector machine with recursive feature elimination (SVM–RFE) [17]), and two unsupervised feature selection models (multi-cluster feature selection (MCFS) [27] and unsupervised discriminative feature selection (UDFS) [28]). The same numbers of features were selected using all feature selection approaches for comparison, and those features were used to evaluate the classification performance in classifying 12 different cancer types. A stratified 5-fold cross-validation using SVM with linear kernel was conducted to evaluate the classification performance. Four different evaluation metrics—accuracy, precision, recall, and f1 score—were used to record the classification performance.

4.4. Implementation of Feature Selection Algorithms

All feature selection algorithm except for mrCAE were implemented using the scikit-learn framework (https://scikit-learn.org/ Accessed: Jun'20), whereas mrCAE was implemented using a deep learning framework named Keras (https://keras.io/ Accessed: Jun'20). Experiments were parallelized on NVIDIA Quadro K620 GPU with 384 cores and 2 GB memory devices. The dataset was split into the training and testing set according in an 80/20 ratio to avoid overfitting. The training set was used to estimate the learning parameters, and the testing set was used for performance evaluation.

5. Conclusions

The authors of this study proposed a multi-run concrete autoencoder (mrCAE) to identify prognostic lncRNAs for multiple cancers. We tested the proposed model in analyzing the lncRNA expression profiles of 12 cancers. The model selected a stable set of lncRNAs that could differentiate 12 cancers with high accuracy and provide subsets of prognostic lncRNAs for 12 cancers. Though the proposed mrCAE model was applied to multiple cancers here, it can also be used on a single cancer dataset, such as when it was used to identify informative features for single-digit MNIST data by the developer of CAE.

The lncRNAs selected by the proposed mrCAE outperformed the lncRNAs selected by the single-run CAE and other feature selection approaches. Additionally, the proposed mrCAE outperformed the standard autoencoder, which selected the latent features and was thought to be the upper limit in dimension reduction. Since the proposed mrCAE outperformed AE and can select actual features in contrast to latent features by AE, it can provide meaningful information that can be used for precision medicine, such as identifying prognostic lncRNAs for different cancers. The same approach can be used in identifying salient features in other omics data.

Supplementary Materials: Supplemental Information (SI) is available at https://www.mdpi.com/article/10.3390/ijms222111919/s1.

Author Contributions: Conceptualization, A.A.M., M.S., R.B.T. and A.M.M.; methodology, A.A.M., M.S., R.B.T. and A.M.M.; software, A.A.M., M.S. and R.B.T.; validation, A.A.M., R.B.T., G.N., K.M. and G.E.H.; formal analysis, A.A.M., M.S., R.B.T., A.M.M. and G.E.H.; investigation, A.A.M., M.S. and R.B.T.; resources, A.M.M.; data curation, A.A.M.; writing—original draft preparation, A.A.M., R.B.T. and A.M.M.; writing—review and editing, M.S., A.M.M., G.N., K.M. and G.E.H.; visualization, A.A.M., M.S. and R.B.T.; supervision, A.M.M.; project administration, A.M.M.; funding acquisition, A.A.M.; All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the NSF CAREER grant, IIS 1651917 (transferred to #1901628) to AMM; NSF RAPID grant, CNS 2037374 to GN, KM, and AMM; Pilot award, NIH/NIMHD U54 MD012393 to AMM. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: https://xenabrowser.net/ Accessed: Jan'19.

Data Availability Statement: The data used in this experiment were collected from *Xena browser*.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cheetham, S.W.; Gruhl, F.; Mattick, J.S.; Dinger, M.E. Long noncoding RNAs and the genetics of cancer. *Br. J. Cancer* **2013**, *108*, 2419–2425. [CrossRef] [PubMed]
2. Fang, Y.; Fullwood, M.J. Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. *Genom. Proteom. Bioinform.* **2016**, *14*, 42–54. [CrossRef]
3. Zhang, X.; Wang, W.; Zhu, W.; Dong, J.; Cheng, Y.; Yin, Z.; Shen, F. Mechanisms and functions of long non-coding RNAs at multiple regulatory levels. *Int. J. Mol. Sci.* **2019**, *20*, 5573. [CrossRef] [PubMed]
4. Tao, H.; Yang, J.-J.; Zhou, X.; Deng, Z.-Y.; Shi, K.-H.; Li, J. Emerging role of long noncoding RNAs in lung cancer: Current status and future prospects. *Respir. Med.* **2016**, *110*, 12–19. [CrossRef]
5. Schmitt, A.M.; Chang, H.Y. Long Noncoding RNAs in Cancer Pathways. *Cancer Cell* **2016**, *29*, 452–463. [CrossRef]
6. Hanahan, D.; Weinberg, R.A. Hallmarks of Cancer: The Next Generation. *Cell* **2011**, *144*, 646–674. [CrossRef]
7. Hoadley, K.A.; Yau, C.; Hinoue, T.; Wolf, D.M.; Lazar, A.J.; Drill, E.; Shen, R.; Taylor, A.M.; Cherniack, A.D.; Thorsson, V.; et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **2018**, *173*, 291–304. [CrossRef] [PubMed]
8. Abid, A.; Balin, M.F.; Zou, J. Concrete autoencoders: Differentiable feature selection and reconstruction. In Proceedings of the 36th International Conference on Machine Learning, PMLR, San Francisco, CA, USA, 27–30 June 2019; pp. 694–711.
9. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417. [CrossRef]
10. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef]
11. Mirzaei, A.; Pourahmadi, V.; Soltani, M.; Sheikhzadeh, H. Deep feature selection using a teacher-student network. *Neurocomputing* **2020**, *383*, 396–408. [CrossRef]
12. Lu, Y.; Fan, Y.; Lv, J.; Noble, W.S. DeepPINK: Reproducible feature selection in deep neural networks. *arXiv* **2018**, arXiv:1809.01185.
13. Borisov, V.; Haug, J.; Kasneci, G. CancelOut: A Layer for Feature Selection in Deep Neural Networks. In Proceedings of the International Conference on Artificial Neural Networks, Munich, Germany, 17–19 September 2019; pp. 72–83.
14. Al Mamun, A.; Duan, W.; Mondal, A.M. Pan-cancer Feature Selection and Classification Reveals Important Long Non-coding RNAs. In Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2020), Seoul, Korea, 16–19 December 2020; pp. 2417–2424. [CrossRef]
15. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Ournal R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
16. Genuer, R.; Poggi, J.-M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [CrossRef]
17. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
18. Chen, J.; Zhang, J.; Gao, Y.; Li, Y.; Feng, C.; Song, C.; Ning, Z.; Zhou, X.; Zhao, J.; Feng, M.; et al. LncSEA: A platform for long non-coding RNA related sets and enrichment analysis. *Nucleic Acids Res.* **2021**, *49*, D969–D980. [CrossRef]
19. Li, Y.; Li, L.; Wang, Z.; Pan, T.; Sahni, N.; Jin, X.; Wang, G.; Li, J.; Zheng, X.; Zhang, Y.; et al. LncMAP: Pan-cancer Atlas of long noncoding RNA-mediated transcriptional network perturbations. *Nucleic Acids Res.* **2018**, *46*, 1113–1123. [CrossRef]
20. Cui, T.; Zhang, L.; Huang, Y.; Yi, Y.; Tan, P.; Zhao, Y.; Hu, Y.; Xu, L.; Lin, Z.; Wang, D. MNDR v2.0: An updated resource of ncRNA-disease associations in mammals. *Nucleic Acids Res.* **2018**, *46*, D371–D374. [CrossRef] [PubMed]
21. Chen, G.; Wang, Z.; Wang, D.; Qiu, C.; Liu, M.; Chen, X.; Zhang, Q.; Yan, G.; Cui, Q. LncRNADisease: A database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* **2013**, *41*, D983–D986. [CrossRef] [PubMed]
22. Ning, S.; Zhang, J.; Wang, P.; Zhi, H.; Wang, J.; Liu, Y.; Gao, Y.; Guo, M.; Yue, M.; Wang, L.; et al. Lnc2Cancer: A manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.* **2016**, *44*, D980–D985. [CrossRef]
23. Zhou, B.; Zhao, H.; Yu, J.; Guo, C.; Dou, X.; Song, F.; Hu, G.; Cao, Z.; Qu, Y.; Yang, Y.; et al. EVLncRNAs: A manually curated database for long non-coding RNAs validated by low-throughput experiments. *Nucleic Acids Res.* **2018**, *46*, D100–D105. [CrossRef]
24. Wang, J.; Zhang, X.; Chen, W.; Li, J.; Liu, C. CRlncRNA: A manually curated database of cancer-related long non-coding RNAs with experimental proof of functions on clinicopathological and molecular features. *BMC Med. Genom.* **2018**, *11*, 29–37. [CrossRef] [PubMed]
25. Goldman, M.; Craft, B.; Brooks, A.; Zhu, J.; Haussler, D. The UCSC Xena Platform for cancer genomics data visualization and interpretation. *BioRxiv* **2019**. [CrossRef]
26. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.

27. Cai, D.; Zhang, C.; He, X. Unsupervised feature selection for Multi-Cluster data. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 24–28 July 2010; pp. 333–342.
28. Yang, Y.; Shen, H.T.; Ma, Z.; Huang, Z.; Zhou, X. L2, 1-Norm Regularized Discriminative Feature Selection for Unsupervised Learning. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011.

Explainable Machine Learning to Identify Patient-specific Biomarkers for Lung Cancer

Masrur Sobhan
Knight Foundation School of Computing and Information Sciences
Florida International University
Miami, Florida, USA
msobh002@fiu.edu

Ananda Mohan Mondal
Knight Foundation School of Computing and Information Sciences
Florida International University
Miami, Florida, USA
amondal@fiu.edu

2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995516

***Abstract— Background:* Lung cancer is the leading cause of death compared to other cancers in the USA. The overall survival rate of lung cancer is not satisfactory even though there are cutting-edge treatment methods for cancers. Genomic profiling and biomarker gene identification of lung cancer patients may play a role in the therapeutics of lung cancer patients. The biomarker genes identified by most of the existing methods (statistical and machine learning based) belong to the whole cohort or population. That is why different people with the same disease get the same kind of treatment, but results in different outcomes in terms of success and side effects. So, the identification of biomarker genes for individual patients is very crucial for finding efficacious therapeutics leading to precision medicine. *Methods*: In this study, we propose a pipeline to identify lung cancer class-specific and patient-specific key genes which may help formulate effective therapies for lung cancer patients. We have used expression profiles of two types of lung cancers, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), and Healthy lung tissues to identify LUAD- and LUSC-specific (class-specific) and individual patient-specific key genes using an explainable machine learning approach, SHaphley Additive ExPlanations (SHAP). This approach provides scores for each of the genes for individual patients which tells us the attribution of each feature (gene) for each sample (patient). *Result*: In this study, we applied two variations of SHAP - tree explainer and gradient explainer for which tree-based classifier, XGBoost, and deep learning-based classifier, convolutional neural network (CNN) were used as classification algorithms, respectively. Our results showed that the proposed approach successfully identified class-specific (LUAD, LUSC, and Healthy) and patient-specific key genes based on the SHAP scores. *Conclusion*: This study demonstrated a pipeline to identify cohort-based and patient-specific biomarker genes by incorporating an explainable machine learning technique, SHAP. The patient-specific genes identified using SHAP scores may provide biological and clinical insights into the patient's diagnosis.**

Keywords— explainable machine learning, lung cancer, patient-specific biomarkers, precision medicine

I. Introduction

Cancer is a disease in which some cells of the body grow uncontrollably and spread to other organs of the body. Cancer is a genetic disease that is caused by the changes in the genes which control the cell's function, especially the growth and division of cells [1]. Three different kinds of genes are responsible for cancer - proto-oncogenes, tumor suppressor genes, and DNA repair genes [2], [3]. There are more than 100 types of cancers, but carcinomas are the most common type of cancer [1].

Lung cancer is the second most prevalent type of cancer [4], and it is the leading cause of death related to cancer in the United States [5]. There are mainly two types of lung cancer - non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) [6]. Two subtypes of NSCLC are lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). There have been many studies where lung cancer biomarkers were identified. Some studies identified race-related biomarkers [7]–[9]. The researchers also applied various well-known and novel machine learning and deep learning techniques for feature selection and classification of lung cancers and other cancer types [9]–[11]. But they have mostly used machine learning and deep learning models as "black boxes." Recently, researchers have been using various approaches to explain the black box models. Several methods have been proposed to make the machine learning and deep learning models explainable, including Shapley Sampling [12], Relevance Propagation [13], LIME [14], ANCHOR [15], and DeepLIFT [16]. But it is not clear how these methods are related and which method to select for a particular problem. To overcome this issue, Lundberg and Lee developed a unified framework for interpreting predictions, SHAP [17]. Recently there has been adequate work to explain the machine learning models using SHAP. Levy et al. used SHAP to discover important methylation states in different cell types and cancer subtypes [18]. In a more recent study, SHAP was used to explain the deep learning model which classified the cancer tissues using RNA-sequence data [19]. Most of the studies identified the global features using SHAP values which represent the average impacts of the genes on that model [20].

Researchers also use various statistical tools, such as DESeq2 [21], edgeR [22], or LIMMA [23] to identify biologically significant genes or differentially expressed genes (DEGs) [24]–[26] from differential gene expression (DGE) analysis by comparing patient cohort with healthy cohort. The DGE analysis helps to identify potential genes associated with the pathogenesis and prognosis of lung cancer [27]. The study [27] developed an integrated approach for identifying genes associated with pathogenesis and prognosis from four different sets of DEGs from four different cohorts of lung cancer patients and corresponding normal cohorts, which means that DEGs are cohort-dependent biomarker genes and do not reflect the patient-specific heterogeneity. A recent study [28] used DGE analysis to find African American (AA) and European

American (EA) cohort-based lung cancer biomarkers where they showed that using principal component analysis (PCA), AA genes are able to distinguish the normal and tumor group of AA lung cancer cohort. But surprisingly, the same AA genes are also able to distinguish the normal and tumor group of EA lung cancer cohort. This observation suggests that this cohort-based study failed to discover biomarkers for a particular cohort. Another recent study [29] also used DGE analysis to find biomarkers for lung cancer using two sets of datasets- tumor and normal samples for non-treatment studies, and cell lines after treatment and cell lines before treatment for treatment studies. The hypothesis of this study is the up-regulated genes in non-treatment studies should be down-regulated in treatment studies and vice versa. But the authors found two different sets of Biomarkers without any common genes which implies that this cohort-based study failed to discover expected biomarker. Researchers also used genome wide association studies (GWAS) to find the biomarker. In one of the studies researchers found two key loci 15q25 and 5p15 for AA cohort [30]. Another study also found eighteen key loci including 15q25 and 5p15 [31]. From these two studies, we can conclude that these GWAS studies failed to identify cohort-based biomarkers. Researchers also used machine learning-based feature selection algorithms [32]–[34] to identify biomarkers for pan cancer classification which do not belong to any cancer cohort or any specific patient. These studies (DGE analysis, GWAS, and Machine Learning-based feature selection) are similar to the population-based studies where the aim is to find cohort-based genetic changes. . As a result, the same treatment provided to the patients with the same cancer type shows different outcomes among the patients [35]. This is because each patient has unique combination of genetic changes and specific genetic changes require specific treatments. That is why it is necessary to identify the patient-specific biomarkers, which we can accomplish by identifying local interpretable features by explaining the machine learning models. The patient-specific biomarkers can be used for targeted therapy leading to precision medicine which the earlier computational studies fail to identify.

We hypothesize that biomarker genes may express differently in different patients due to the variability of mutations of genes for which cohort-based gene therapy may not be beneficial to most of the patients. To solve this issue, identifying patient-specific biomarker genes is very crucial and it may aid in precision medicine or personalized medicine. In this study, we developed a pipeline to discover global and local NSCLC-associated genes using an explainable machine learning tool, SHAP. This study identified both class-specific and patient-specific genes based on SHAP scores by calculating global and local SHAP scores, respectively. To our knowledge, there has not been any study identifying lung cancer patient-specific genes using SHAP.

The later part of this paper is ordered as follows. The "Materials and Methods" section includes the cohort analysis, preparation of the dataset, and the methods used for the research. The "Experimental Results" section provides the outcome of the research and analysis of the results. We briefly discussed our result in the "Discussion" section. Finally, conclusions and the future scope is discussed in the "Conclusion" section.

II. Materials and Methods

A. Workflow of the study

The overall workflow of this study is shown in Fig. 1.

