ThinMLA:Late-Fusion for Graph Attention Layer Feature Based on Co-expression Gene Modules for Disease Diagnosis

Ziyun Zou[1], **Shuxian Cai**[2], **Xinxin Xiong**[3], **Shiyu Xu**[3], **Zelong Chen**[3]
School of Informatics Xiamen University[1]
School of Medicine Xiamen University[2]
4221 Xiang'an South Road, Xiamen, China
{23020221154156,23020221154069, 23020221154129, 24520221154586, 23020221154076}@stu.xmu.edu.cn

Abstract

In the progression of the disease, genes usually work interactively and change synergistically, which leads to aberrations in clinical omics and difficulties in medical judgment. Therefore, there is no doubt that grasping the knowledge of co-expression gene situations could help in disease state prediction. Despite traditional methods, rapid-developing deep learning technologies have also prompted this area in recent years. In this work, we apply a late-fused graph attention layers-based model to conduct disease diagnosis using co-expression gene data. Our model first uses several graph attention layers to extract high-level features and build corresponding high-dimensional layers, then subtract high-level features with original input, and finally applies a Squeeze-and-Excitation attention module-based late-fusion paradigm to integrate them. In addition, we conduct some experiments and demonstrate the redundancy within different dimensions to mitigate the effects.

Introduction

Genomics omics refers to a comprehensive and global assessment of entire genomes instead of genetics that only stands for individual variants or single genes(Hasin Y. and A. 2017). The purpose of genomics is to collectively characterize and quantify all the genes of an organism, and to study their interrelationships and their impact on the organism. With disease progression, genomics omics go through a series of changes and aberrations, which makes it possible to serve as biomarkers that could help in observing and indicating disease(Veličković et al. 2017). Therefore, new genomic tools for omics analysis will improve our ability to identify diseases in the presymptomatic phase and guide the therapeutic procedures(J 2004).

Research and analysis on genomics omics have been widely carried out together with the development of computer science technologies, from machine learning to deep learning nowadays. Machine learning methods can be used to "learn" how to identify the location of transcription start sites (TSS) in genomic sequences(M. and W. 2015). (Iorio et al. 2016) used elastic net models to predict the drug IC50 of cancer cell lines given their profiles of gene mutations and expression levels, depending on the compound, a range of predictive accuracy is observed. With the same dataset, (Cortés-Ciriano et al. 2015) showed that predictive performance could in some cases be improved using a random forest model linked to a measure of statistical confidence in each prediction. (Kuzmin et al. 2020) used a variety of machine learning methods (support vector machine, logistic regression, decision tree, random forest) to predict the host specificity of coronaviruses, and their scheme has a high accuracy rate by analyzing the host specificity of multiple spike sequences. (Gupta et al. 2022) developed a machine learning tool that used support vector machines to build a model that predicts therapeutic proteins and classifies therapeutic categories. Using this tool can accurately predict disease-causing proteins in genome and metagenomic datasets, providing evidence in real validation, but the improved tool does not perform well on real data.

With the gradual rise of more powerful technology with deep learning, researches on disease diagnosis has been prompted. (Aramburu A and et al 2015)used a semi-supervised deep learning approach to analyze the prognosis of lung cancer, (RJ et al. 2022) utilized multimodal technology for survival outcome prediction. When wider studies have been carried out, these methods are said to ignore the topological information among genes. Concurrently, more recent researches find that co-functional gene modules could show the disease status and biological processes better and clearer(Muzio, O'Bray, and Borgwardt 2020). (M. et al. 2022) developed a priori attribution based on Fourier transform, trained a more stable and interpretable deep neural network model, and achieved consistent and reliable revelation of biological patterns that drive various genomic regulatory events, while refining the decision-making process of neural networks. Due to the specialty, the co-expression data holds, one of the widely adopted methods is graph neural network(GNN). Xing et al. proposed to use a multi-level attention graph neural network (MLA-GNN) to explore the gene modules and topological information in omics data, but it did not fully use the hierarchical information among layers, which might waste a lot of extracted features and information.

Thus, in this work, we decide to integrate a Squeeze-and-Extraction(SE) (Jie, Shen, and Sun 2018) attention module with the feature extraction network of MLA-GNN, which could not only gain the hidden relationship between genes, but also take better usage of the hierarchical relationships between gotten graph neural layers. What's more, inspired by the shortcut strategies of ResNet, we concern that the usage of GAT might lead to the redundancy of information and carried out trial and error, finally proposing our model, which gains better performance results based on evaluation metrics. We call the proposed model ThinMLA since it eases the redundancy of the information.

