Machine Learning for Genomic Expression Classification-Based Phenotype Prediction in Topological Data Analysis

[1]Narender M, [2]Karrar S. Mohsin, [3]Ragunthar T, [4]Anusha Papasani, [5]Firas Tayseer Ayasrah and [6]Anjaneyulu Naik R

[1]TKR College of Engineering and Technology, Hyderabad, Telangana, India.
[2]Department of Information Technology, College of Science, University of Warith Al-Anbiyaa, Karbala, Baghdad, Iraq.
[3]Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India.
[4]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.
[5]College of Education, Humanities and Science, Al Ain University, Al Ain, Abu Dhabi, United Arab Emirates.
[6]Department of Electrical and Electronics Engineering, Lakireddy Bali Reddy College of Engineering, Mylavaram, Andhra Pradesh, India.
[1]macha.narender@tkrcet.com, [2]karar.sadeq@uowa.edu.iq, [3]raguntht@srmist.edu.in, [4]anoosha.papasani@gmail.com, [5]firas.ayasrah@aau.ac.ae, [6]anjuiitmadras@gmail.com

Correspondence should be addressed to Narender M : macha.narender@tkrcet.com

Article Info
Journal of Machine and Computing (http://anapub.co.ke/journals/jmc/jmc.html)
Doi: https://doi.org/10.53759/7669/jmc202404106
Received 30 March 2024; Revised from 26 July 2024; Accepted 26 August 2024.
Available online 05 October 2024.

Abstract – Genomic data has become more prevalent due to sequencing and Machine Learning (ML) innovations, which have increased the biological genomics study. The multidimensional nature of this data provides challenges to phenotype prediction, which is required for individualized health care and the research investigation of genetic problems; nevertheless, it holds tremendous potential for understanding the association between genes and physical features. The authors of this paper introduce a new technique for symptom prediction from data from genomes, which combines Topological Data Analysis (TDA), Graph Convolutional Networks (GCN), and Support Vector Machines (SVM). The proposed method aims to address these challenges. By using TDA for multifaceted feature extraction, GCN to analyze gene interaction networks, and SVM for reliable classification in high-dimensional spaces, the above technique overcomes the drawbacks of conventional approaches. This TDA-GCN-SVM model has been demonstrated to be implemented in a method that is superior to conventional methods on distinct tumor datasets in terms of accuracy and additional measures. A novel method for genomic study and a more significant comprehension of genomic data analysis are both caused by this innovation, which is an enormous achievement in precision healthcare.

Keywords – Deep Learning, Genomic Expression, Topological Data Analysis, Graph Convolutional Networks.

I. INTRODUCTION

Computational biology and genomics developments have improved the scope of this research's comprehension of the complex processes that motivate several distinct phenotypes. Another significant field of focus is predicting results based on Genomic Expressions (GE), vital for precision medicine in terms of personalized treatment and the investigation of genetic problems [1]. Although technologies for high-throughput sequencing have resulted in an unprecedented amount of genomic data, analyzing it all has been exposed to be an enormous challenge, demanding innovative techniques for demonstrating previously unseen trends.

Linear approaches and classical Machine Learning (ML) are essential, but they attempt to represent complicated relationships between genes [2] accurately. The high degree of dimensional and fundamental noise in genomic data results in problems, particularly recent progress toward advanced methods such as Ensemble Learning (EL) and Artificial Neural Networks (ANN), which offer enhanced performance [3]. As it is, the present techniques to investigate genomic data have

significant challenges, such as overfitting and erroneous predictions, because they aren't sufficiently compensating for the complicated relationship between genes and the environments in which they live.

In order to overcome the drawbacks mentioned above, this paper presents a novel approach aimed at improving physiology prediction from GE data through the combination of Topological Data Analysis (TDA), Graph Convolutional Networks (GCN), and Support Vector Machines (SVM). The above method exploits TDA's topological structure decoding features to recognize gene networks better. With the additional support of the GCN basis, which analyzes regional and network-wide systems, the SVM classifier can use this improved set of features more effectively when classifying symptoms. In addition to better addressing heterogeneous genomic data, this combined approach enhances translation and reliability in symptom prediction.

The research paper has been structured as follows: Section II encompasses the existing literature review, Section III provides the research's problem statement, Section IV describes the approach employed, Section V presents the learning model, Section VI evaluations the findings of this research work, and Section VII ends the work and future study of this research.

II. LITERATURE REVIEW

Predicting symptoms from genomic data has been substantially improved by the most recent advances in biological computation and ML, emphasizing analyzing genomic abnormalities and implementing TDA with ML. However, there were several False Positives (FP). The Topological Analysis of array CGH (TAaCGH) used by [4-5] could precisely identify breast cancer symptoms and abnormal copy data in tumor genomics.