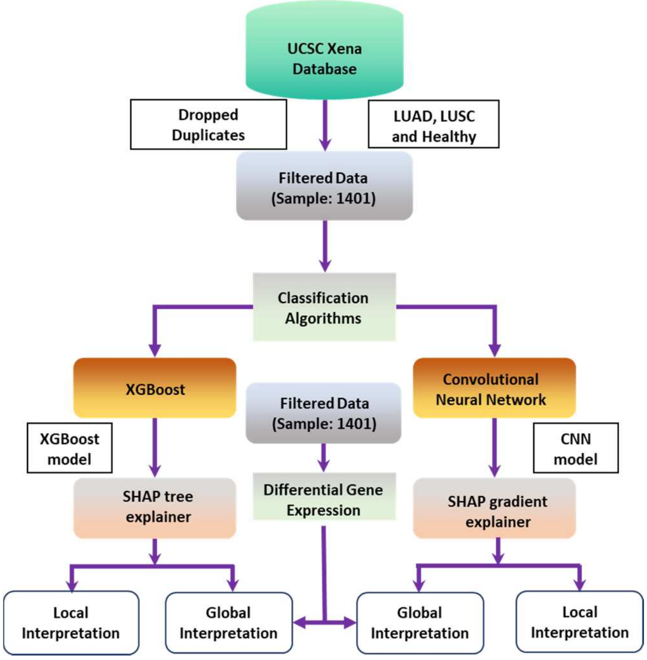

Fig. 1. Workflow of the study to identify patient-specific and class-specific genes.

The overall workflow of this study is as follows. At first, the lung cancer tumor (LUAD and LUSC), and healthy tissue samples were downloaded from the UCSC Xena database. Next, the dataset was filtered by dropping the duplicate records of the same patients having the same tumor type. The filtered data were used to classify three different classes (LUAD, LUSC, and healthy) using two different algorithms- XGBoost and CNN. Hyperparameters were tuned to achieve a higher classification accuracy. 5-fold cross-validation was performed to measure the performance of the two algorithms. Then the two models from two different genres of classification algorithms – XGBoost from tree-based and CNN from deep learning-based classifiers - were used for interpretation using SHAP. As such, we used the tree explainer technique for the XGBoost and the gradient explainer for the CNN model for interpretation. Next, we analyzed the two different interpretation techniques to get class-specific genes and patient-specific genes. We also used a statistical tool DESEq2 to get the important genes across the populations.

B. Data Collection and Cohort Analysis

To characterize the lung-cancer-associated mRNA, the expression profiles and clinical data associated with lung cancer were collected from the UCSC Xena database [36]. The normal tissue samples were downloaded from the Xena database and the mskcc GitHub repository [37], [38]. There are 1415 samples, including 503 LUAD, 489 LUSC, and 423 healthy, as shown in Table I.

TABLE I. Sample distribution and cohort analysis of lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC) and healthy samples

LUAD Tumor Samples	LUSC Tumor Samples	Healthy Samples
503	**489**	**423**
Total= 1415		

C. *Data Preparation*

Fourteen of 1415 samples were duplicates. We kept only one record of the same patients. So, the final cohort size for this analysis was 1401 with 492 LUAD tumors, 486 LUSC tumors, and 423 healthy samples, respectively. We used the dataset with FPKM values which were already log-normalized. The raw gene count dataset was also considered in this study. The data distribution of the three categories is well distributed and there is little chance of bias towards the larger group. The final dataset consists of 1401 samples with 19,648 mRNA expression values. Then we used this dataset to classify LUAD, LUSC, and healthy using a tree-based machine learning algorithm and a deep learning algorithm.

D. *Classification Algorithms*

We used two algorithms in our analysis - Extreme Gradient Boosting (XGBoost) [39] and Convolutional Neural Network (CNN) [40]. XGBoost is a decision tree-based machine learning algorithm that uses a process called boosting to help improve performance. It is an optimized gradient-boosting algorithm through parallel processing, tree pruning, handling missing values, and regularization to avoid bias or overfitting.

CNN is a deep neural network primarily used in image classification or computer vision applications. But it has also wide applications in analyzing tabular data. The convolution layers extract features from the samples. A small filter or kernel scans through the samples and extracts features from the samples. The following layer is the pooling layer which down-samples the feature map extracted by the convolution layer. The pooling layer runs a filter across the feature map and takes the specific information from that filter. It translates the features' exact spatial information to latent information. The final pooling layer is then flattened out and transformed into a one-dimensional array and fed to the fully connected layers that predict the output.

The samples of each class were divided into 80/20 split in a stratified manner for training and testing respectively. 5-fold cross-validation was used for measuring the classification performance. For the stratification of the samples, we used StratifiedKFold from the scikit-learn library. Hyperparameters were also tuned to get optimized results from both XGBoost and CNN classifiers.

Next, the contribution of all the features of individual samples for the two classifiers was determined. We wanted to identify the reasons for the machine learning models' success or accuracy. Feature contributions, both globally and locally, can decipher the models' accomplishment. That is why we applied an explainable machine learning tool that can identify the feature contribution that caused the models' success.

E. *Global and Local Feature Interpretation*

Global Feature Interpretation

The global features are a set of features that reflect the average behavior of a cohort of samples or patients. For global feature interpretation, we used two techniques: (a) DESeq2, a statistical tool, and (b) SHAP (SHapley Additive exPlanations) a game theoretic approach. DESeq2 is a tool for differential gene expression analysis of RNA-seq data. It provides a list of important genes for a cohort of patients, which reflects the average or global impact of genes across the cohort.

SHAP is a game theoretic approach to explain the output of any machine learning model. It takes the machine learning or deep learning algorithms into account and then calculates a score for each feature. The first step to calculate the SHAP score is taking the differences in the model's prediction with and without a feature from all the coalition sets. Then taking the average of all the values from each of the coalition sets provides the SHAP score. In short, the average marginal contribution of a feature value across all possible coalitions is the SHAP score. The collective SHAP values can show how much each predictor or feature contributes, either positively or negatively, to the target variable or output of the model. The collective SHAP values refer to the global features of the dataset.

Local Feature Interpretation

Local features are a set of features that reflect the characteristics or behavior of an individual sample or patient. Along with identifying global important features, SHAP identifies local important features as well. Each sample for each feature or predictor gets its SHAP value. It increases transparency by calculating the contributions of the predictors. Traditional feature importance or selection algorithms produce results across the entire population, not on each individual. The idea of local interpretability of SHAP was used for identifying patient-specific genes which may help devise the strategy for personalized treatment.

III. Experimental Result

A. *Classification Accuracy*

The dataset was divided into 80/20 split for training and testing. Also, 5-fold cross-validation was performed to measure the performance of the models. The testing accuracy of algorithms from 5 folds was measured and then the average was calculated to finalize the accuracy. Table II summarizes the results of 5-fold cross-validation. The testing accuracy of XGBoost and CNN were 96.3% and 92.6%, respectively.

TABLE II. Results of 5-Fold Cross-Validation. First Row: Distribtuiion of Actual Labeled data; Second row: Distribtuiion of correctly predicted data; Third row: average classification accuracy of 5-fold cross-validation.

	XGBoost			CNN		
	LUAD	LUSC	Healthy	LUAD	LUSC	Healthy
Actual Data	492	486	423	492	486	423
Correct Prediction	473	453	423	468	450	379
Accuracy	96.3%			92.6%		

B. Differential Gene Expression Analysis

Differential gene expression (DGE) analysis was performed using the statistical tool DESEq2. Raw counts of gene expression value were used in this analysis. We used 492 LUAD tumor samples, 486 LUSC tumor samples, 59 healthy tissues from LUAD patients, and 51 healthy tissues from LUSC patients. We ran the DGE analysis tool on LUAD and LUSC samples separately to get the most important lung cancer subtype (LUAD and LUSC) specific genes. These genes represent the average behavior of the population related to the subtypes (LUAD and LUSC). These genes can also be named global features as they represented the average importance of the cohorts. We identified LUAD-specific differentially expressed genes (LUAD-DEGs) and LUSC-specific differentially expressed genes (LUSC_DEGs) based on the thresholds |log2Fold-change| >3 and adjusted p-value < 0.001, which provided us 1,037 and 1,773 genes, respectively.

C. Global Interpretability using SHAP

We used the explainable machine learning tool, SHAP, to identify the important genes by leveraging XGBoost and CNN classifier models. The important genes were compared with the differential gene expression genes derived from the DESeq2 tool discussed in section 'B'. SHAP and DESeq2 tools both were used to identify the important genes across the population.

In our analysis total number of features (genes) used for XGBoost and CNN algorithms was 19,648. SHAP generates a shapely score for each gene for each patient. The scores were then averaged across the samples of correctly classified classes. Thus, we got three sets of genes (LUAD-specific, LUSC-specific, and healthy-specific) with scores. We sorted the genes of each class based on the shapely values. Both XGBoost and CNN generated 5 different models because of five-fold validation. For global interpretation, we considered the average of the five models' output (five sets of test data from 5-fold) from XGBoost and CNN. Next, we took the top 1037 genes from LUAD and 1773 genes from LUSC class each, the same as the number of DEGs. The top genes of LUAD and LUSC classes were compared with LUAD-DEGs and LUSC-DEGs, respectively. From the analysis we noticed that the tree explainer leveraging the XGBoost model and gradient explainer for the CNN classifier model were able to identify a significant number of global genes for both LUAD and LUSC classes which are shown in Fig. 2. From the figure it is clear that XGBoost model identified 89 LUAD and 214 LUSC common genes with LUAD DEGs and LUSC DEGs respectively. Whereas CNN only identifies 68 LUAD common and 218 LUSC common genes.

<u>Optimal Genes for global interpretation</u>

To find the optimal number of genes for global interpretation, we ran four classifiers- three variants of SVM (linear, rbf and polynomial) and logistic regression with different set of top genes. To identify the top genes, at first, the genes were sorted in a descending order based on SHAP score and then picked up the important genes. Genes having higher SHAP score are considered as the important genes. For example, top 25 genes indicate the most important 25 genes from each of the classes (LUAD, LUSC and Healthy). The criteria to select optimal number of genes was to find a minimal number of genes with high accuracy. We found out that SVM rbf and SVM polynomial are not good classifiers for the three classes. Logistic regression and SVM linear were good at classifying the three classes using the top genes. But unfortunately, SVM linear failed to classify using top 25 genes. Logistic regression and SVM linear showed that the classification accuracies were high using top 50 genes. Thus, for this study we chose top 50 genes as the optimal number of genes for global interpretation. This scenario is shown in Fig. 3.

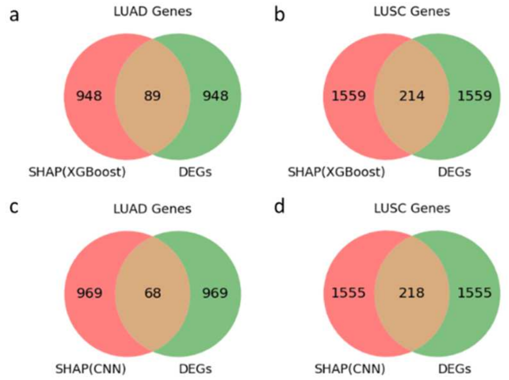

Fig. 2. Venn diagram of SHAP genes and DEGs. (a) and (c) represent the SHAP genes and DEGs for LUAD tumor. (b) and (d) represent the SHAP genes and DEGs for LUSC tumor.

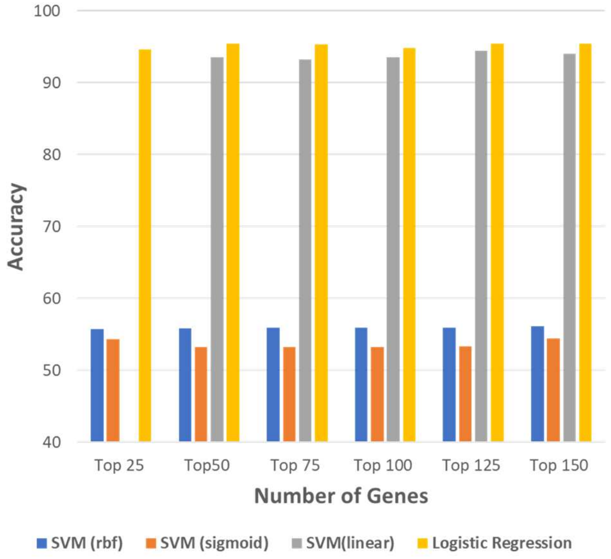

Fig. 3. Classification accuracy of four classifiers using top genes to find optimal set of genes for global interpretation . Top 50 genes from each classes (LUAD, LUSC and Healthy) are the optimal number of genes for global interpretation as top 50 genes has a minimum number of genes with high accuracy.

Next, we examined whether the top 50 genes are truly class-specific genes. If these genes are truly class-specific then there must be few overlaps among the three groups (LUAD, LUSC, and Healthy). This scenario is shown in Fig. 4 (a). We considered the top 50 genes from LUAD, LUSC, and Healthy samples separately. We found that there is no common gene among the three sets derived from both XGBoost and CNN. There are very few or no common genes when considering two of the three classes. Also, t-SNE plot shows that, using the top 50 genes from three classes, there are three clusters for the three different classes shown in Fig. 4(b). Thus, we can conclude that the identified top 50 genes for three classes are truly class-specific. Fig. 4. only represents the genes identified by tree explainer. Similar scenarios were observed for the genes identified by gradient explainer.

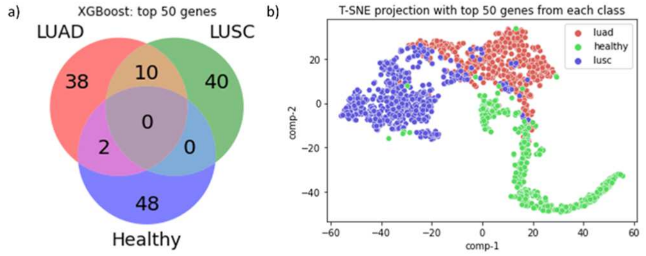

Fig. 4. Venn diagram and t-SNE plot of class-specific genes. (a) Top 50 genes of LUAD, LUSC and Healthy from tree-explainer shows mimum overlap among the genes (b) Top 50 genes of LUAD, LUSC and Healthy from tree-explainer shows three clusters for three classes.

SHAP also provides us the information on important genes that contributed most to the model along with its shapely scores and class impact of the genes. Fig. 5 shows the top 10 genes that contributed most to the XGBoost model (CNN is not shown). It also provides information about the class-specific impact of the genes. For example, ACVRL1 (gene) contributed most to both healthy class and model output. TP63 contributed most to the LUSC class and slightly contributed to LUAD class. This means that the TP63 gene could be an important biomarker for LUSC. Similarly, we can say that GOLM1 is an important biomarker gene for LUAD.

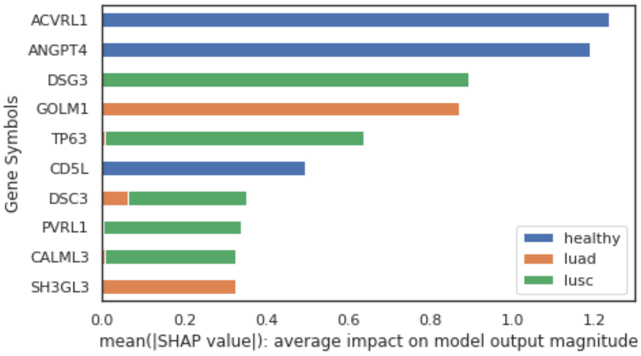

Fig. 5. Barplot of top 10 genes. The X-axis is the mean SHAP values scored by XGBoost. The values indicate the average score of the model output for the genes. Blue, orange, and green color represent three different classes-healthy, luad, and lusc. The Y-axis represents the top gene symbols determined by the tree explainer.

D. *Local Interpretability using SHAP*

We identified the most salient genes from the XGBoost and Convolutional neural network (CNN) model using SHAP for each gene and each sample. This level of local interpretability helped to identify patient-specific biomarkers which may be used as personalized medicine or therapy. To get the scores for each of the genes and samples, we trained both XGBoost and CNN with 80% of the data and tested with the rest 20% of the data. We followed this procedure five times and in each of the cases, there was a new 20% of the data in the testing set, thus providing 100% of data after testing. But XGBoost and CNN have 96.3% and 92.6% accuracy respectively which indicates that there are few false predictions. Next, we discarded the false predicted samples and kept only the true prediction. Out of 1,401 samples, the numbers of correctly classified samples were 1,349 and 1,297 for XGBoost and CNN, respectively. So, each of the samples has all the genes scored based on shapely values. Next, we sorted all the genes in descending order based on the score. Fig. 6 shows the most important genes for a single patient. This figure is a force plot for a particular LUSC tumor patient. The predicted SHAP score of this sample is 6.23 where the base value is 0.8992. This score indicates that the expression values of the genes for this patient have a higher influence on the model. The base value is the average of the model output of LUSC class. The red arrow indicates that the genes pushed the model score higher and the blue arrow indicates the genes that pushed the model score lower. From the gene expression values, we also see that DSG3 has a high expression value and SLC4A4 has a low expression value thus the former is red and the latter is blue.

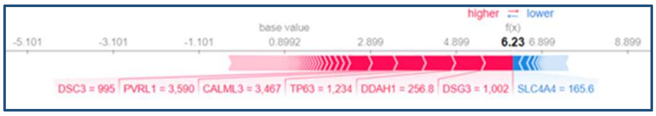

Fig. 6. Force plot of a single LUSC patient. The numerical values along with the genes are the expression values for this patient. This plot shows the most important genes for this particular patient.