Related Work

High-throughput technologies have revolutionized medical research continuously(Hasin Y. and A. 2017), and both traditional statistics and machine learning approaches as well as a deep neural network are adopted. (Sidharth S Prakash 2020) built a usual deep neural network that can predict the malignancy of breast cancer. (Yue Zhao and Andrew Pattison 2020) developed an RNA-based classifier that uses a 1D inception convolutional neural network model to infer the primary tissue of origin for tumors. (Iorio et al. 2016) applied an elastic net model to predict the drug IC50 of cancer cell lines given their profiles of gene mutations and expression levels. (Capper D and et al 2018) used the random forest algorithm to train the whole genome information of 2810 cancer patients to realize the classification of central nervous system tumors. (Aramburu A and et al 2015) exploited a semi-supervised deep learning approach to analyze the prognosis of early-stage non-small cell lung cancer based on gene copy number variation combined with clinical information and gene expression, which is demonstrated to be quite robust. (Kong and Yu 2018) constructed a graph based on HINT dataset(H. 2012), and then conduct disease outcomes prediction using a feedforward GNN. Nevertheless, these methods only consider gene seperately whereas ignore the topological information among them.A stochastic block model neural network was introduced by (Fanfani et al. 2021) and then conduct disease outcomes prediction using a feedforward GNN. Nevertheless, these methods only consider genes separately whereas ignoring the topological information among them.

To include the topological relationships, instead of identifying markers as individual genes, (Chuang HY, Liu YT, and T 2007) made breast cancer predictions through subnetworks extracted from protein interaction databases, but it lacks higher-level feature extraction. In contrast, (ie Hao, Song, and Kang 2020) propose a convolutional neural network combined with a gene pathway, based on patch texture for cross-modal analysis, which can extract global survival discriminant features without manual annotation of pathologically specific layers. (Strand and et al 2020) presented a prognostic classifier that predicts both precancer recurrence and invasive progression. They compared and correlated breast cancer data with routine pathological findings, clinical outcomes, and disease status to generate spatial resolution maps of pre-breast cancer lesions based on a multiscale approach. Similarly, to address the problem that general graphs may not reflect gene interactions in specific diseases, (Veličković et al. 2017) proposed an interpretable multi-level attention graph neural network (MLA-GNN) to explore the gene modules and topological information in omics data. Specifically, it uses graph attention layers(GAT) to extract higher features and figure out the relationships between nodes. Thus, each node in the higher-level layer will contain information about its neighbors together with the topological information among them. Then, they simply use linear projection to normalize the dimensions, and finally, simply concatenate the features and transfer them into several fully convolutional layers to do the prediction. However, it did not fully take advantage of the relationship between layers, which may lead to the extraction effect wastage.

Squeeze-and-Excitation Networks

(J. Hu and Sun 2018) focused on the channel relationship and proposed a new architecture unit called "squeeze and excitation" (SE) block, that adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels. Experiments show that these blocks can be stacked together to form SENet architectures that generalize extremely effectively across different datasets. In addition, SE blocks bring significant improvements in performance for existing state-of-the-art CNNs at a slight additional computational cost.

Deep Residual Learning for Image Recognition

(K. He and Sun 2016) proposed ResNet (Residual Networks) in 2015 to solve the problem that the training effect of neural networks becomes worse when the number of layers is too deep. By directly bypassing the input information to the output, the integrity of information is protected. The entire network only needs to learn the difference between the input and output, simplifying the learning objectives and difficulties.

Proposed Method

The overview of the proposed model is illustrated in Figure 1. Unlike most of the previous works, we decide to construct a graph based on each patient rather than considering all patients as a whole. Given K genes data for N patients, we will first use a classic R package, WGCNA(Langfelder P. 2008), a weighted correlation network analysis for clustering highly correlated genes, is used to build the edge matrix $E^{K\times K}$ for each patient. Thus, a co-expression graph $G_1 = G(V^{K\times 1}, E^{K\times K})$ is obtained, where $V^{K\times 1}$ stands for the features of K gene nodes and the edge matrix $E^{K\times K}$ indicates the correlation relationships of genes. Next, we apply two graph attention layers (GAT) to construct hierarchical features G_2 and G_3 respectively. In order to ease the redundancy, we subtract each head of the G_2 and G_3 layers with G_1, which makes the G_2 and G_3 retain the high-level information better. In the proposed multi-level graph late-fusion module, a SE attention toolkit is added to catch the importance of different levels and channels and weighted them. And then, graph features from different levels are concatenated after linear projection(LP) and vectorization. Finally, the fused feature is sent to the last stage of the pipeline which is used to conduct downstream tasks, in this case, disease classification and survival prediction.