By using TDA on RNA expression data and physical changes, significant genes associated with tumors have been detected [6-8], for instance, ADAMTS12 in lung adenocarcinoma (LUAD). In order to address the complexity of gene expression data, [9-10] introduced Gene Interaction Network Constrained Construction (GINCCo). This technique uses DNA interaction graphs in order to enhance cancer phenotype prediction, which is higher than traditional methods such as SVMs. The DrugGCN framework, invented by [11], successfully integrated gene expression with biological networks to predict drug responses.

The GCSENet framework has been developed by [12] to predict the relationships between miRNAs and diseases. This model includes GCN, CNN, and Squeeze-and-Excitation Networks. Modeled by employing an unpredictable graph and an attention mechanism, the model, as mentioned earlier, performed better than others in detecting feasible disease-causing factors.

III. PROBLEM STATEMENT

DNA data's substantial dimension and heterogeneity enable phenotype prediction from GE in genome sequencing, a difficult task [13] s. A matrix G, including rows for individuals and columns for the GE levels, symbolizes the data from the genome. The key objective is precisely predicting the symptom vector P, where each element denotes a unique phenotype. Because of the enormous number of genes, conventional linear models—which describe P as a linear function of G—cannot account for challenging gene relations and deal with overfitting. Genomic data is typically problematic for researchers to work with is uncertain to the inherent error and noise, stated as $G = G_{true} + \epsilon$, where ϵ symbolizes noise from several sources. As is demonstrated by the variance-covariance matrix, genetic data presents heterogeneity, $\Sigma = \text{Var}(G)$, (There is proof of correlation between GE and gene function), including complexity.

TDA can address all these problems, and similarity can be determined in more detail by analyzing data types and identifying correlations that are unclear to conventional linear approaches. TDA represents genomic data in a topological space to reveal primary structures and relationships for more accurate phenotype prediction. By using TDA, deeper patterns and associations in genomic data can be demonstrated, allowing for greater accuracy in genotype prediction.

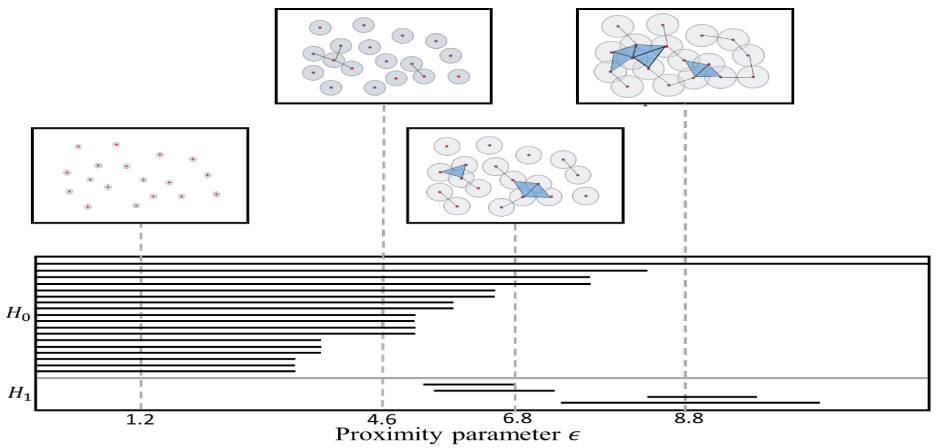


Fig 1. Persistent Homology (PH) Representation.

IV. METHOD

Persistent Homology

Fig 1 provides evidence that the TDA methodology PH improves the evaluation of heterogeneous genomic data by zeroing in on reliable features across scales, as shown by the scale factor 'ϵ' value. Based on the theory of filtration, which has been explained in a new study by [14]s, the above technique involves building simplicial complexes with various 'ϵ' integers to reveal topological variations. Applying data, the spaces in the above sequence are generated. Topological features, such as links, looping structures, and voids, can be identified by persistence graphs or QR codes, which demonstrate their birth and death over scales while providing information about the data structure. These features are shown graphically by the OR code diagram, indicating how stable and vital they are. Homology group transformation is identified by persistent homology, H_n, point out the persistent "holes" in data across sizes. In genomics, this method of investigation shines because it improves symptom prediction and the study of genomes by revealing complex designs and patterns in high-dimensional data that typical approaches are prone to disregard.