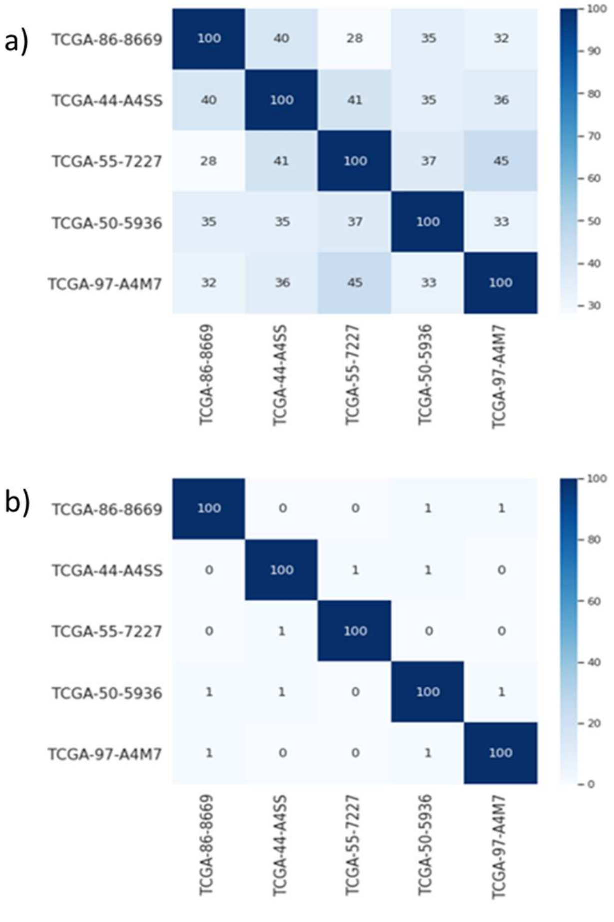

Fig. 7. Heatmap of LUAD patients with corresponding top 100 genes. (a) Heatmap of 100 genes derived from tree explainer (XGBoost model) across 5 LUAD patients. (b) Heatmap of 100 genes derived from gradient explainer (CNN model) across 5 LUAD patients.

Next, we tried to interpret the patient-specific genes of each of the samples. We wanted to make sure whether these genes are really patient-specific or not. To prove it we considered randomly chosen five LUAD and five LUSC samples. For each of the patients, we picked the top 100 genes based on the SHAP score (higher SHAP-scored genes were chosen). Our hypothesis was that if these genes are really patient-specific then there should be very few overlapping genes as each individual has different mutations of genes and different expression profiles. To validate this hypothesis, we plotted a heatmap for LUAD samples which is shown in Fig. 7. From the figure, we see that there are very few overlapping genes from the tree explainer output leveraging the XGBoost model. On the other hand, the gradient explainer was able to find totally unique genes or almost zero overlapping genes among the five patients. The same scenario was observed with the LUSC patients as well (not shown). Also, the heatmap was plotted across all the patients and very few overlapping genes were found. This indicates that, even though these samples are coming from the same class, SHAP was able to score patient-specific genes.

Next, we hypothesized that there should be many overlapping genes in the healthy samples. This is because there should be very few mutations of genes as the tissue samples are not affected by the tumor. Again, we plotted a heatmap with randomly chosen 5 healthy patients shown in Fig. 8. From the heatmap, it is evident that there are lots of overlapping genes which proves our hypothesis.

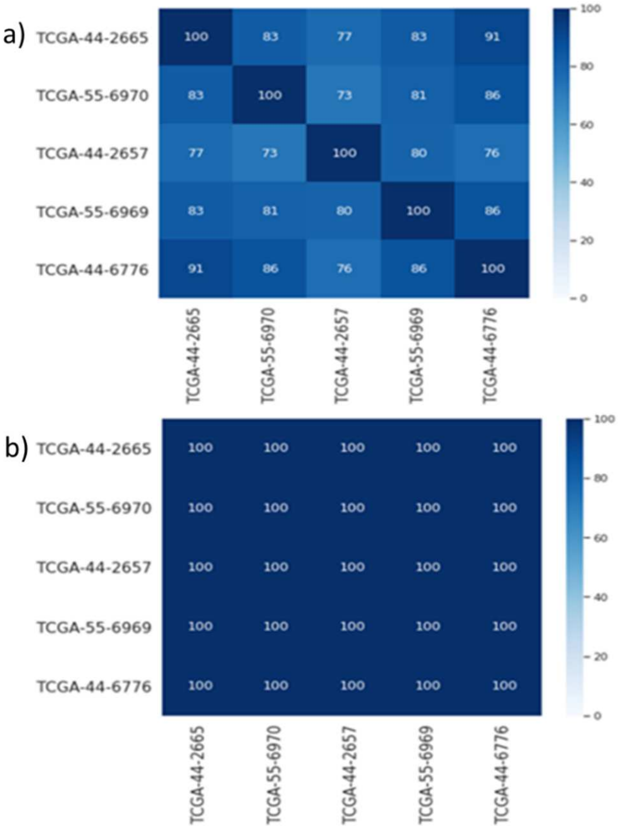

Fig. 8. Heatmap of healthy samples with corresponding top 100 genes. (a) Heatmap of 100 genes derived from tree explainer (XGBoost model) across 5 healthy samples. (b) Heatmap of 100 genes derived from gradient explainer (CNN model) across 5 healthy samples.

IV. Discussion

Most of the prior machine learning and deep learning works were involved in cancer classification and the algorithms were used as a "black box." But recently a few algorithms like- SHAP, LIME, ANCHOR, DeepLIFT, etc. algorithms have been introduced to explain the black box. In this study, we used a tree-based algorithm, XGBoost, and a deep learning classifier CNN to classify the two types of lung cancer (LUAD and LUSC) and Healthy cohorts. Then the models generated by the classifiers were used in SHAP to explain the output of the models. SHAP is a unified approach to explaining machine learning models which addresses the limitations of the black box models by explaining local and global features. We used two different explainers- a tree explainer for the XGBoost model and a gradient explainer for the CNN model. Tree explainer is a fast and exact method to estimate SHAP values for the tree models. Gradient explainer is another kind of SHAP explainer that can handle neural network models. In this study, we tried to address an important task that may play a vital role in the field of healthcare, personalized medicine, by adopting the proposed pipeline.

SHAP is able to identify global features that explain the impact of the model output on the whole population. To identify whether the SHAP explainability model was able to identify plausible features, we compared the output of the two explainers with the differential gene expression (DGE) analysis tool DESEq2 output. The DGE tool was used as the reference to assess the correctness of the predicted genes from the explainers. Unlike DESEq2, there is no standard approach for selecting SHAP features (genes). That is why we ranked the genes based on the SHAP values and considered the only top-ranked genes to compare with differentially expressed genes (DEGs). The outputs of both the explainers had some common genes with the output DGE analysis. We also tried to find out the common genes among the three classes (LUAD, LUSC, and healthy) and found very few genes overlapping among the two classes, and none of the genes overlapped among the three classes. It tells us that SHAP was able to identify biologically significant class-specific genes.

One of the greatest challenges in healthcare is to identify patient-specific important biomarkers which can aid in personalized medicine. In this study, we addressed this issue by explaining the local interpretability of SHAP output. SHAP scores were assigned to every gene of every sample leveraging the modification of the game theoretic approach. So, each of the genes of every sample consists of a SHAP score which is then ranked based on the score. To explain the local interpretability, we considered the top 100 genes of each patient. We tried to find out the common genes among the samples of the same classes and found that tree explainer output has very few common genes across the samples, whereas gradient explainer has almost zero overlapping genes across the samples. It tells us that SHAP can identify patient-specific important genes in the tumor classes (LUAD and LUSC) as the tumor is more likely to work differently in different patients. We also noticed that there are lots of overlapping genes across the healthy samples. It is understandable because there is no mutation or few genomic alterations in the patients.

V. Conclusion

Majority of previous studies identified only cohort-based important genes or population-based important genes. But it was observed that different patients require different kinds of treatment for the same disease due to the various genomic alterations and mutations. In this study, we addressed two important issues of therapeutics- the identification of subtype-specific (class-specific) and patient-specific genes. To solve these issues, we developed a pipeline that can identify both subtype-specific and patient-specific genes leveraging SHAP scores. For this analysis, we used RNA-seq data of lung cancer to show that SHAP was able to identify both class-specific and patient-specific genes. This study shows that SHAP can be used to find many biological insights by identifying local (patient-specific) and global (class-specific) genes which may help to develop better therapeutics for individual patients.

All the output shown in this analysis is machine learning and deep learning-based computational outcome. These outcomes should be verified in the wet lab to strongly validate our result. If they can be verified in the wet lab, the pipeline can be used to identify important genes for any type of disease.

Acknowledgment

The work has been supported by the NSF CAREER (#1901628) to AMM.

References

[1] "What Is Cancer? - National Cancer Institute." https://www.cancer.gov/about-cancer/understanding/what-is-cancer (accessed Apr. 11, 2022).

[2] G. Mendiratta, E. Ke, M. Aziz, D. Liarakos, M. Tong, and E. C. Stites, "Cancer gene mutation frequencies for the U.S. population," Nat. Commun. 2021 121, vol. 12, no. 1, pp. 1–11, Oct. 2021.

[3] P. A. Futreal et al., "A CENSUS OF HUMAN CANCER GENES," Nat. Rev. Cancer, vol. 4, no. 3, p. 177, 2004.

[4] "Lung and Bronchus Cancer — Cancer Stat Facts." https://seer.cancer.gov/statfacts/html/lungb.html (accessed Oct. 28, 2020).

[5] "Common Cancer Sites — Cancer Stat Facts." https://seer.cancer.gov/statfacts/html/common.html (accessed Oct. 28, 2020).

[6] K. Inamura, "Lung Cancer: Understanding Its Molecular Pathology and the 2015 WHO Classification," Front. Oncol., vol. 7, no. AUG, p. 193, Aug. 2017.

[7] S. D. Stellman et al., "Lung cancer risk in white and black Americans," Ann. Epidemiol., vol. 13, no. 4, pp. 294–302, 2003.

[8] J. D. Campbell et al., "Comparison of Prevalence and Types of Mutations in Lung Cancers Among Black and White Populations," JAMA Oncol., vol. 3, no. 6, pp. 801–809, Jun. 2017.

[9] M. Sobhan, A. Al Mamun, R. B. Tanvir, M. J. Alfonso, P. Valle, and A. M. Mondal, "Deep Learning to Discover Genomic Signatures for Racial Disparity in Lung Cancer," Proc. - 2020 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2020, pp. 2990–2992, 2020.

[10] B. Shin et al., "Cascaded Wx: A novel prognosis-related feature selection framework in human lung adenocarcinoma transcriptomes," Front. Genet., vol. 10, no. JUN, pp. 1–11, 2019.

[11] S. Tian, H. H. Chang, and C. Wang, "Weighted-SAMGSR: Combining significance analysis of microarray-gene set reduction algorithm with pathway topology-based weights to select relevant genes," Biol. Direct, vol. 11, no. 1, pp. 1–15, 2016.

[12] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," Knowl. Inf. Syst., vol. 41, no. 3, pp. 647–665, Nov. 2014.

[13] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," PLoS One, vol. 10, no. 7, p. e0130140, Jul. 2015.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," NAACL-HLT 2016 - 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Demonstr. Sess., pp. 97–101, Feb. 2016.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," Accessed: Apr. 11, 2022. [Online]. Available: www.aaai.org.

[16] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," 34th Int. Conf. Mach. Learn. ICML 2017, vol. 7, pp. 4844–4866, Apr. 2017.

[17] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," Adv. Neural Inf. Process. Syst., vol. 2017-December, pp. 4766–4775, May 2017.

[18] J. J. Levy, A. J. Titus, C. L. Petersen, Y. Chen, L. A. Salas, and B. C. Christensen, "MethylNet: An automated and modular deep learning approach for DNA methylation analysis," BMC Bioinformatics, vol. 21, no. 1, pp. 1–15, Mar. 2020.

[19] M. Yap et al., "Verifying explainability of a deep learning tissue classifier trained on RNA-seq data," Sci. Reports 2021 111, vol. 11, no. 1, pp. 1–12, Jan. 2021.

[20] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," Entropy 2021, Vol. 23, Page 18, vol. 23, no. 1, p. 18, Dec. 2020.

[21] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," Genome Biol., vol. 15, no. 12, pp. 1–21, Dec. 2014.

[22] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," Bioinformatics, vol. 26, no. 1, p. 139, Nov. 2010.

[23] G. K. Smyth, "Limma: Linear Models for Microarray Data."

[24] E. Porcu et al., "Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome," Nat. Commun. 2021 121, vol. 12, no. 1, pp. 1–9, Sep. 2021.

[25] N. Shriwash, P. Singh, S. Arora, S. M. Ali, S. Ali, and R. Dohare, "Identification of differentially expressed genes in small and non-small cell lung cancer based on meta-analysis of mRNA," Heliyon, vol. 5, no. 6, Jun. 2019.

[26] J. M. Xue, Y. Liu, L. H. Wan, and Y. X. Zhu, "Comprehensive Analysis of Differential Gene Expression to Identify Common Gene Signatures in Multiple Cancers," Med. Sci. Monit., vol. 26, pp. e919953-1, Feb. 2020.

[27] M. Ni et al., "Identification of candidate biomarkers correlated with the pathogenesis and prognosis of non-small cell lung cancer via integrated bioinformatics analysis," Front. Genet., vol. 9, no. OCT, pp. 1–14, 2018.

[28] K. A. Mitchell, A. Zingone, L. Toulabi, J. Boeckelman, and B. M. Ryan, "Comparative Transcriptome Profiling Reveals Coding and Noncoding RNA Differences in NSCLC from African Americans and European Americans," Clin. Cancer Res., vol. 23, no. 23, pp. 7412–7425, Dec. 2017.

[29] M. Maharjan, R. B. Tanvir, K. Chowdhury, W. Duan, and A. M. Mondal, "Computational identification of biomarker genes for lung cancer considering treatment and non-treatment studies," BMC Bioinformatics, vol. 21, no. 9, pp. 1–19, Dec. 2020.

[30] K. A. Zanetti et al., Genome-wide association study confirms lung cancer susceptibility loci on chromosomes 5p15 and 15q25 in an African-American population, vol. 98. 2016.

[31] J. D. McKay et al., "Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic

susceptibility across histological subtypes," Nat. Genet. 2017 497, vol. 49, no. 7, pp. 1126–1132, Jun. 2017.

[32] A. Al Mamun, W. Duan, and A. M. Mondal, "Pan-cancer Feature Selection and Classification Reveals Important Long Non-coding RNAs," Proc. - 2020 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2020, pp. 2417–2424, Dec. 2020.

[33] A. Al Mamun et al., "Multi-run Concrete Autoencoder to Identify Prognostic lncRNAs for 12 Cancers," 2021.

[34] A. Al Mamun, M. Sobhan, R. B. Tanvir, C. J. Dimitroff, and A. M. Mondal, "Deep Learning to Discover Cancer Glycome Genes Signifying the Origins of Cancer," Proc. - 2020 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2020, pp. 2425–2431, Dec. 2020.

[35] W. Lee, D. S. Huang, and K. Han, "Constructing cancer patient-specific and group-specific gene networks with multi-omics data," BMC Med. Genomics, vol. 13, no. 6, pp. 1–12, Aug. 2020.

[36] "UCSC Xena." https://xenabrowser.net/datapages/ (accessed Apr. 16, 2022).

[37] Q. Wang et al., "Unifying cancer and normal RNA sequencing data from different sources," Sci. Data 2018 51, vol. 5, no. 1, pp. 1–8, Apr. 2018.

[38] "GitHub - mskcc/RNAseqDB." https://github.com/mskcc/RNAseqDB (accessed Sep. 20, 2022).

[39] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-August-2016, pp. 785–794, Mar. 2016.

[40] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," Nov. 2015.

Quantifying Intratumor Heterogeneity by Key Genes Selected Using Concrete Autoencoder

Raihanul Bari Tanvir(✉), Ricardo Ruiz, Samuel Ebert, Masrur Sobhan, Abdullah Al Mamun, and Ananda Mohan Mondal

Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, FL 33199, USA
{rtanv003,rruiz101,seber007,msobh002,mmamu009,amondal}@fiu.edu

Abstract. The tumor cell population in cancer tissue has distinct molecular characteristics and exhibits different phenotypes, thus, resulting in different subpopulations. This phenomenon is known as Intratumor Heterogeneity (ITH), a major contributor to drug resistance, poor prognosis, etc. Therefore, quantifying the levels of ITH in cancer patients is essential, and many algorithms do so in different ways, using different types of omics data. DEPTH2 algorithm utilizes transcriptomic data to assess ITH scores and exhibits promising performance. However, it quantifies ITH using all genes, limiting the identification of ITH-related prognostic genes. We hypothesize that a subset of key genes is sufficient to quantify the ITH level, and this subset of key genes could be ITH-related prognostic genes. To prove our hypothesis, we propose an unsupervised deep learning-based framework using Concrete Autoencoder (CAE) to select a subset of cancer-specific key genes for ITH evaluation. For the experiment, we used gene expression profile data of breast, kidney, and lung cancer tumor cohorts from the TCGA repository. Multi-run CAE identified three sets of key genes for each cancer cohort. Comparing ITH scores derived from all genes and CAE-selected key genes showed similar prognostic outcomes. Subtypes of lung cancer displayed consistent ITH distributions for both gene sets. Based on these observations, it can be concluded that a subset of key genes, instead of all, is sufficient for ITH quantification. Our results also showed that many key genes are prognostically significant and can be used as therapeutic targets.