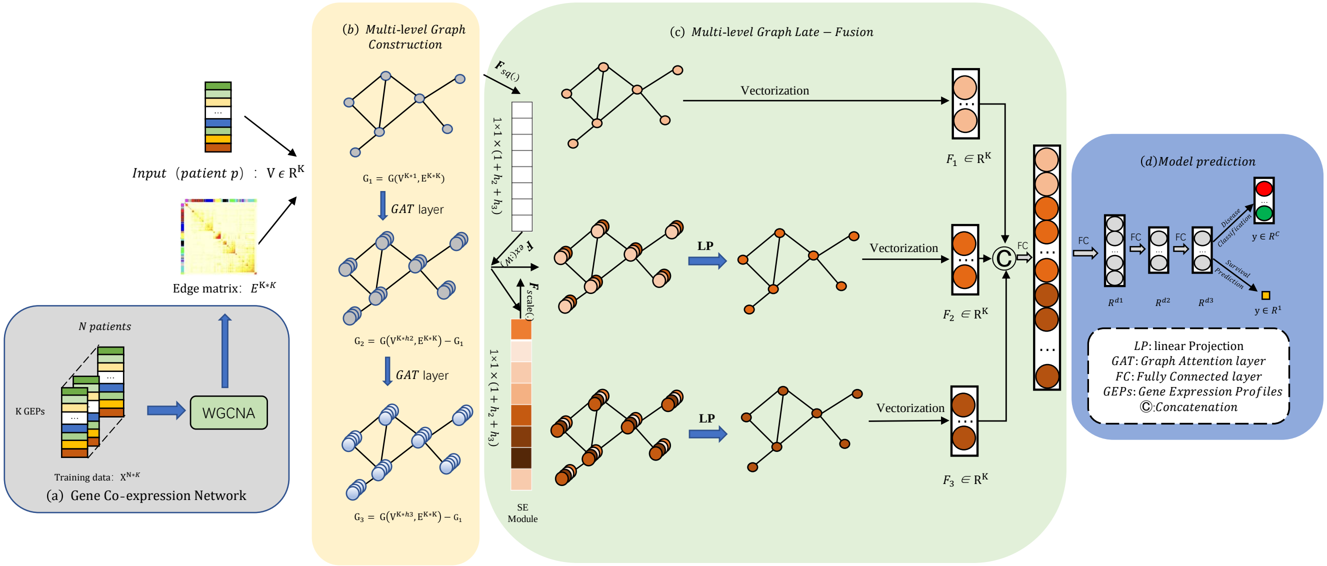

Figure 1: Proposed model

Gene co-expression computation

To start with, we use GNNs to format the omic data of each patient into a graph represented by feature V and an edge matrix E. Expect that each patient has K genes, the feature can be implemented by $V^{K\times1}$, which indicates to a K nodes graph with each node contains the expression of a gene. As mentioned, we use the WGCNA analysis package to conduct the gene co-expression analysis and get the edge matrix E. Specifically, the expression profile of each gene node in training data $X^{K\times K}$ is represented by an N-dimensional vector, where N is the number of patients in the training set. The correlation calculation A_{ij} between two nodes v_i and v_j is

$$A_{ij}=\left(\frac{1}{2}\left(1+\frac{\sum_{n=1}^{N}\left(v_{i,n}-\bar{v}_i\right)\left(v_{j,n}-\bar{v}_j\right)}{\sqrt{\sum_{n=1}^{N}\left(v_{i,n}-\bar{v}_i\right)^2}\sqrt{\sum_{n=1}^{N}\left(v_{j,n}-\bar{v}_j\right)^2}}\right)\right) \tag{1}$$

where $\bar{v}_i$ and $\bar{v}_j$ means average features of the node v_i and v_j. The WGCNA package's 'pickSoftThreshold' function transforms the correlation matrix into an adjacency matrix using a power function of the soft threshold β. By this means, we obtain an adjacency matrix $A \in R^{K\times K}$, which shows that genes with similar expressions have higher adjacency values.

To create the edge matrix E, we binarize the continuous values in the adjacency matrix A through

$$E_{ij}=\begin{cases}1, & A_{ij} > adj_{thresh}\\ 0, & otherwise\end{cases} \tag{2}$$

where the hyperparameter adj_{thresh} is tuned using an automated machine learning technique [20]. Notice that, the edge matrix E has universality and does not necessitate repeated calculations because it is computed using all training data rather than patient-specific data.

Using the edge matrix E, each patient is considered as a gene co-expression graph $G_1 = G(V^{K\times1}, E^{K\times K})$. In this graph, edges connect nodes with comparable gene expression while leaving other nodes unconnected. By processing co-functional gene modules connected in the coexpression graph, enhanced features of biological gene modules can be recovered, and illness prediction performance can be improved.

β For multi-level graph construction, the gene co-expression graph $G_1 = G(V^{K\times1}, E^{K\times K})$ of each patient is loaded into a stack of GAT layers. The GAT layer is a more advanced graph convolution layer that outputs each node's features as a weighted combination of adjacent nodes and nodes themselves. Co-functional genes, as previously stated, are more likely to be linked in the gene coexpression graph. As a result, the output feature of each node following the GAT layer is a weighted mixture of gene characteristics on the co-functional gene module, which can more accurately depict illness status (G and et al 2021).

Multi-level graph feature late-fusion

There are three different levels of graphs in total, including the input graph G_1, and the high-level graphs generated through GAT layers G_2 and G_3. Although they have the same number of nodes, they hold different hierarchical information that the node in G_1 stands for the expression of an individual gene whereas nodes in G_2 and G_3 contain knowledge from many co-functional genes and their neighbors. For the sake of the significance, both gene features and gene-group module features hold in omics representation learning(Ben-Hamo and et al 2020), we fuse the multi-level graph features to produce more discriminative feature representations.

However, since different layers have different dimensions both in information and nodes, simply concatenating them might lead the network to put more attention on the graphs at a higher level, we further deal with the layers before fusion. Firstly, since G_2 and G_3 both include the basic information of G_1, we consider that it is a huge unnecessary redundancy. Thus, inspired by the shortcut concept of ResNet (He Kaiming 2016), each block obtained from the heads of GAT is subtracted by G_1. Secondly, to better balance the importance of G_1, G_2, and G_3 to make the network more effective, a SE module (Hu Jie and Sun 2018) is added to attach different weights on different channels. Lastly, to concatenate layers, we reduce the node dimension of G_2 and G_3 through linear projection by fully connected layers to get G_2' and G_3'. Then, features G_1 , G_2' and G_3' are vectorized and concatenated at the end to produce the fused feature $F \in R^{3k}$:

$$F = \begin{matrix} G_1 \\ [G_2 \cdot T_2] \\ G_3 \cdot T_3 \end{matrix} \tag{3}$$

where $T_2 \in R^{b_2 \times 1}$ and $T_3 \in R^{b_3 \times 1}$ denote the weight parameters in he linear projection layers. Thus, this late-fusion module could utilize information from different dimensions with higher efficiency. The multi-level graph feature, which accurately mimics the biological regulatory process and can thus better reveal the disease mechanism, is of utmost importance to bioinformatics research and clinical applications. Disease progression is a complex biological process in which genes interact or cooperate.

Unlike other research that extracts representations by pooling across all the nodes, it is also noticeable that this method compresses the features inside each node while keeping the node structure in the graph, we believe it remains the biological meaning of each individual gene better rather than destroy it by compressing across nodes.

Model prediction

The proposed method is intended to address a wide range of therapeutic tasks. In the prediction module, the fusion feature F is stored by a continuous fully connected layer for disease classification and survival prediction, as shown in Figure 1c.

For the illness categorization task, the output $y \in R^c$ reflects the probability score for c classes. The problem is optimized via cross-entropy loss. The output $y \in R^1$ indicates the ”risk ratio” in the survival prediction task. And the cox loss is computed as:

$$L_{cox} = -\sum_{C(p)=1} \left(y_p - \log_{y_q \geq y_p} \exp\left(y_q\right) \right) \tag{4}$$

where $C(p) = 1$ denotes the non-censored patient set, and the cox loss calculation includes just the non-censored patients.

Experiment

To find the best way to ease the effect of redundancy between different layers and show the influence of the SE module, we conduct several groups of contrast and ablation experiments. To better compare with (Xing et al. 2022), we use the RNAseq of glioma cases from TCGA-GBM and TCGA-LGG projects as the basic datasets. The following section will discuss the implementations in detail.