Keywords: Concrete Autoencoder · Deep Learning · Gene Expression · Intratumor Heterogeneity · ITH

1 Introduction

Intratumor Heterogeneity (ITH) refers to different types of tumor cell subpopulations within a tumor [1]. Even though these cell subpopulations have the same origin (tumor tissue, patient), they exhibit different phenotypes and molecular characteristics. ITH is one of the main challenges for targeted cancer therapy, as the difference in tumor cells and their microenvironments makes it harder for targeted cancer therapy to eradicate cancer cells [2, 3]. Therefore, an accurate assessment of ITH is essential to understand

P. Maji et al. (Eds.): PReMI 2023, LNCS 14301, pp. 844–852, 2023.
https://doi.org/10.1007/978-3-031-45170-6_88

the tumor dynamics and the development of effective and durable therapeutic strategies. ITH causes can vary depending on different levels, such as the genome, epigenome, transcriptome, etc. [4]. For example, reduced DNA damage mechanisms, microenvironmental factors (hypoxia, acidosis, etc.) [5], subclonal evolution [2], etc., contribute to ITH at the genomic level. The methylation of tumor suppressor genes is an example of ITH at the epigenomic level [6]. Different gene expression patterns contribute to ITH at the transcriptome level, which is observed to mirror ITH at the genomic or epigenomic level or both [5, 7]. This makes transcriptomic data suitable for quantifying ITH.

Different algorithms for quantifying ITH exist, such as ABSOLUTE [8], MATH [9], EXPANDS [10], and PhyloWGS [11]. These algorithms use genomic data, such as – copy number alterations (CNA), somatic mutation profiles, etc. Some algorithms take advantage of transcriptome profile that mirrors ITH at the genomic and epigenomic level, such as – tITH [12], sITH [13], DEPTH [14], and DEPTH2 [15]. In contrast to other ITH evaluation techniques, such as DEPTH and others, the DEPTH2 method assesses ITH independently of normal controls. This implies that it can be utilized for all tumor gene expression profiles regardless of the availability of corresponding normal samples' gene expression data. tITH requires protein-protein interaction (PPI) network along with gene expression data. Unlike tITH, DEPTH2 calculates the ITH score using only gene expression data.

Though the DEPTH2 method is statistically sound, the drawbacks are- (i) it uses expression values of all genes (~20,000) in calculating the ITH score and (ii) it cannot guide finding the prognostically significant genes. We argue that not all genes are related to ITH, and a subset of key genes is sufficient to calculate the ITH score at the transcriptome level.

This study presents a deep learning-based computational framework that utilizes an unsupervised concrete autoencoder (CAE) to identify key genes for quantifying Intratumor Heterogeneity (ITH). The framework selects a subset of key genes from Breast Invasive Carcinoma (BRCA), Kidney Renal Carcinoma (KIRC), and Lung Adenocarcinoma (LUAD) using expression profile data from the TCGA repository. The ITH scores are then calculated using all genes and the selected key genes. The results demonstrate that using the subset of 100 key genes outperforms all ~20,000 genes in terms of survival and prognostic outcomes for the three cancer types. The key genes exhibit consistent levels of ITH across cancer subtypes and show potential as prognostic markers and therapeutic targets. This study highlights the effectiveness of a reduced set of key genes in quantifying ITH at the transcriptome level. The overall framework is depicted in Supplementary Fig. S1.

2 Materials and Methods

2.1 Dataset Collection

We collected gene expression datasets of BRCA,LUAD, and KIRC cancers from the UCSC Xena Browser database [16]. Each dataset contains expression profiles of 20,530 mRNAs. The number of tumor samples for each cancer type was as follows: BRCA (1097 samples), LUAD (533 samples), and KIRC (517 samples).

2.2 Concrete Autoencoder to Select Cancer-Specific Key Genes

Concrete Autoencoder (CAE) [17], an unsupervised deep learning approach is used to identify cancer-specific key genes. CAE identifies features most informative for a given dataset [18–23]. CAE differs from the standard Autoencoder in the encoder part, where CAE employs a concrete selector layer (See Fig. S2). This selector layer is based on Concrete distribution [24], a relaxed variant of discrete distribution. Unlike the encoder part of the CAE, the decoder part resembles closely with the standard Autoencoder. The selector layer is used to incorporate discrete distribution into deep learning algorithms. For example, CAE uses it to learn a subset of the most informative features and produce minimum reconstruction error. In the learning phase, the selector layer learns a subset of features, which depends on a hyperparameter called Temperature (T), which is gradually lowered during the training phase to a low value using a simple annealing scheduling. This gradual decrease in temperature helps the concrete distribution to learn and select a definite subset of features [17]. In the selector layer, each unit selects a unique feature with the highest probability from the original feature space. Thus, CAE selects the most informative subset of features, and the reconstruction of the original feature space using the selected subset of features produces minimum reconstruction error. In the original Autoencoder, the features learned at the encoder part are latent features, whereas those learned at CAE are actual features. CAE was trained on each gene expression data of BRCA, KIRC, and LUAD, and 100 features were selected in each run. While training CAE, the dataset was divided randomly into 80/20 split for training and testing. Details of hyperparameter tuning are in Table S1.

2.3 Training CAE

Figure S3 shows the characteristics curve for CAE or an instance of the training behaviors of CAE for the LUAD dataset. The hyperparameter, Temperature (T), was reduced using a simple annealing schedule from 10 to 0.1 from the start epoch to the last. The reconstruction errors (loss) for the training and validation sets are plotted using blue and red curves, respectively. It shows that both errors were relatively high during the early training phase, as expected, and both reached a minimum plateau at the end. Also, the mean-max probability finally approaches 1.0 (yellow curve). The CAE was implemented using Keras (https://keras.io/). Experiments were conducted on the high-performance cluster with NVIDIA Quatro K620 GPU with 384 cores and 2 GB memory devices.

2.4 ITH Level Estimation Method

To calculate the Intratumor Heterogeneity (ITH) score, we used a scoring method named - Deviating Gene Expression Profiling Tumor Heterogeneity, or DEPTH2 in short [15], defined as –

$$\sqrt{\frac{\sum_{i=1}^{m}\left(z(ex(G_i, T)) - \frac{1}{m}\sum_{j=1}^{m} z(ex(G_j, T))\right)^2}{m-1}} \tag{1}$$

where,

$$z(ex(G_i, T) = \frac{\left|(ex(G_i, T) - \frac{1}{t}\sum_{j=1}^{t} ex(G_i, TS_j)\right|}{SD_i} \tag{2}$$

and,

$$SD_i = \sqrt{\frac{\sum_j^t \Big((ex(G_i, T) - \frac{1}{t}\sum_{j=1}^{t} ex(G_i, TS_j)\Big)^2}{t-1}} \tag{3}$$

where T is the tumor sample for which the score is being calculated. G_i is the i-th gene, and m is the number of genes. $ex(G_i, T)$ expression of gene G_i in sample T. It assigns a score to each patient. It is based on standard deviations of the z-score of the gene expression value variations. If a tumor displays similar z-scored expression values across most genes, it will have a low DEPTH2 score and a lower ITH level. In contrast, if there is variation in gene expression alterations, the tumor will receive a higher DEPTH2 score. This score indicates how much the gene expressions deviate from the norm for all tumors and genes within the matrix. We calculated the ITH score for each cancer patient of BRCA, KIRC, and LUAD employing DEPTH2 using two sets of genes. One score uses all the genes, and the other uses only the key genes selected by multi-run CAE.

2.5 Survival Analysis

Survival Analysis was performed to check whether two groups of patients based on high and low ITH scores are significantly distinguishable in prognosis. In our analysis, the event of interest is the death of cancer patients.

Survival Analysis Based on ITH Scores: Samples were sorted in descending order of the ITH score, and then the top and bottom of the total samples were taken as two groups. This analysis compared the prognostic importance of ITH scores derived using all genes and key genes (our study).

Survival Analysis Based on Each Key Gene: The cohort was divided into two groups based on the median gene expression values. This survival analysis helped identify prognostically significant genes.

After forming two groups, the Kaplan-Meier curves were plotted, and the Log-rank test was performed to check the statistical significance of the difference in survival function.

3 Results and Discussion

3.1 Multi-run CAE to Select Key Genes

Due to the stochastic nature of CAE, the model was trained ten times, and in each run, 100 features were selected for each cancer cohort - BRCA, KIRC, and LUAD. Figure 1(a) shows the stochastic nature of CAE since only 16 genes are common between three single-run CAE. In the case of 10-run CAE, the top 100 features were selected from the combined list sorted in descending order based on the frequency of appearance of a feature in 10 runs. It is clear from Fig. 1(b) that there are 53 genes common between three batches of 10 runs, which is more than the common genes (16 genes) in three single runs. Thus, a multi-run approach was adopted to select the robust set of features.

The top 100 frequent features were chosen to select the key features based on the assumption that the more frequent a feature in different runs, the more informative the feature is. The combined lists of features derived from 10-run CAE consist of 469, 527, and 435 genes for BRCA, KIRC, and LUAD, respectively. The frequency range of the top 100 features is 3 to 10 for each cancer cohort, which means that the most frequent features appeared in all ten runs, and the least frequent one appeared in 3 runs.

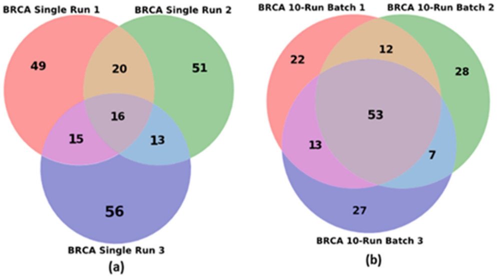

Fig. 1. Selecting the robust set of features. (a) Venn diagram of three sets of 100 genes from three single-run CAE; CAE produces only 16 features common between single runs. (b) three sets of most frequent 100 features from 10-run CAE. 10-run CAE produces more features (53 genes) common between three batches of runs. Thus, multi-run CAE produces a robust set of features.

3.2 Multi-run CAE Selects Cancer-specific Genes

We investigated whether there were any common genes between two sets or among the three sets of key genes derived from three cancers, shown by the Venn diagram in Fig. S4. It shows that there is no common gene between the three gene sets. However, a few genes are common between each pair of gene sets: 5 between BRCA and LUAD, 3 between KIRC and BRCA, and 3 between KIRC and LUAD. Since the size of each set is 100 and there are only a few genes common between two sets and none between the three sets, thus, the key gene sets are cancer-specific.

3.3 All Genes vs. Key Genes in ITH Scoring: Whole Cancer Cohorts

We compared the ITH scores calculated for BRCA, KIRC, and LUAD cohorts using two different sets of genes: (i) DEPTH2 score calculated using all genes and (ii) DEPTH2 score calculated using only the key genes selected by the multi-run CAE system (our work). Survival analysis is used to compare the two ITH scores. Figure 2 presents the results of survival analyses, Kaplan Meier plots, for cancer cohort - BRCA based on ITH scores derived from all genes (Fig. 2a) and key genes (Fig. 2b).

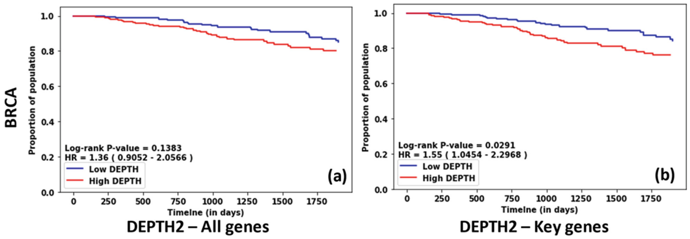

Fig. 2. Survival analysis of BRCA cohort. Kaplan Meier plots based on DEPTH2 score calculated using all genes (a) and key genes (b).

It is evident from Kaplan Meier plots that high DEPTH2 scores are related to poor prognosis, and low DEPTH2 scores have a higher chance of survival.

Survival analysis of BRCA showed a P-value of 0.1383 (not significant) and Hazard Ratio (HR) of 1.36 using all genes (Fig. 2a), while key genes produced a significant result with a P-value of 0.0291 and HR of 1.55 (Fig. 2b). The latter case is prognostically significant (P-value $\leq$ 0.05) compared to the former, thus validating our claim.

Similarly, better results were found using key genes than all genes both in LUAD (P-value: 0.0019 vs. 0.109; HR: 1.79 vs. 1.34) and KIRC (P-value: 5.18e−07 vs. 0.0018, HR: 2.67 vs. 1.78), as shown in Fig. S5a–b and S5c–d.

Our investigation showed that 100 key genes produced better results than all genes (~20,000) in three types of cancers. Thus, we do not need all genes to evaluate the ITH scores.

3.4 All Genes vs. Key Genes in ITH Scoring: LUAD Subtypes

In this section, we show the comparison of ITH scores (DEPTH2) for LUAD subtypes calculated using all genes versus key genes. Of 435 LUAD patients, 55, 34, and 54 are labeled as Terminal Respiratory Unit (TRU), Proximal Proliferation (PP), and Proximal Inflammation (PI), respectively. The remaining patients did not have any subtype-based labels. This molecular subtyping was done in [25]. It is evident from survival analysis that the TRU subtype is prognostically favorable and has a higher chance of survival than the PI and PP subtypes combined (Fig. S6).

Figure 3 shows the ITH score distribution for three subtypes, using all genes and key genes. Min-max normalization on DEPTH2 scores was performed to bring the distribution to the same scale. It is seen that the subtype TRU has comparatively lower values in ITH score than other subtypes, which supports the higher chance of survival for the TRU subtype than PI and PP combined (Fig. S6). It is also clear that the distribution of ITH scores for three subtypes remained the same for all genes and key genes.

We performed correlation analysis to compare the distribution of the DEPTH2 scores using all genes and key genes, and the results are shown in Table S2. It is observed that there is a relatively high correlation between DEPTH2 scores of each subtype of LUAD cancer using all and key genes. It is clear from the P-values (ns: not significant) in Fig. 3 that the two scores for each subtype derived using all genes and key genes are not significantly different. Both all genes and key genes produced the same level of

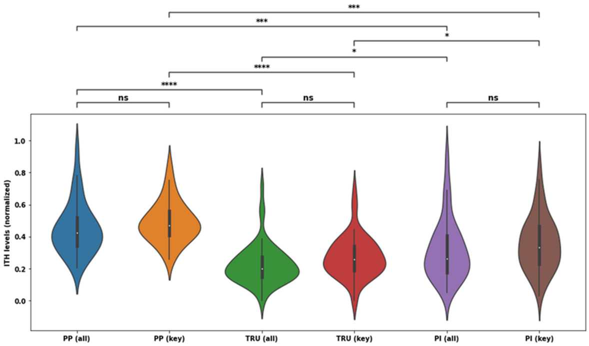

Fig. 3. Comparison of ITH scores of three molecular subtypes of LUAD using all genes (labeled as 'all') and key genes (labeled as 'key'). Distribution of Min-max normalized ITH score (DEPTH2) in three molecular subtypes of LUAD in violin plots. The Mann-Whitney-Wilcoxon test between two distributions was performed, and stars marked the p-value significance. P-value annotation legend: ns (not significant): $0.05 < p \leq 1$, *: $0.01 < p \leq 0.05$, **: $0.001 < p \leq 0.01$, ***: $0.0001 < p \leq 0.001$, ****: $p \leq 0.0001$

difference in ITH between two subtypes. For example, PP and TRU (****), TRU and PI (*), and PP and PI (***).

Based on these observations, we do not need ~20,000 genes to calculate the ITH score; only 100 key genes will suffice.

3.5 Survival Analysis of Key Genes

Survival analysis was performed on each key gene from their respective cancer cohort to identify whether they possess prognostic capabilities. Figure S5 shows the forest plot of the prognostically significant genes and the summary of survival analyses in terms of Logrank P-value and Hazard Ratio with a 95% confidence interval. The thresholds for prognostically significant genes are Logrank P-value ≤ 0.05 and Hazard Ratio, HR $\neq 1$. Of 100 key genes for BRCA, 15 were prognostically significant, as shown in the forest plot in Fig. S7. Similarly, for KIRC and LUAD, 30 and 61 genes were prognostically significant. The list of genes with prognostically significant genes marked as bold is given in Table S3.

4 Conclusion and Future Direction

This study proposes that a subset of key genes instead of all genes (~20,000) is adequate for evaluating the ITH scores of individual tumors. To test this hypothesis, a computational framework was developed using a multi-run concrete autoencoder to select the key genes from gene expression profile data. Results showed that using only the selected 100 key genes instead of all ~20,000 genes produced better survival and prognostic outcomes for three cancers (BRCA, KIRC, and LUAD). Our investigation showed that key genes produce the same levels of ITH at the cancer subtype levels. We also showed that many of these key genes are prognostically significant, which can be investigated further as possible therapeutic targets. This study concludes that a subset of key genes is sufficient to quantify the ITH at the transcriptome level.