Dataset

The dataset contains the archives and 20531 expressed genes for 769 patients. Because the training samples are very limited and the specialty of datasets in bioinformatics fields holds, including too many features may lead to the 'curse of dimensionality' or 'large p, small n" problem, we need to select features included carefully. We use semi-supervised top 240 genes as the input following the instruction of (Xing et al. 2022). The clinical information includes survival outcomes and histological grading(Grade II, Grade III, and Grade IV), as the labels to be predicted by the models.

Implementation details

We implement the proposed model with Pytorch and Pytorch Geometric library. The model is trained for 50 epochs with Adam optimizer and with batch size set to 8. The learning rate is initialized as 0.002 and linearly decayed in the training process. Adj_{thresh}, defines the density of the gene co-expression graph, which is set as 0.08. As for the dataset set-up, to validate the effectiveness of our proposed model, we conduct Monte Carlo 15-fold cross-validation in the experiments and the split is strictly consistent with (Xing et al. 2022).

Contrast Experiments

In order to find out the best way to apply the shortcut strategies, based on the model of (Xing et al. 2022), we set up three control experiments and all are based on both the grading classification task and survival prediction task. Firstly, after vectorization, we subtracted F_1 from F_2 and F_3 and named the model MLA-GNN-F1. Similarly, let F_3 minus F_2 whereas F_2 still deducting F1 we get MLA-GNN-Fp. And, to maintain the node structure of graphs, the third model is constructed by docking G_1 directly from G_2 and G_3 before linear projection, and we get MLA-GNN-G1. The experiment results are shown in Table 1 and Table 2. We train and test all the mentioned models together with the original MLA-GNN model for comparison.

Experimental results in Table 1 and Table 2 show that when the deduction of features from lower levels might decrease the performance slightly towards grading classification tasks, it improves the ability when conducting survival prediction jobs at a higher level, which might prove our suppose that there exists the redundancy of information. Also, results indicate that subtracting the original G_1 graph rather

Model	Accuracy
MLA-GNN	0.6096 ± 0.1488
MLA-GNN-F_1	0.5976 ± 0.1683
MLA-GNN-F_p	0.6030 ± 0.1527
MLA-GNN-G_1	$\mathbf{0.6174 \pm 0.0781}$

Table 1: Model performance on the histological grading task of glioma dataset, evaluated by average accuracy. The best performance is highlighted by bold text.

Model	c-index
MLA-GNN	0.6320 ± 0.1508
MLA-GNN-F_1	0.6388 ± 0.1717
MLA-GNN-F_p	0.6460 ± 0.1827
MLA-GNN-G_1	$\mathbf{0.6515 \pm 0.1595}$

Table 2: Model performance on survival prediction task on glioma dataset, measured by the average c-index metric. The best performance is highlighted by bold text.

than the transformed vector helps more, we attribute this to the necessity to hold the structures of the graph.

Ablation Experiments

After multi-level graph construction, the dimension of G_1 , G_2 and G_3 is ($V^{K\times 1}$, $E^{K\times K}$), ($V^{K\times h2}$, $E^{K\times K}$) and ($V^{K\times h3}$, $E^{K\times K}$) respectively. Thus, we have (1+h2+h3) channels in total, where 1 channel contains the original nodes' information, and $h2$ and $h3$ channels include higher-level knowledge. So as to verify whether a SE attention module is useful or not, we add a SE module based on the MLA-GNN model, named MLA-GNN-SE, and our final proposed model to perform ablation experiments.

As shown in Table 3, SE module helps to increase the performance of the model both on classification and survival prediction tasks. Together with the shortcut strategies, our proposed model outperforms more than the original MLA-GNN.

Conclusion

In this study, we propose ThinMLA based on omics data to replicate biological processes and explicitly investigate the topological data present in the co-expression gene graphs. By combining extracted features from co-expression genes, the model can hierarchically extract functional gene module-level features. However, through experiments, we display the redundancy inside different layers of features obtained from GAT layers and show the importance to construct connections between layers with different dimensions using SE attention module. Both on grading classification and survival prediction tasks on glioma dataset, our model outperforms better than the original MLA-GNN model.

Model	Classification Accuracy	Survival c-index
MLA-GNN	0.6096 ± 0.1488	0.6320 ± 0.1508
MLA-GNN-SE	0.6159 ± 0.1344	0.6397 ± 0.1554
Proposed	$\mathbf{0.6242 \pm 0.1269}$	$\mathbf{0.6551 \pm 0.1812}$

Table 3: Model performance on classification and survival prediction task on glioma dataset, measured by the average accuracy and c-index metric. The best performance is highlighted by bold text.

References

Aramburu A, Zudaire I, P. M., and et al. 2015. Combined clinical and genomic signatures for the prognosis of early stage non-small cell lung cancer based on gene copy number alterations. *BMC Genomics* 16:752.

Ben-Hamo, and et al. 2020. Predicting and affecting response to cancer therapy based on pathway-level biomarkers. *Nat Commun* 1–16.

Capper D, Jones DTW, S. M., and et al. 2018. DNA methylation-based classification of central nervous system tumours. *Nature* 555(7697):469–474.

Chuang HY, L. E.; Liu YT, L. D.; and T, I. 2007. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 3:140.

Cortés-Ciriano, I.; van Westen, G. J. P.; Bouvier, G.; Nilges, M.; Overington, J. P.; Bender, A.; and Malliavin, T. E. 2015. Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics* 32(1):85–95.

Fanfani, V.; Torne, R. V.; Lio', P.; and Stracquadanio, G. 2021. Discovering cancer driver genes and pathways using stochastic block model graph neural networks. *bioRxiv*.

G, M., and et al. 2021. Biological network analysis with deep learning. *Brief Bioinf* 1515–1530.

Gupta, A.; Malwe, A.; Srivastava, G.; and et al. 2022. MP4: a machine learning based classification tool for prediction and functional annotation of pathogenic proteins from metagenomic and genomic datasets. *BMC Bioinformatics* 23,507.

H., D. J. Y. 2012. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol* 6:92.

Hasin Y., S. M., and A., L. 2017. Multi-omics approaches to diseas. *Genome Bio* 18:83.

He Kaiming, e. a. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Hu Jie, L. S., and Sun, G. 2018. Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

ie Hao, S. C. K.; Song, N. Z. T. D. H.; and Kang, M. 2020. PAGE-Net: Interpretable and Integrative Deep Learning for Survival Analysis Using Histopathological Images and Genomic Data. *Pacific Symposium on Biocomputing*.

Iorio, F.; Knijnenburg, T. A.; Vis, D. J.; and et al. 2016. A landscape of pharmacogenomic interactions in cancer. *Cell* 166(3):740–754.

J. Hu, L. S., and Sun, G. 2018. Squeeze-and-Excitation Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7132–7141.

J, B. 2004. Predicting disease using genomics. *Nature* 429:453–456.

Jie, H.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

K. He, X. Zhang, S. R., and Sun, J. 2016. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kong, Y., and Yu, T. 2018. A graph-embedded deep feed-forward network for disease outcome classification and feature selection using gene expression data. *Bioinformatics* 34(21):3727–3737.

Kuzmin, K.; Adeniyi, A. E.; DaSouza, A. K.; Lim, D.; Nguyen, H.; Molina, N. R.; Xiong, L.; Weber, I. T.; and Harrison, R. W. 2020. Machine learning methods accurately predict host specificity of coronaviruses based on spike sequences alone. *Biochemical and Biophysical Research Communications* 533,3:553–558.

Langfelder P., H. S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559.

M., L., and W., N. 2015. Machine learning applications in genetics and genomics. *Nat Rev Genet* 16:321–332.

M., T. A.; A., K.; P., F.; and M., H. 2022. Improving and leveraging the interpretability of deep neural networks for genomics. *Stanford University*.

Muzio, G.; O'Bray, L.; and Borgwardt, K. 2020. Biological network analysis with deep learning. *Briefings in Bioinformatics* 22(2):1515–1530.

RJ, C.; MY, L.; J, W.; DFK, W.; SJ, R.; NI, L.; and F, M. 2022. Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. *IEEE Trans Med Imaging* 41(4):757–770.

Sidharth S Prakash, V. K. 2020. Breast Cancer Malignancy Prediction Using Deep Learning Neural Networks. *Proceedings of the Second International Conference on Inventive Research in Computing Applications*.

Strand, S. H., and et al. 2020. Molecular classification and biomarkers of clinical outcome in breast ductal carcinoma. *Analysis of TBCRC 038 and RAHBT cohorts*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2017. Graph attention networks.

Xing, X.; Yang, F.; Li, H.; Zhang, J.; Zhao, Y.; Gao, M.; Huang, J.; and Yao, J. 2022. Multi-level attention graph neural network based on co-expression gene modules for disease diagnosis and prognosis. *Bioinformatics* (8):8.

Yue Zhao, Ziwei Pan, S. N., and Andrew Pattison, A. P. 2020. CUP-AI-Dx: A tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. *EbioMedicine*.