However, this study has its limitations. The intratumor heterogeneity (ITH) is determined by genetic and epigenetic variation within an individual's tumor. The transcriptome reflects both types of heterogeneity, meaning that a unique set of genes may dictate ITH for each patient. However, our study used the same key genes to assess ITH across all patients for a specific type of tumor, which presents a limitation. The selection of 10 runs in multi-run CAE was arbitrary and may not be optimal for identifying a stable set of features for BRCA, KIRC, and LUAD cohorts. Despite these limitations, the study demonstrated that a short list of key genes is effective in assessing ITH levels. In future research, we will extend this study to determine the ideal number of runs needed to select a reliable feature set across different cohorts using multi-run CAE. Additionally, we aim to create an approach that identifies patient-specific key genes for evaluating ITH.

Acknowledgment. This work has been partially supported by the NSF CAREER Award #1901628.

Supplementary Materials. The supplementary materials are available at GitHub. https://github.com/mldag/ITH-Key-Genes-mrCAE/blob/main/supplementary.pdf.

References

1. Jamal-Hanjani, M., Quezada, S.A., Larkin, J., Swanton, C.: Translational implications of tumor heterogeneity. Clin. Cancer Res. **21**, 1258–1266 (2015)
2. Qazi, M.A., et al.: Intratumoral heterogeneity: pathways to treatment resistance and relapse in human glioblastoma. Ann. Oncol. **28**(7), 1448–1456 (2017). https://doi.org/10.1093/annonc/mdx169
3. Reinartz, R., et al.: Functional Subclone profiling for prediction of treatment-induced intratumor population shifts and discovery of rational drug combinations in human glioblastoma. Clin. Cancer Res. **23**(2), 562–574 (2017). https://doi.org/10.1158/1078-0432.CCR-15-2089
4. Grzywa, T.M., Paskal, W., Włodarski, P.K.: Intratumor and intertumor heterogeneity in melanoma. Transl. Oncol. **10**(6), 956–975 (2017). https://doi.org/10.1016/j.tranon.2017.09.007
5. Gillies, R.J., Verduzco, D., Gatenby, R.A.: Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. Nat. Rev. Cancer **12**(7), 487–493 (2012). https://doi.org/10.1038/nrc3298
6. Takamizawa, J., et al.: Reduced expression of the let-7 MicroRNAs in human lung cancers in association with shortened postoperative survival. Cancer Res. **64**(11), 3753–3756 (2004)
7. Sigalotti, L., et al.: Intratumor heterogeneity of cancer/testis antigens expression in human cutaneous melanoma is methylation-regulated and functionally reverted by 5-Aza-2′-deoxycytidine. Cancer Res. **64**(24), 9167–9171 (2004). https://doi.org/10.1158/0008-5472.CAN-04-1442
8. Carter, S.L., et al.: Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. **30**(5), 413–421 (2012). https://doi.org/10.1038/nbt.2203
9. Mroz, E.A., Rocco, J.W.: MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. Oral Oncol. **49**(3), 211–215 (2013). https://doi.org/10.1016/j.oraloncology.2012.09.007

10. Andor, N., Harness, J.V., Müller, S., Mewes, H.W., Petritsch, C.: Expands: expanding ploidy and allele frequency on nested subpopulations. Bioinformatics **30**(1), 50–60 (2014). https://doi.org/10.1093/bioinformatics/btt622
11. Deshwar, A.G., Vembu, S., Yung, C.K., Jang, G.H., Stein, L., Morris, Q.: PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. **16**, 1–20 (2015)
12. Park, Y., Lim, S., Nam, J.-W., Kim, S.: Measuring intratumor heterogeneity by network entropy using RNA-seq data. Sci. Rep. **6**(1), 37767 (2016). https://doi.org/10.1038/srep37767
13. Kim, M., Lee, S., Lim, S., Kim, S.: SpliceHetero: an information theoretic approach for measuring spliceomic intratumor heterogeneity from bulk tumor RNA-seq. PLoS ONE **14**(10), e0223520 (2019). https://doi.org/10.1371/journal.pone.0223520
14. Li, M., Zhang, Z., Li, L., Wang, X.: An algorithm to quantify intratumor heterogeneity based on alterations of gene expression profiles. Commun. Biol. **3**(1), 505 (2020). https://doi.org/10.1038/s42003-020-01230-7
15. Song, D., Wang, X.: DEPTH2: an mRNA-based algorithm to evaluate intratumor heterogeneity without reference to normal controls. J. Transl. Med. **20**(1), 150 (2022)
16. Goldman, M., et al.: The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. bioRxiv (2018)
17. Abid, A., Balin, M.F., Zou, J.: Concrete autoencoders: differentiable feature selection and reconstruction. In: 36th International Conference on Machine Learning, ICML 2019 (2019)
18. Tanvir, R.B., Sobhan, M., Mondal, A.M.: An autoencoder based bioinformatics framework for predicting prognosis of breast cancer patients. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 3160–3166 (2022)
19. Sobhan, M., Al Mamun, A., Tanvir, R.B., Alfonso, M.J., Valle, P., Mondal, A.M.: Deep learning to discover genomic signatures for racial disparity in lung cancer. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2990–2992 (2020)
20. Sobhan, M., Kalie, K., Al Mamun, A., Godavarty, A., Mondal, A.M.: Skin tone benchmark dataset for diabetic foot ulcers and machine learning to discover the salient features. In: International Conference on Image Processing, Computer Vision, & Pattern Recognition (2022)
21. Al Mamun, A., et al.: Multi-run concrete autoencoder to identify prognostic lncRNAs for 12 cancers. Int. J. Mol. Sci. **22**, 11919 (2021)
22. Al Mamun, A., Sobhan, M., Tanvir, R.B., Dimitroff, C.J., Mondal, A.M.: Deep learning to discover cancer glycome genes signifying the origins of cancer. In: Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020 (2020)
23. Al Mamun, A., Duan, W., Mondal, A.M.: Pan-cancer feature selection and classification reveals important long non-coding RNAs. In: Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020, pp. 2417–2424 (2020)
24. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: a continuous relaxation of discrete random variables. In: 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings (2017)
25. Comprehensive molecular profiling of lung adenocarcinoma. Nature (2014)

International Journal of
Molecular Sciences

MDPI

Article

MOGAT: A Multi-Omics Integration Framework Using Graph Attention Networks for Cancer Subtype Prediction

Raihanul Bari Tanvir, Md Mezbahul Islam , Masrur Sobhan , Dongsheng Luo * and Ananda Mohan Mondal *

Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, FL 33199, USA; rtanv003@fiu.edu (R.B.T.); misla093@fiu.edu (M.M.I.); msobh002@fiu.edu (M.S.)
* Correspondence: dluo@fiu.edu (D.L.); amondal@fiu.edu (A.M.M.)

Abstract: Accurate cancer subtype prediction is crucial for personalized medicine. Integrating multi-omics data represents a viable approach to comprehending the intricate pathophysiology of complex diseases like cancer. Conventional machine learning techniques are not ideal for analyzing the complex interrelationships among different categories of omics data. Numerous models have been suggested using graph-based learning to uncover veiled representations and network formations unique to distinct types of omics data to heighten predictions regarding cancers and characterize patients' profiles, amongst other applications aimed at improving disease management in medical research. The existing graph-based state-of-the-art multi-omics integration approaches for cancer subtype prediction, MOGONET, and SUPREME, use a graph convolutional network (GCN), which fails to consider the level of importance of neighboring nodes on a particular node. To address this gap, we hypothesize that paying attention to each neighbor or providing appropriate weights to neighbors based on their importance might improve the cancer subtype prediction. The natural choice to determine the importance of each neighbor of a node in a graph is to explore the graph attention network (GAT). Here, we propose MOGAT, a novel multi-omics integration approach, leveraging GAT models that incorporate graph-based learning with an attention mechanism. MOGAT utilizes a multi-head attention mechanism to extract appropriate information for a specific sample by assigning unique attention coefficients to neighboring samples. Based on our knowledge, our group is the first to explore GAT in multi-omics integration for cancer subtype prediction. To evaluate the performance of MOGAT in predicting cancer subtypes, we explored two sets of breast cancer data from TCGA and METABRIC. Our proposed approach, MOGAT, outperforms MOGONET by 32% to 46% and SUPREME by 2% to 16% in cancer subtype prediction in different scenarios, supporting our hypothesis. Our results also showed that GAT embeddings provide a better prognosis in differentiating the high-risk group from the low-risk group than raw features.

Keywords: cancer subtype prediction; graph neural network; graph attention network; multi-omics integration

check for updates

Citation: Tanvir, R.B.; Islam, M.M.; Sobhan, M.; Luo, D.; Mondal, A.M. MOGAT: A Multi-Omics Integration Framework Using Graph Attention Networks for Cancer Subtype Prediction. *Int. J. Mol. Sci.* **2024**, *25*, 2788. https://doi.org/10.3390/ijms25052788

Academic Editor: Giovanni Cugliari

Received: 1 October 2023
Revised: 15 February 2024
Accepted: 22 February 2024
Published: 28 February 2024

1. Introduction

Integrating multi-omics data is crucial for gaining a comprehensive understanding of complex diseases like Alzheimer's [1], Parkinson's [2,3], and cancer [4]. However, it is a difficult task that requires advanced computational methods. New analytical tools and methods are needed to effectively extract biologically relevant information from multi-omics data and integrate it into a comprehensive understanding of the disease. Despite the challenges of the high dimensionality and complexity of data, the integration of multi-omics data holds great potential for understanding the biology of cancer.

Graph-based learning models are used in many proposed models to get hidden representations and graph structures from different omics data. This helps us learn more about Alzheimer's, Parkinson's, cancer prediction, patient categorization, and other topics. Wang et al. [1] used multi-omics integration for Alzheimer's disease patient classification.

Researchers also used multi-omics integration to find molecular biomarkers [3] and disease classification [2] for Parkinson's disease. Much multi-omics research has been conducted to predict cancer subtypes and patient categorization. Li et al. [5] utilized a graph convolutional network (GCN) [6] to classify 28 different cancer types from pan-cancer data using gene expression and copy number alteration as features and three knowledge networks as the input graphs, including gene–gene interaction (GGI) networks, protein–protein interaction (PPI) networks, and gene co-expression networks. Zhou et al. [7] used gene expression, DNA methylation, and miRNA expression as features for multi-omics analysis. They used anchors to derive sample similarity networks and a graph convolutional autoencoder for clustering cancer samples to identify novel subtypes for breast, brain, colon, and kidney cancer. Guo et al. [8] used GCN by taking the PPI network as a graph and gene expression, copy number alterations, and DNA methylation as features. Finally, they applied attention on top of embeddings generated by GCN to classify breast cancer subtypes. Li et al. proposed the MoGCN [9], which uses an autoencoder for feature extraction and similarity network fusion to construct the patient similarity network. It applies GCN to classify breast cancer subtypes and pan-kidney cancer type classification using gene expression, copy number alterations, and phase protein array data as input. M-GCN [10] is another multi-omics framework based on GCN to classify breast and stomach cancer subtypes. They use the Hilbert-Schmidt independence criterion-based least absolute shrinkage and selection operator (HSIC LASSO) to select the molecular subtype-related transcriptomic features and then use those features to construct a patient similarity graph applying Pearson's correlation. It takes gene expression, single nucleotide variation, and copy number alterations as input for multi-omics data. MOGONET [1] inputs gene expression, DNA methylation, and miRNA. Unlike other methods, it uses GCN to learn omics-specific embeddings and uses network and node features for particular omics data. Then, it combines the embeddings using a view correlation discovery network (VCDN) to classify cancer subtypes for breast, brain, and pan-kidney cancer. The SUPREME [11] method utilizes GCN for analyzing breast cancer subtypes. It integrates seven types of data—gene expression, miRNA expression, DNA methylation, single nucleotide variation, copy number alteration, co-expression module eigengenes, and clinical data—for constructing the network and determining node features. Additionally, it combines GCN embeddings with node features and employs a multi-layer perceptron (MLP) as a classifier.

In summary, the existing GNN-based multi-omics integration approaches to predict cancer subtypes apply GCN to extract salient features from different omics data. However, GCN-based frameworks cannot determine the relative significance of neighboring samples regarding downstream analyses, including cancer subtype prediction, patient stratification, etc. It is also noticeable that none of the existing studies considered long non-coding RNA (lncRNA) expression data in multi-omics integration. However, lncRNAs play important regulatory roles in various cellular processes, including gene expression and epigenetic regulation [12–15].

This research presents MOGAT, illustrated in Figure 1, a novel multi-omics integration-based cancer subtype prediction leveraging a graph attention network (GAT) [16] model that incorporates graph-based learning with an attention mechanism for analyzing multi-omics data. The proposed MOGAT utilizes a multi-head attention mechanism that can extract information for a specific patient more efficiently by assigning unique attention coefficients to its neighboring patients, i.e., obtaining the relative influence of neighboring patients in the patient similarity graph. We also include lncRNA expression in the multi-omics integration process. Altogether, eight different data types are integrated, including mRNA expression, miRNA expression, lncRNA expression, DNA methylation, single nucleotide variation, copy number alteration, co-expression module eigengenes, and clinical data. Based on our knowledge, only one other multi-omics integration framework utilizes GAT to identify cancer driver genes but not for cancer subtype prediction [17].

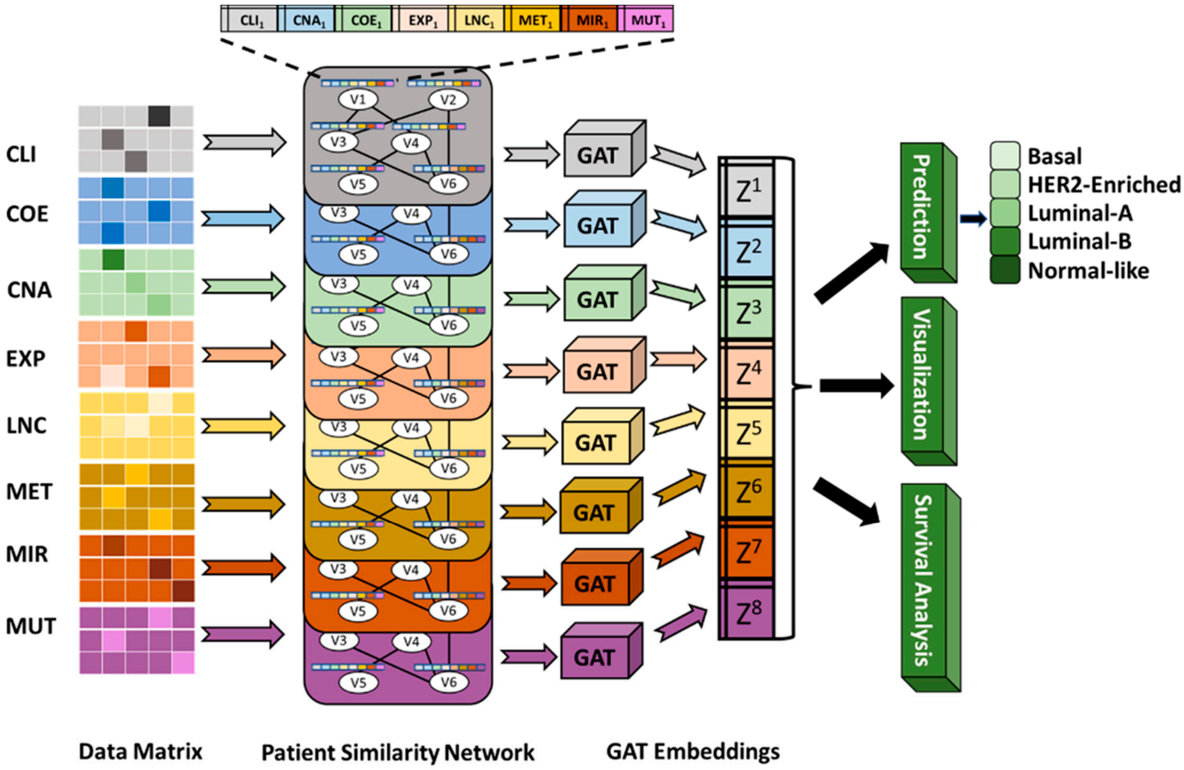

Figure 1. Illustration of the MOGAT framework. The MOGAT framework processes patient similarity networks constructed from eight datatypes, including CLI (clinical), CNA (copy number alteration), COE (co-expression), EXP (mRNA expression), LNC (lncRNA expression), MET (DNA methylation), MIR (miRNA expression), and MUT (simple nucleotide variation). Nodes in each patient similarity network are annotated with features from eight data types. By applying GAT to each patient similarity network, the framework generates embeddings. These embeddings are then used for subtype prediction, visualization, and survival analysis.

The salient features of this study are enumerated below.

- Our group is the first to explore graph attention network-based multi-omics integration for cancer subtype prediction.
- The proposed approach, MOGAT, provides better embeddings than MOGONET and SUPREME for multi-omics integration, which results in improved accuracy for cancer subtype prediction.
- MOGAT embeddings provide a better prognosis in differentiating the high-risk group from the low-risk group, which will help the physician devise an appropriate treatment strategy for an individual patient depending on the location of the patient on the prognostic curve.
- Our group is the first to incorporate lncRNA expression in multi-omics integration studies.
- We provided detailed information so that the results can be reproduced, such as (a) handling duplicate samples coming from the same patient and (b) providing the number of features in each step of preprocessing, from raw features to cleaned features to selected features.
- The interactions between different omics types are considered during the node feature engineering by concatenating features from different omics types.

2. Results

2.1. Comparison of Performance

To assess the performance of the proposed MOGAT framework, we compared it with two state-of-the-art frameworks that integrate multi-omics data for cancer subtype prediction, namely, MOGONET and SUPREME. The macro-F1 score is used to compare the performance, as illustrated in Table 1 with average (avg) and standard deviation (SD), as well as in Figure 2 with violin plots for different combinations of omics data. Three omics data were used to show the performance comparison of MOGAT with MOGONET and SUPREME: gene expression, DNA methylation, and miRNA expression, as they were originally used in MOGONET. For the same three omics data, the test macro-F1 scores on seven ($2^3 - 1$) combinations of three omics data were calculated and plotted in Figure 2a. For 3-omics analysis, we observed that (Table 1) MOGAT has higher macro-F1 scores with an average of 0.804 compared to 0.550 and 0.732 for MOGONET and SUPREME, respectively.

Table 1. Macro-F1 score to compare MOGAT with state-of-the-art multi-omics integration frameworks, MOGONET and SUPREME. The first two rows correspond to TCGA-BRCA with three omics (as shown in Figure 2a) and eight omics (as shown in Figure 2b) data. The last row corresponds to METABRIC with six omics data (as shown in Figure 2c). For each scenario, the average with standard deviation is listed. The last column shows the percentage improvement of the proposed MOGAT over MOGONET and SUPREME.

Data	Model	Avg ± SD	Improvement
TCGA: 3 omics	**MOGAT**	**0.804 ± 0. 017**	---
	SUPREME	0.732 ± 0.019	10%
	MOGONET	0.550 ± 0.145	46%
TCGA: 8 omics	**MOGAT**	**0.797 ± 0.019**	---
	SUPREME	0.686 ± 0.062	16%
METABRIC: 6 omics	**MOGAT**	**0.745 ± 0.012**	---
	SUPREME	0.733 ± 0.008	2%
	MOGONET	0.566 ± 0.056	32%

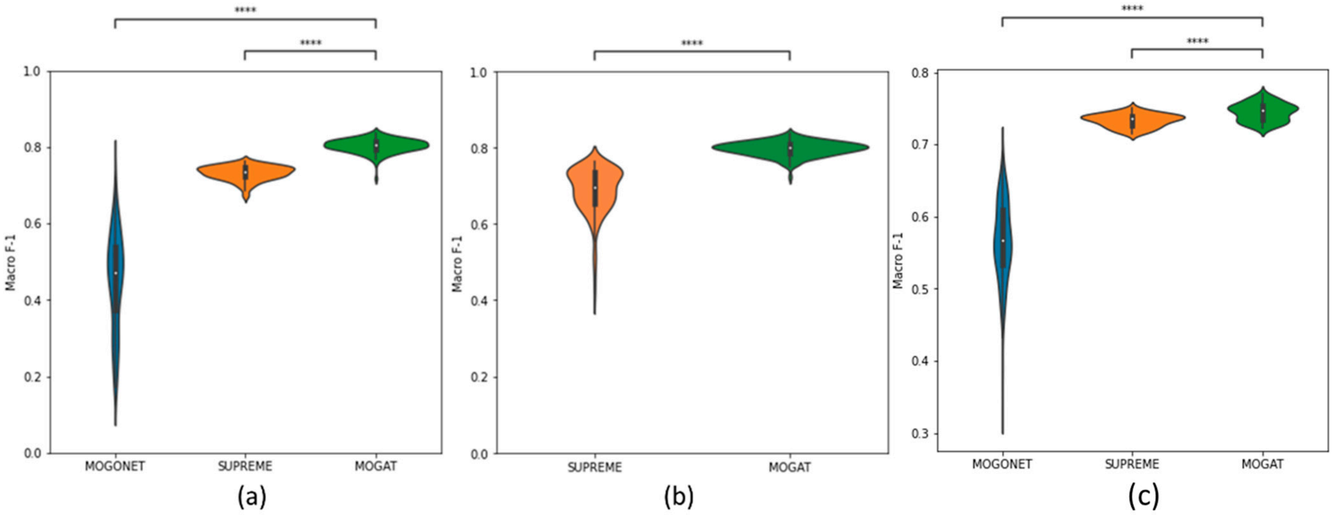

Figure 2. Macro-F1 score to compare MOGAT with state-of-the-art multi-omics integration frameworks, MOGONET and SUPREME. (**a**,**b**) with TCGA-BRCA data and (**c**) with METABRIC data. Comparison between (**a**) MOGONET, SUPREME, and MOGAT using seven ($2^3 - 1$) combinations of three omics data (EXP, MET, MIR); (**b**) SUPREME and MOGAT using 255 ($2^8 - 1$) combinations of eight omics data; (**c**) MOGONET, SUPREME, and MOGAT using 63 ($2^6 - 1$) combinations of six omics data. A pairwise statistical comparison was performed using the Mann-Whitney Wilcoxon test with a two-sided Bonferroni correction. The *p*-value annotations are—****: $p \leq 0.0001$.

However, for all the omics data, a comparison between SUPREME and MOGAT was performed, as we found that MOGONET is incompatible with eight datatypes. The macro-F1 scores of 255 ($2^8 - 1$) different combinations of eight datatypes are calculated (Table 1) and plotted using violin plots in Figure 2b. MOGAT outperforms with an average score of 0.797 compared to 0.686 for SUPREME.

For the METABRIC cohort, macro-F1 scores of 63 ($2^6 - 1$) combinations of six datatypes were calculated (Figure 2c). We observed that MOGAT has higher macro-F1 scores with an average of 0.745 compared to 0.566 and 0.732 for MOGONET and SUPREME, respectively. Overall, our proposed approach, MOGAT, outperforms MOGONET by 32% to 46% and SUPREME by 2% to 16% in cancer subtype prediction in different combinations of multi-omics data, supporting our hypothesis.

Omics-Specific Contribution in Prediction

We also investigated the contribution of each omics data type in cancer subtype prediction, and the results are shown in Table 2. Eight combinations were considered for eight different datatypes, where each combination constitutes all datatypes except the one whose contribution will be investigated. The last row shows the performance of MOGAT using all data types. The performance was estimated in terms of accuracy, weighted-F1 score, and macro-F1 score. The performance metrics were calculated from ten runs, and their mean and standard deviations are reported in Table 2. We observed that MOGAT performs better using all types of data than the other eight combinations where one data type is absent, which means that each data type contributes toward subtype prediction. It is noticeable that the performance without data type EXP (i.e., mRNA expression) is the lowest compared to the performance using all data types, with accuracy 0.837 vs. 0.861, weighted-F1 score 0.831 vs. 0.861, and macro-F1 score 0.766 vs. 0.826. This means that mRNA expression contributes the most toward subtype prediction. On the other hand, the data type MIR (i.e., miRNA expression) has the lowest contribution towards subtype prediction.

Table 2. Contribution of individual omics datatype. Performance of MOGAT with different combinations of datatypes in terms of accuracy, weighted-F1, and macro-F1 for TCGA-BRCA and METABRIC. The last row contains the result of integrating/embedding all data types. Each of the other rows contains embeddings from every datatype except the one noted in the first column.

	TCGA-BRCA			**METABRIC**		
Used Embeddings	**Accuracy**	**Weighted F1**	**Macro F1**	**Accuracy**	**Weighted F1**	**Macro F1**
All except CLI	0.842 ± 0.023	0.840 ± 0.026	0.790 ± 0.038	0.754 ± 0.02	0.755 ± 0.02	0.736 ± 0.02
All except CNA	0.837 ± 0.03	0.832 ± 0.041	0.775 ± 0.088	0.784 ± 0.012	0.782 ± 0.013	0.753 ± 0.016
All except COE	0.856 ± 0.02	0.853 ± 0.024	0.792 ± 0.044	0.790 ± 0.008	0.788 ± 0.008	0.759 ± 0.009
All except EXP	0.837 ± 0.012	0.831 ± 0.016	0.766 ± 0.041	0.775 ± 0.012	0.772 ± 0.012	0.748 ± 0.014
All except LNC	0.848 ± 0.023	0.849 ± 0.025	0.792 ± 0.043	N/A	N/A	N/A
All except MET	0.850 ± 0.013	0.847 ± 0.016	0.802 ± 0.056	0.782 ± 0.012	0.779 ± 0.012	0.75 ± 0.013
All except MIR	0.859 ± 0.013	0.856 ± 0.01	0.814 ± 0.029	N/A	N/A	N/A
All except MUT	0.848 ± 0.01	0.848 ± 0.018	0.798 ± 0.024	0.778 ± 0.024	0.776 ± 0.025	0.751 ± 0.027
All Datatypes	**0.861 ± 0.024**	**0.861 ± 0.03**	**0.826 ± 0.069**	**0.791 ± 0.009**	**0.790 ± 0.01**	**0.762 ± 0.013**

Red: lowest performance among All except single omics; Purple: highest performance among All except single omics; Bold: all omics provide best performance.

For METABRIC, it was also observed that all datatypes produce the highest performance compared to other combinations of datatypes where only one type of data is excluded. Unlike TCGA, the combination without datatype CLI (clinical) has the lowest compared to performance using all datatypes, with accuracy 0.754 vs. 0.791, weighted-F1 0.755 vs. 0.790, and macro-F1 0.736 vs. 0.762, meaning it has the highest contribution towards prediction. On the other hand, COE has the lowest contribution towards subtype prediction.

2.2. Visualization

To investigate whether the embeddings can capture the underlying insights of the data, we used principal component analysis (PCA) [18] and tSNE [19], which are dimensionality reduction techniques commonly used to reduce the high-dimensional data to a lower dimensional space to visualize the data. Figure 3 shows the PCA and tSNE plots for the learned GAT embeddings, with their counterpart raw feature matrix for TCGA-BRCA and METABIRC. We observed that for both cohorts, the embeddings learned the underlying structure of the data. The PCA and tSNE plots of GAT embeddings make it easier to tell the difference between groups of points that represent different types of breast cancer than the raw feature matrix plots that were used to train the GATs.

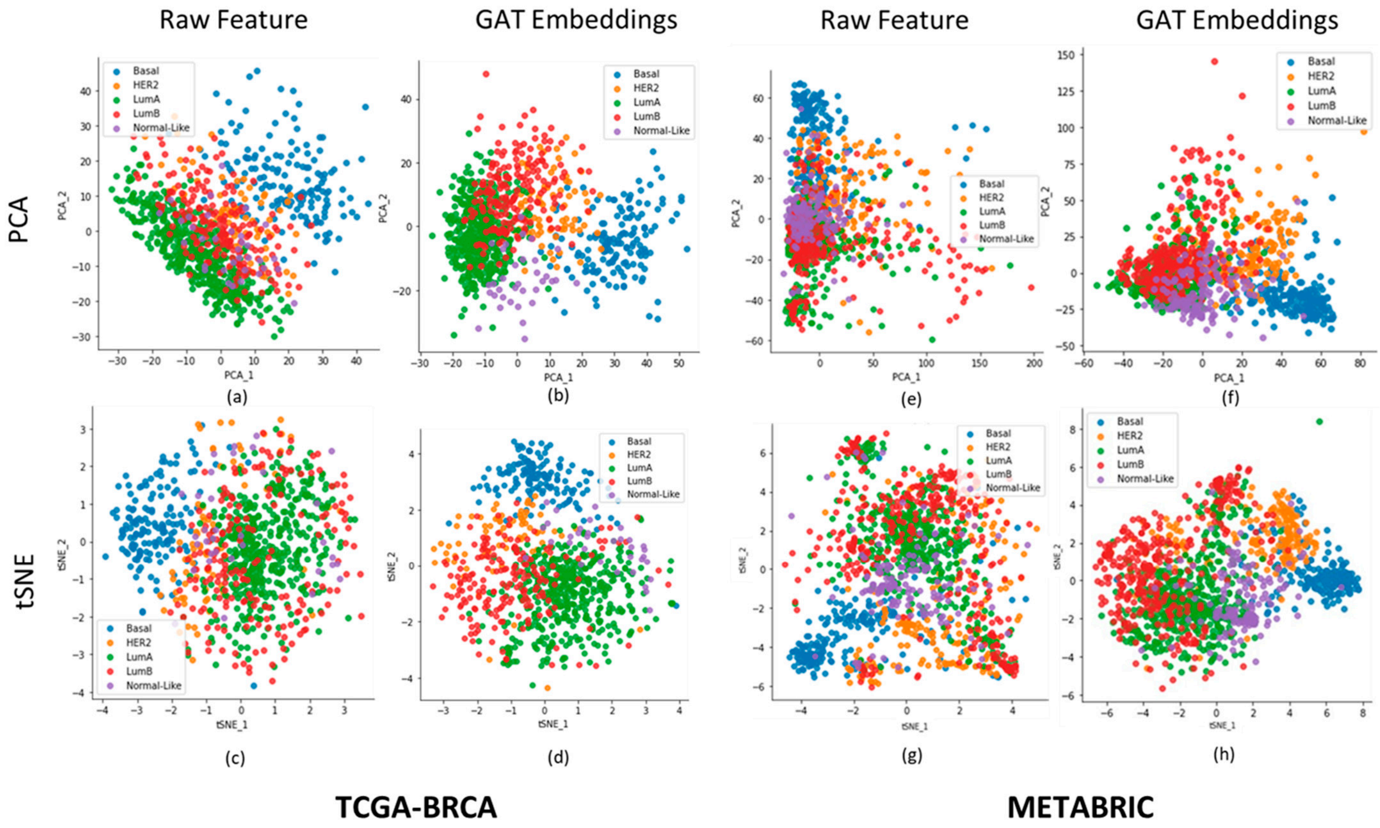

Figure 3. PCA and tSNE plots of breast cancer patients using raw features and embeddings. (**a**–**d**): for TCGA-BRCA. (**e**–**h**): for METABRIC. The *X*-axis and *Y*-axis correspond to the first and second components of PCA or tSNE.

2.3. Survival Analysis to Evaluate GAT Embeddings

Survival analysis was performed for TCGA-BRCA and METABRIC using raw features and GAT embeddings separately, following the methods described in Section 4.14, to evaluate the performance of our framework, MOGAT. The high-risk group contains patients with a risk score higher than the median, and the low-risk group has a score less than or equal to the median. The Kaplan-Meier curves using raw features and GAT embeddings are shown in Figure 4. It is observed that, in both cases, the difference in survival between high-risk and low-risk groups is significant. However, GAT embeddings can distinguish the high-risk and low-risk groups with higher significance than the raw features, as denoted by the log-rank p-value (2.10×10^{-30} vs. 7.85×10^{-3} for TCGA-BRCA and 2.03×10^{-27} vs. 2.46×10^{-16} for METABRIC).

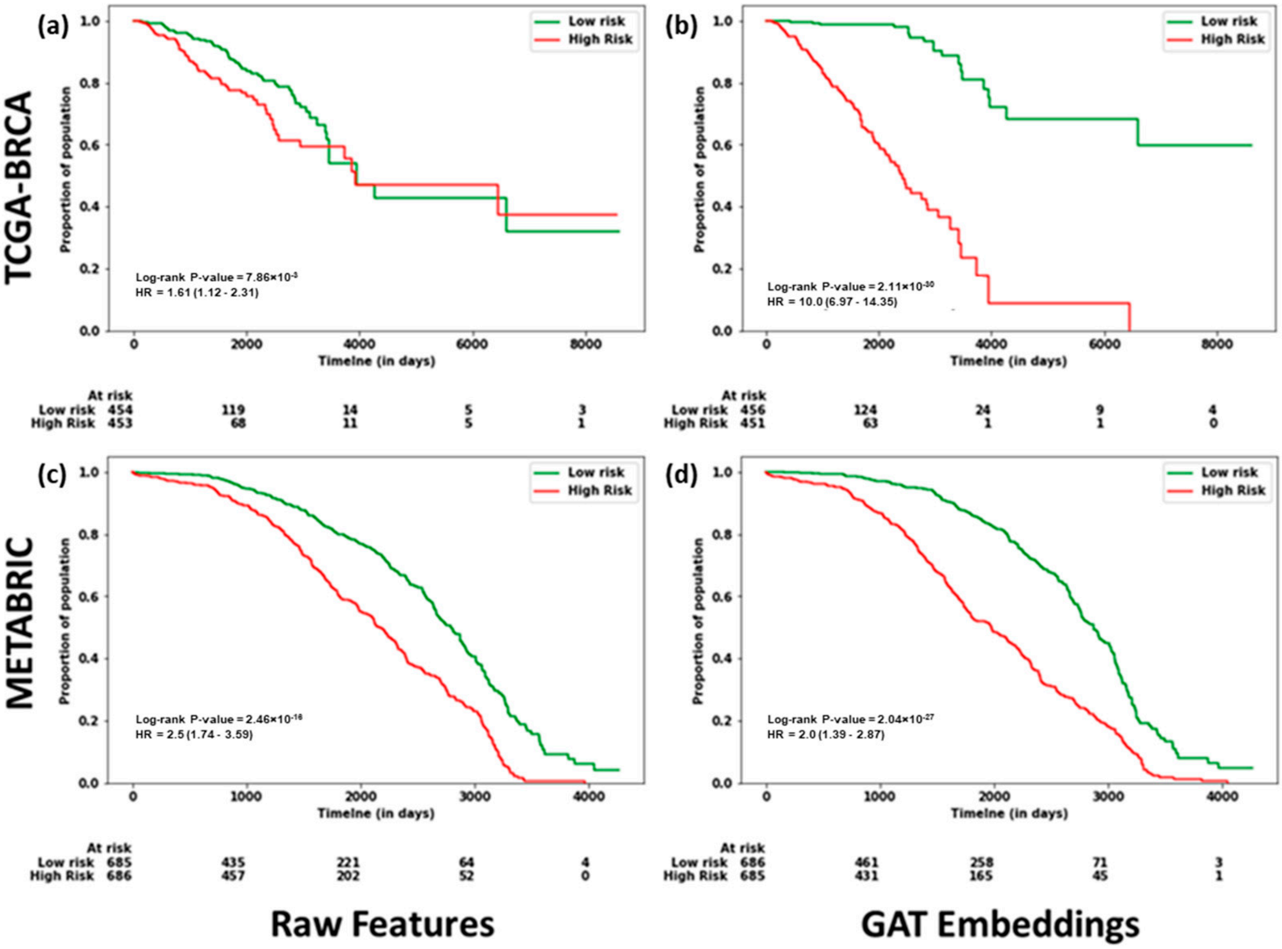

Figure 4. Survival analysis using raw features and GAT embeddings. (**a**,**b**): TCGA-BRCA. (**c**,**d**): METABRIC. Kaplan–Meier curves showing the proportion of population of low-risk and high-risk groups at different observation times. The low-risk and high-risk groups were determined by their risk scores calculated using raw features (**a**,**c**) and GAT embeddings (**b**,**d**).

3. Discussion

The hypothesis of the present study was that the attention-based graph neural network would provide better embeddings compared to the graph convolutional neural network (GCN). Our proposed MOGAT model for integrating multi-omics data based on graph attention network (GAT) provides better embeddings and performs better than the GCN-based approaches, such as MOGONET and SUPREME. MOGAT outperforms MOGONET by 32% to 46% and SUPREME by 2% to 16% in cancer subtype prediction in different combinations of multi-omics data, thus supporting our hypothesis.

The rationale for proposing GAT in multi-omics integration is that it has the built-in advantage of employing attention mechanisms to weigh neighbors' contributions, allowing each node to adaptively focus on its most informative neighbors during message passing. This can lead to better model generalization. In our study, nodes are patients. For example, if a node represents a patient with the basal subtype of breast cancer and has five neighbors, of which two are the basal subtype, it would be realistic to assign more attention to neighbors with the basal subtype than other subtypes. On the other hand, attention mechanisms introduce additional computational overhead. For large-scale graphs, this can make training and inference slower compared to simpler aggregation methods.

The GAT is a specific type of graph neural network (GNN) that utilizes attention mechanisms to dynamically weigh the importance of neighboring nodes during message passing. This means that GNN represents a broad family of neural network architectures designed for graph-structured data. This family includes various architectures and mechanisms, such as graph convolutional networks (GCNs), spectral-based GNNs, message-passing neural networks (MPNNs), and, of course, GATs, among others. Different GNN models have different mechanisms for aggregating information from neighbors. For instance, GCNs use a fixed weight averaging scheme, while MPNNs can employ more general message and update functions.

In the present study, we included eight types of data, including mutations, copy number alterations, mRNA expression, lncRNA expression, miRNA expression, co-expression eigengenes, DNA methylation, and clinical data. Note that we did not include protein expression for analysis. The reason is that it might reflect the same information as mRNA expression since mRNAs are translated into amino acids to form proteins. As a result, both mRNA expression and protein expression might generate similar patient similarity networks. There is already a concern that analysis by integrating seven omics might be an overkill. Adding protein expression could bemore overkill. Whether analysis using too many omics is an overkill deserves further investigation, which will be addressed in our future work.

The study presented in this work has some limitations that need to be addressed in future work. It was restricted to using only the TCGA-BRCA and METABRIC cohorts to predict its five different cancer subtypes. To check the efficacy of the proposed methodology, we will consider subtypes of other cancers in a pan-cancer analysis as well as other diseases, such as Alzheimer's and Parkinson's.

The current study also presents opportunities for future research. The framework used in this study utilized a patient similarity network as the input graph, with each node representing a patient. However, it is possible to reorganize the framework so that each node represents a gene, making the task of the graph neural network a graph classification instead of a node classification. While some existing methods follow this approach [5,8], it limits the number of omics data that can be incorporated as node features. For instance, when using genes as nodes, gene expression, somatic mutation, copy number variation, and DNA methylation can be incorporated, but not lncRNAs, miRNAs, and co-expression eigengenes due to the absence of a one-to-one association with genes. To address this, separate graph attention networks with different network and node features are required to incorporate these additional omics data.

Our framework is based on supervised machine learning, where the number of subtypes must be known beforehand. An unsupervised machine learning-based framework would allow for scenarios where the number of subtypes is not known. We envisage the integration of MOGAT with clustering or other unsupervised learning methods to tackle such scenarios.

4. Materials and Methods

4.1. Dataset Preparation and Preprocessing: TCGA

To develop and investigate the MOGAT approach, we downloaded omics and clinical data for breast invasive carcinoma (BRCA) from the GDC portal (https://portal.gdc.cancer.gov, accessed on 16 December 2022) of The Cancer Genome Atlas (TCGA). The RNAseq gene (mRNA, miRNA, and lncRNA) expression, DNA methylation, copy number variation, simple nucleotide variation, and clinical data were collected for this cohort. Table 3 summarizes the processed omics data with the number of features in different preprocessing steps. The preparation and preprocessing of different types of data are outlined in Figure 5.

Table 3. The summary of each datatype for the TCGA-BRCA cohort. Row 1 (Original Features): number of original features, Row 2 (Cleaned Features): number of features after cleaning by filtering, Row 3 (Selected Features): number of features after applying the Boruta feature selection approach, Row 4 (All Tumor Samples): Number of tumor samples, including duplicates, Row 5 (Unique Tumor Samples): number of tumor samples after removing duplicates, Row 6 (Common Samples): subtype distribution of the tumor samples common across all datatypes, and Row 7 (Network): number of nodes and edges for the patient similarity network for each datatype.

Datatype	CLI	CNA	COE	EXP	LNC	MET	MIR	MUT
Original Features	31	28,918	40	19,962	16,901	25,978	1881	16,662
Cleaned Features	31	28,918	40	5343	3398	25,978	306	16,662
Selected Features	31	500	40	1000	500	1000	306	200
All Tumor Samples	1089	1106	1212	1212	1212	1107	1069	992
Unique Tumor Samples	1089	1096	1076	1076	1076	1097	1057	969
Common Samples	920 Samples; Basal: 158 (17.24%) HER2: 73 (8.02%) LumA: 467 (50.65%) LumB: 188 (20.39%) NL: 34 (3.68%)							
Network (Nodes, Edges)	(920, 2398)	(920, 2346)	(920, 2122)	(920, 2391)	(920, 2218)	(920, 2675)	(920, 2108)	(920, 2752)

Note: CLI: clinical, CNA: copy number alteration, COE: co-expression, EXP: gene expression, LNC: lncRNA expression, MET: DNA methylation, MIR: miRNA expression, MUT: simple nucleotide mutation. LumA: Luminal A, LumB: Luminal B, NL: Normal-like. Blue: Number of features after cleaning; Red: Number of features after feature selection using Boruta.

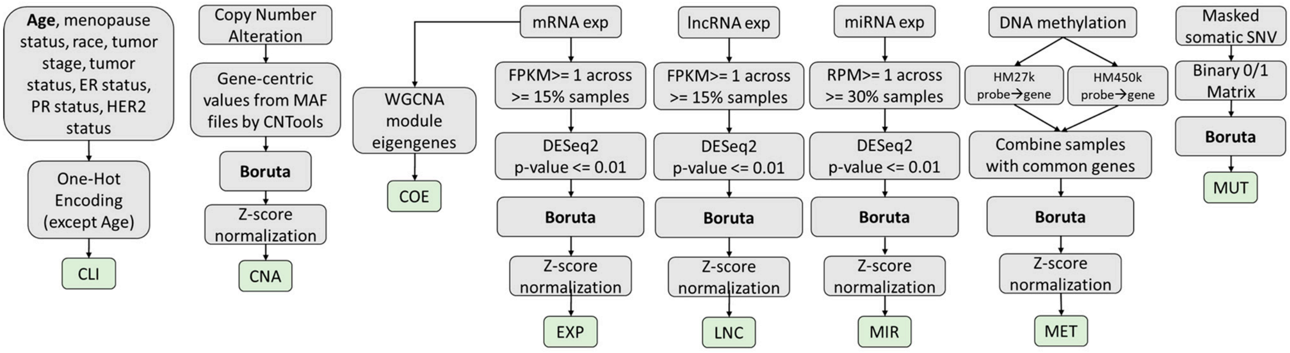

Figure 5. Flowcharts for data preprocessing for each datatype used in the MOGAT framework.

4.1.1. Features Based on Clinical Data (CLI Features)

The clinical data consists of age, race, neoplasmic cancer status, tumor stage, menopause status, estrogen receptor status, progesterone receptor status, and HER2 receptor status. The variables were one-hot encoded except for age. Finally, the number of features remained at 31, as shown in Table S1 in Supplementary Material S1.

4.1.2. Features Based on Copy Number Alterations (CNA Features)

Supplementary Material S2 provides the details of processing CNA features from copy number segment mean to a gene-centric matrix with an example. Copy number alteration (CNA) data came as TSV files (Figure S1 in Supplementary Material S2), each for one sample. Each row of the TSV file corresponds to a genomic coordinate for which the copy number alterations were observed, and the corresponding segment mean value is defined as: $log_2(\text{CopyNumber}/2)$. These files were combined into a single TSV file for all patients containing 1,268,167 rows. Then, CNTools [20] was used to obtain gene-centric values from the segmented copy number variation data, as explained in Figure S2, Tables S1 and S2 in Supplementary Material S2, and the number of genes was 28,918.

4.1.3. Features Based on Gene Co-Expression (COE Features)

For generating a gene co-expression network, WGCNA [21] was used. The optimal soft-threshold power β was selected by evaluating its effect on the scale-free topology model fit R^2. From the beta values from 1 to 30, 4 was the lowest while maintaining the high R^2 values (threshold 0.90) (Supplementary Material S3, Figure S1). The adjacency matrix was first transformed into a topological overlap matrix (TOM) -based similarity matrix for module detection. Then, it was converted into a TOM-based dissimilarity matrix by subtracting from unity (1). This matrix was used as a distance metric to perform the average linkage hierarchical clustering algorithm, which outputs a dendrogram. Then, the dynamic tree cut [22] method was performed for branch cutting to generate network modules. This process identified 40 modules. Supplementary Material S3, Table S1, shows the number of genes for each of the 40 modules identified by WGCNA. The values of eigengenes for each patient are provided in Supplementary Material S4.

4.1.4. Features Based on mRNA Expression (EXP Features)

RNAseq expression data contain the expression of 60,660 genes (including mRNAs, miRNAs, and lncRNAs), from which expression values of 19,962 mRNAs were isolated. The expression values were in FPKM (fragments per kilobase of transcript per million mapped reads). We employed three different approaches in sequence to reduce the original high-dimensional feature space to a meaningful low-dimensional space. First, some of the original features have very small values, such as FPKM ≤ 1 for many samples, which do not carry signals for analysis. The mRNAs are filtered out if their expression values do not meet the threshold of FPKM ≥ 1 in $\geq$15% of samples (as used in [11]), which resulted in 13,503 mRNAs. Second, these mRNAs were used to perform differential gene expression analysis using DESeq2 [23]. After using the criteria of an adjusted p-value ≤ 0.01, the number of remaining mRNAs was 5343, which we referred to as cleaned features, Table 3. Third, a well-known random forest-based feature selection package, BORUTA [24], was used to identify 1000 significant mRNAs. The first 3 rows of Table 3 summarize the feature selection results.

4.1.5. Features Based on lncRNA Expression (LNC Features)

Applying the similar preprocessing and feature selection approaches used for mRNA, we selected 500 significant lncRNAs and corresponding expression values from the original dataset of 60,660 gene expressions.

4.1.6. Features Based on miRNA Expression (MIR Features)

The miRNA expression data were in reads per million (RPM) units. The miRNAs were filtered out if their expression values did not meet the threshold of RPM $\geq$1 in $\geq$ 30% of samples, which resulted in 393 miRNAs. Then, differential gene expression analysis using an adjusted p-value ≤ 0.01 resulted in 306 miRNAs. The BORUTA feature selection was not used for MIR since the number of features is already low.

4.1.7. Features Based on DNA Methylation (MET Features)

HumanMethylation 27 k (HM27) and HumanMethylation 450 k (HM450) data were collected for DNA methylation. The samples and probes were 343 and 27,578 for HM27 and 895 and 485,577 for HM450, respectively. After combining both datasets by keeping the same probes, the samples and probes were 1238 and 25,978, respectively.

4.1.8. Features Based on DNA Mutation (MUT Features)

For simple nucleotide mutation data, there were 992 samples, and each sample contained a different set of genes for which one or more mutations were observed. The sample mutation data were converted into a vector of genes, where 1 signifies a mutation occurred and 0 signifies no mutation. The size of this vector is the union of all the genes from all samples, which is 16,662.

4.2. Original Features to Cleaned Features

Original features with gene expression values (i.e., EXP, LNC, and MIR features) have very small values, such as FPKM ≤ 1 or RPM ≤ 1 for many samples, which do not carry signals for analysis. Those features were filtered out. Then, differential gene expression analysis was conducted to determine the significant features, which we referred to as "Cleaned Features". The numbers in blue in the "Cleaned Features" row in Table 3 represent the cleaned EXP, LNC, and MIR features. The number of features from other data types remained the same as "Original Features" since they do not require any filtering.

4.3. Cleaned Features to Selected Features

Among the cleaned features in Table 3, copy number alteration (CNA), mRNA expression (EXP), lncRNA expression (LNC), DNA methylation (MET), and mutation (MUT) are high-dimensional compared to sample size. Focusing on a subset of significant features can reduce potential noise and overfitting often associated with high-dimensional data. Thus, the Boruta package [24], a feature selection method based on the random forest algorithm, was used for the feature selection process. Without this feature selection, node features would have very high dimensions (~29 K for CNA, ~5 K for EXP, ~3 K for LNC, ~26 K for MET, and ~17 K for MUT; in total, ~80 K features for TCGA-BRCA) as opposed to the sample size (920 patients). This step was not used for co-expression (COE), miRNA expression (MIR), and one-hot encoded clinical features (CLI), as they did not have many features like their other counterparts. The numbers in red in the "Selected Features" row of Table 3 are features selected by Boruta. The other three types of features remained the same as "Cleaned Features".

4.4. Dataset Preparation and Preprocessing: METABRIC

The METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) [25] cohort data were collected from cbiportal.com. The summary of the preparation and preprocessing of METABRIC data is given in Table 4. The repository contained clinical (CLI), copy number alterations (CNA), mRNA expression (EXP), DNA methylation (MET), and somatic mutation (MUT) data. For the clinical dataset, age at diagnosis, menopause status, estrogen receptor status, progesterone status, HER2 status, and ethnicity were considered for analysis. The variables were converted into a one-hot vector, except for age. Finally, the number of features remained at 14. The list of clinical features is given in Supplementary Material S1, Table S2. The other datasets were already preprocessed into a gene matrix, unlike the TCGA-BRCA data from the GDC portal.

Table 4. The summary of each datatype for the METABRIC cohort. Row 1 (Original Features): number of original features, Row 2 (Selected Features): number of features after applying the Boruta feature selection approach, Row 3 (All Samples): number of samples (no duplicates), Row 4 (Common Samples): subtype distribution of the tumor samples common across all datatypes, and Row 5 (Network): number of nodes and edges for the patient similarity network for each datatype.

Datatype	CLI	CNA	COE	EXP	MET	MUT
Original Features	14	22,544	49	24,368	13,188	173
Selected Features	14	1003	49	1048	1058	173
All Samples	2508	1905	1905	1418	2509	2174

Common Samples	1372 Samples; Basal: 218 (15.89%); HER2: 181 (13.19%);
	LumA: 500 (36.44%); LumB: 335 (24.42%);
	Normal-like: 138 (15.89%)					

Network (Nodes, Edges)	(1372, 5055)	(1372, 4853)	(1372, 4593)	(1372, 5023)	(1372, 5329)	(1372, 4765)

The co-expression eigengenes (COE) as the features from the gene co-expression network were generated using WGCNA, following a similar procedure for the TCGA-BRCA cohort. Supplementary Material S3, Figure S2 shows that 6 was the lowest value of soft-threshold power β while maintaining the high scale-free topology model fit R^2

value at 0.90. There were 49 co-expression modules, and 49 features (module eigengene, which is the 1st PCA component using expression of genes in a module) were found. Supplementary Material S3, Table S2 shows the number of genes in each of the 49 modules identified by WGCNA. The values of eigengenes for each patient are provided in Supplementary Material S5.

4.5. Duplicate Data Handling

It was observed that some patients have more than one sample in different omics datasets for the TCGA-BRCA cohort. For example, a patient with ID: TCGA-A7-A0DB contains three samples of mRNA expression, miRNA expression, and mutation data. Therefore, to make sure each patient has one corresponding sample, an extra preprocessing step was taken. For mRNA expression, miRNA expression, co-expression, DNA methylation, and copy number alterations data, the multiple samples corresponding to single patients were replaced with one sample by taking the average value for each feature. Since the mutation data are in binary (0/1), the Boolean OR operation was performed instead of the average for the patients with more than one sample.

4.6. Missing Data Handling

In multi-omics integration analysis, data could be missing both across and within omics [26]. In multi-omics study designs, it is common for individuals to be represented for some omics layers but not all, which results in across-omics missing data. The same is true for both datasets used for breast cancer, one from TCGA and the other from METABRIC. For TCGA breast cancer data, the number of samples for eight omics varies from 969 for mutation to 1097 for DNA methylation (row “Unique Tumor Samples” of Table 3), and the six omics for METABRIC vary between 1418 for gene expression and 2509 for DNA methylation (row “All Samples” in Table 4). This means that both datasets have across-omics missing data. We used the common samples across omics to avoid bias due to the across-omics missing data, which are 920 and 1372 for the TCGA and METABRIC datasets, respectively.

The major issue with within-omics missing data is that true zeros (representing the true gene expression levels, for example) are mingled with dropout zeros (representing the actual missing data) [27], which is altogether a different topic and beyond the scope of this study. The multi-omics integration pipelines—MOGONET and SUPREME—we are comparing did not consider the handling of within-omics missing data.

4.7. Normalizing Feature Values

The selected feature values in categories CNA, EXP, LNC, MIR, and MET are z-score normalized. The CLI (one-hot encoded), COE features (eigengenes), and MUT features (binary) do not require normalization.

4.8. Final Set of Patients for Analysis

Our objective is to integrate the most available number of omics data and to investigate the effects of each data type on a patient’s outcome. However, different omics datasets contain different numbers of samples. The union of samples would lead to some patients not having features from all types of omics data; thereby, incorporating them would not meet the study objective. Thus, we used the intersection of samples from all types of data, which resulted in 920 and 1372 tumor samples common in all types for the TCGA-BRCA and METABRIC cohorts, respectively. Therefore, these common tumor samples were used to create the patient similarity networks (last row in Tables 3 and 4) and the feature matrices.

4.9. Network and Feature Matrix Construction

The input to the graph attention network is a network and the feature matrix for the nodes in the given network. In this case, the training would be in single mode, where each sample corresponds to a node in the network, and the whole network represents a data type (gene expression, mutation, etc.).

4.9.1. Network Construction

In the present study, a network is a patient similarity network, where a node represents a patient and an edge represents the similarity between two patients. The similarity matrix was created for each data type before constructing the corresponding patient similarity network. The similarity was computed using Pearson's correlation for mRNA expression, miRNA expression, lncRNA expression, co-expression features, copy number alterations, and DNA methylation. The Jaccard similarity was used instead of Pearson's correlation network for mutation data, as it is binary. The Gower metric [28] was used to compute patient similarity using clinical data since it combines categorical and continuous features. Then, from each similarity matrix, the top 3 similar samples for each sample were selected as edges to construct the similarity network.

4.9.2. Feature Matrix Construction

Each row of the feature matrix represents the feature vector for a node (here, a patient), which is the concatenation of all types of features for that patient coming from eight types (for the TCGA cohort) or six types (for the METABRIC cohort) of data. Note that only the selected features from each data type are concatenated, which resulted in a feature matrix of 920 × 3577 and 1372 × 3345 for the TCGA and METABRIC cohorts, respectively.

4.10. Interaction between Omics Data

The interactions between different omics types are considered during node feature engineering. The patient similarity network was constructed based on omics-specific data, but the node features contain features from all the omics data for each node or patient. While individual graph attention networks operate independently, they do so on an integrated foundation based on node features from all omics data. The post-analysis concatenation represents an integration of these learned representations, not a simplistic merger of isolated data types.

4.11. Graph Attention Network

The utilized GAT model is based on the idea of the self-attention mechanism, where embeddings are created from eight different types of data (Table 3) with the assumption that samples with similar characteristics (such as gene expression or DNA methylation) are likely to have similar disease outcomes and are, therefore, related to each other. However, not all related samples should be given equal importance. Some samples may have a greater impact on the prediction or clustering of a target sample, which cannot be accurately determined by similarity metrics. To address this, the GAT model assigns varying levels of attention to a target sample's neighboring samples, allowing it to capture the significance of each one.

For each data type, let n be the number of samples or patients and m be the number of features (concatenated from different omics types). The input feature matrix is given by $X = [x_1, x_2, \ldots, x_n]$, where $x \in \mathbb{R}^{1\times m}$ represents a sample feature vector. While generating the embedding of sample x_i, the attention given to it from its neighbor x_j can be calculated as:

$$c_{ij} = LeakyReLU\left(a^T \left[Wx_i \,||\, Wx_j\right]\right) \tag{1}$$

where $W \in \mathbb{R}^{p\times m}$ and $a \in \mathbb{R}^{2p\times 1}$ are learnable weight parameters, shared across all samples and p is the embedding size; | | symbol denotes the concatenation of two vectors; and *LeakyReLU* is the non-linear activation function. c_{ij} describes the importance of sample j's feature to sample i. We then normalize attention coefficients by applying a SoftMax function:

$$\alpha_{ij} = Softmax(c_{ij}) = \frac{exp(c_{ij})}{\sum_{k\in N_i} exp(c_{ik})} \tag{2}$$

where N_i is the set of neighboring nodes of sample i. With the normalized attention coefficients being the weights, a linear combination of input features is used as the output representation for each data sample. Formally, we have:

$$h_i = \sum_{j \in N_i} \alpha_{ij} W x_i \tag{3}$$

where h_i is the output representation of sample i.

4.12. Training GAT

For training the GATs, the architecture remained the same for all the datatypes, consisting of two graph attention layers. The hidden layer dimension for each GAT model and learning rate were selected based on grid search-based hyperparameter tuning. The ranges of values for hyperparameters are listed in Table 5. For LeakyReLU, the hyperparameter called negative input slope, α, was used as 0.2 following [16].

Table 5. Hyperparameter Tuning for GAT, MLP, and LASSO. The type of hyperparameter and their ranges of values used for tuning are provided. Optimized hyperparameter values for GAT and LASSO are bolded for TCGA-BRCA. For MLP tuning, it has 255 sets of optimized hyperparameter values, one for each combination of multi-omics data.

Hyperparameters	Values
GAT hidden layer dimensions	[128, 256, **512**, 1024]
GAT Learning Rate	[0.01, **0.001**, 0.0001]
GAT # of epochs	[100, **200**, 500]
GAT # of heads	[**1**, 2, 4, 8]
MLP learning rate	[0.1, 0.01, **0.001**, 0.0001, 0.00001]
MLP hidden layer architecture	[(32), (64), (128), (256), (512), (32, 32), **(64, 32)**, (128, 32), (256, 32)]
MLP # of epochs	[200, 500, **1000**, 1500]
LASSO regularizing factor α	[0.001, 0.002, 0.005, 0.01, 0.05, **1.0**]

4.13. Classification

Embeddings generated after training the GATs were concatenated and used as input for classification. A multi-layer perceptron (MLP) was used to classify breast cancer subtypes. The architecture, learning rate, and number of epochs for MLP were selected based on a randomized grid search. The range of values is listed in Table 5.

The classification metrics, including accuracy, weighted-F1 score, and macro-F1 score, were estimated to evaluate the performance of the MOGAT model.

4.14. Implementation

All experiments were conducted on a Linux machine with 8 NVIDIA A100 GPUs, each with 40 GB of memory. The software environment was CUDA 11.6 and Driver Version 520.61.05. We used Python 3.9.13 and Pytorch 1.12.1 to construct our project. Other packages and their versions are available in the GitHub repository.

4.15. Survival Analysis

Survival analysis was performed using raw features (concatenated selected features after cleaning and feature selection) and GAT embeddings separately. Table 6 shows the number of raw features and GAT embeddings at different stages of survival analyses. The initial numbers of raw features were 3577 and 4335 for TCGA-BRCA and METABRIC, respectively. The initial numbers of embeddings were 4096 (8 × 512) and 3072 (6 × 512). For the TCGA-BRCA cohort, LASSO regression with overall survival as output reduced the number of raw features and embeddings to 276 and 2247, respectively. The regularizing factor α for LASSO was selected using a Grid SearchCV approach. The list of values for α used in Grid Search and the optimized value are given in Table 5.

Table 6. The number of variables (Raw features and GAT embeddings) that remained at each step of the survival analysis process. Initial number of raw features is the sum of the reduced set of features from eight or six different data types (Row: "Selected Features" from Tables 3 and 4).

	TCGA-BRCA		**METABRIC**	
Item	**Raw**	**Embeddings**	**Raw**	**Embeddings**
Initial	3577	4096	4335	3072
After LASSO	276	2247	21	21
After Cox-PH	57	542	8	6

Next, a multivariate Cox proportional hazard (Cox-PH) regression analysis [29] was conducted using the selected features in the previous step. This technique examines the influence of multiple variables on the time it takes for a specific event to occur, in this case, death. In the Cox regression model, the coefficients of predictor variables (raw features or embeddings) are related to hazard, i.e., risk of death. A positive coefficient indicates a worse prognosis, and a negative coefficient indicates a protective effect of the variable with which it is associated. The exponent of its coefficient gives the hazard ratio associated with a predictor variable, and the p-value shows the significance of the association between the predictor variable (raw feature or embedding) and the risk of death. The significant predictor variables, 57 raw features and 542 embeddings, with a p-value < 0.05, were selected, and their Cox coefficients were used to calculate the risk scores using raw features and embeddings, respectively.

$$RiskScore = \sum X_i * Coef_i \tag{4}$$

where X_i is the value of i-th predictor (raw feature or embedding) and $Coef_i$ is the corresponding coefficient for the predictor obtained from the Cox regression. This Risk Score is used to divide the cohort into low-risk and high-risk groups using the median as the divider. Then, Kaplan–Meier [30] and logrank tests [31] were performed, and hazard ratios were calculated to see if the two groups were significantly distinguishable.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/ijms25052788/s1.

Author Contributions: Conceptualization, R.B.T., D.L. and A.M.M.; methodology, R.B.T., M.S., M.M.I. and D.L.; software, R.B.T. and M.S.; validation, R.B.T.; investigation, R.B.T. and A.M.M.; resources, R.B.T., D.L. and A.M.M.; data curation, R.B.T.; writing—original draft preparation, R.B.T.; writing—review and editing, D.L. and A.M.M.; visualization, R.B.T. and M.M.I.; supervision, A.M.M.; project administration, A.M.M.; funding acquisition, A.M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the NSF CAREER Award (#1901628) to AMM.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code and data required for training the model are provided in the GitHub link: https://github.com/MezbahJUCSE39/MOGAT.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, T.; Shao, W.; Huang, Z.; Tang, H.; Zhang, J.; Ding, Z.; Huang, K. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Commun.* **2021**, *12*, 3445. [CrossRef]
2. Chan, Y.H.; Wang, C.; Soh, W.K.; Rajapakse, J.C. Combining Neuroimaging and Omics Datasets for Disease Classification Using Graph Neural Networks. *Front. Neurosci.* **2022**, *16*, 866666. [CrossRef] [PubMed]
3. Wang, W.; Zhu, G.; Wang, Y.; Li, W.; Yi, S.; Wang, K.; Fan, L.; Tang, J.; Chen, R. Multi-Omics Integration in Mice with Parkinson's Disease and the Intervention Effect of Cyanidin-3-O-Glucoside. *Front. Aging Neurosci.* **2022**, *14*, 877078. [CrossRef] [PubMed]

4. Subramanian, I.; Verma, S.; Kumar, S.; Jere, A.; Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform. Biol. Insights* **2020**, *14*, 1177932219899051. [CrossRef] [PubMed]
5. Li, B.; Wang, T.; Nabavi, S. Cancer Molecular Subtype Classification by Graph Convolutional Networks on Multi-Omics Data. In Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, Gainesville, FL, USA, 1–4 August 2021. [CrossRef]
6. Kipf, T.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907.
7. Zhou, N.; Wang, S.; Tan, Z. AEMVC: Anchor Enhanced Multi-Omics Cancer Subtype Identification. In Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences, Amsterdam, The Netherlands, 13–15 October 2022; pp. 57–63. [CrossRef]
8. Guo, H.; Lv, X.; Li, Y.; Li, M. Attention-based GCN Integrates Multi-omics Data for Breast Cancer Subtype Classification and Patient-specific Gene Marker Identification. *bioRxiv* **2022**. [CrossRef] [PubMed]
9. Li, X.; Ma, J.; Leng, L.; Han, M.; Li, M.; He, F.; Zhu, Y. MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis. *Front. Genet.* **2022**, *13*, 806842. [CrossRef] [PubMed]
10. Yin, C.; Cao, Y.; Sun, P.; Zhang, H.; Li, Z.; Xu, Y.; Sun, H. Molecular Subtyping of Cancer Based on Robust Graph Neural Network and Multi-Omics Data Integration. *Front. Genet.* **2022**, *13*, 884028. [CrossRef] [PubMed]
11. Kesimoglu, Z.N.; Bozdag, S. SUPREME: Multiomics data integration using graph convolutional networks. *NAR Genom. Bioinform.* **2023**, *5*, lqad063. [CrossRef]
12. Al Mamun, A.; Mondal, A.M. Feature Selection and Classification Reveal Key lncRNAs for Multiple Cancers. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019; pp. 2825–2831. [CrossRef]
13. Kaikkonen, M.U.; Lam, M.T.Y.; Glass, C.K. Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovasc. Res.* **2011**, *90*, 430–440. [CrossRef]
14. Al Mamun, A.; Duan, W.; Mondal, A.M. Pan-cancer Feature Selection and Classification Reveals Important Long Non-coding RNAs. In Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Republic of Korea, 16–19 December 2020; pp. 2417–2424. [CrossRef]
15. Al Mamun, A.; Tanvir, R.B.; Sobhan, M.; Mathee, K.; Narasimhan, G.; Holt, G.E.; Mondal, A.M. Multi-Run Concrete Autoencoder to Identify Prognostic lncRNAs for 12 Cancers. *Int. J. Mol. Sci.* **2021**, *22*, 11919. [CrossRef]
16. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio', P.; Bengio, Y. Graph Attention Networks. *arXiv* **2017**, arXiv:1710.10903.
17. Zhao, W.; Gu, X.; Chen, S.; Wu, J.; Zhou, Z. MODIG: Integrating multi-omics and multi-dimensional gene network for cancer driver gene identification based on graph attention network model. *Bioinformatics* **2022**, *38*, 4901–4907. [CrossRef]
18. Jolliffe, I.T. *Principal Component Analysis*; Springer Science & Business Media: New York, NY, USA, 2002.
19. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
20. Zhang, J. CNTools: Convert Segment Data into a Region by Sample Matrix to Allow for Other High Level Computational Analyses. 2023. Available online: https://git.bioconductor.org/packages/CNTools (accessed on 20 May 2023).
21. Langfelder, P.; Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **2008**, *9*, 559. [CrossRef]
22. Langfelder, P.; Zhang, B.; Horvath, S. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* **2007**, *24*, 719–720. [CrossRef]
23. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
24. Kursa, M.B.; Rudnicki, W.R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
25. Curtis, C.; Shah, S.P.; Chin, S.-F.; Turashvili, G.; Rueda, O.M.; Dunning, M.J.; Speed, D.; Lynch, A.G.; Samarajiwa, S.; Yuan, Y.; et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **2012**, *486*, 346–352. [CrossRef]
26. Song, M.; Greenbaum, J.; Luttrell, J.; Zhou, W.; Wu, C.; Shen, H.; Gong, P.; Zhang, C.; Deng, H.-W. A Review of Integrative Imputation for Multi-Omics Datasets. *Front. Genet.* **2020**, *11*, 570255. [CrossRef]
27. Gong, W.; Kwak, I.Y.; Pota, P.; Koyano-Nakagawa, N.; Garry, D.J. DrImpute: Imputing dropout events in single cell RNA sequencing data. *BMC Bioinform.* **2018**, *19*, 220. [CrossRef] [PubMed]
28. Gower, J.C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **1971**, *27*, 857. [CrossRef]
29. Mauger, E.A.; Wolfe, R.A.; Port, F.K. Transient effects in the cox proportional hazards regression model. *Stat. Med.* **1995**, *14*, 1553–1565. [CrossRef] [PubMed]
30. Kaplan, E.L.; Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **1958**, *53*, 457. [CrossRef]
31. Bland, J.M.; Altman, D.G. The logrank test. *BMJ* **2004**, *328*, 1073. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.