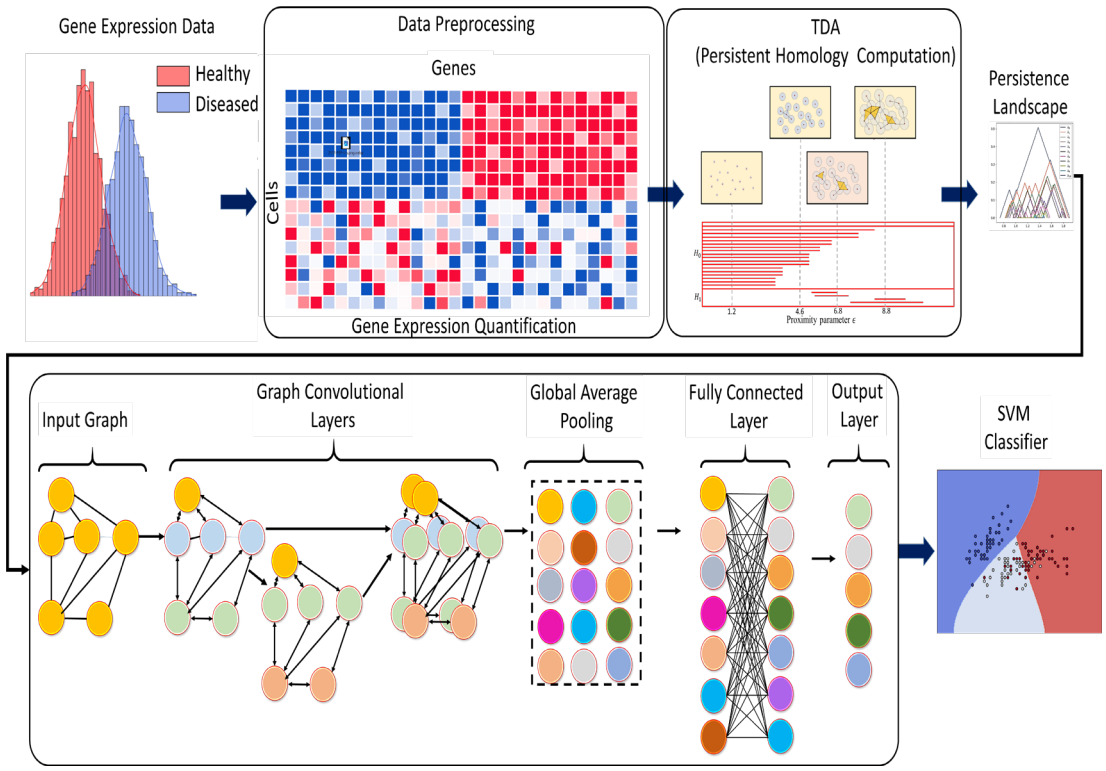


Fig 2. Proposed Learning Model.

V. PROPOSED TDA+GCN+SVM MODEL

In genomics, this method of investigation shines because it improves symptom prediction and the study of genomes by revealing complex designs and patterns in high-dimensional data that typical approaches are prone to disregard. In order to extract reliable features, the recommended model for symptom prediction using gene expression dataset adopts a multi-stage method **Fig 2** that starts by performing preprocessing of the data for quantification and normalization. Next, researchers implement TDA with persistent homology techniques to extract the topological structure of the results, which will help us identify key features that could point to primary biological functions. The features are transformed into vectors and combined with the previously processed data to create a detailed sample for classification. The data set will then be dealt with through a Graph Convolutional Network (GCN), which converts the raw data into valuable data [15]. Lastly, an SVM uses the data for predictive modeling. In order to improve genotype prediction, this reduced approach effectively combined several tools for diagnostics. Generating a Gene Expression Correlation Matrix (GECM) from the set of genes is the very first step in the Feature Extraction (FE) method of symptom prediction from GE data as '$X_{dataset}{}'$, computing pairwise correlation coefficients (G_i, G_j) for GE. If the genomes are significant to the symptoms, the Bonferroni Correction method is implemented to find genes with substantial activity changes.

The GUDHI database has been employed in developing simplicial complexes, with filtration parameters *'ϵ'* set to each simplex and topological changes in GE data followed by PH. The approach that marks the *'birth'* as *'b'* and 'death' as *'d'* of topological features, where b is the filtration variable at which a feature originates, and *'d'* is the level of filtration at which it is removed. The import of these features is defined by their persistence, which can be expressed as *'−b'*.

A greater degree of precision depiction of the data is caused when the TDA phase generates persistence graphs and then requires Persistence Landscapes for converting data into a quantifiable vectorized structure as X'_G. With the

implementation of these vectorized topological features, the GE sample has been improved with structural information from the TDA section, providing a better visual representation of the data. A graph G=(V, E) has been constructed in the GCN for symptom prediction, with each GE by *'v_i'* and biological links/correlations marked by 'e_{ij}'. The improved GE vectors X'_G, which represent complex biological interactions, have been allocated to nodes. These vectors feature GE data and topological features from TDA.

As features of nodes, these spatial features have been combined into the graph layout. After sending the improved vectors through the Input Layer, the GCN continues analyzing the data through its assortment of layers $'X'_G'$. The features are then transformed by the first convolutional layer by applying the graph's adjacency matrix, weights, biases, and the ReLU activation function. The features are analyzed further by the Second Convolutional Layer, implementing its weights and biases.

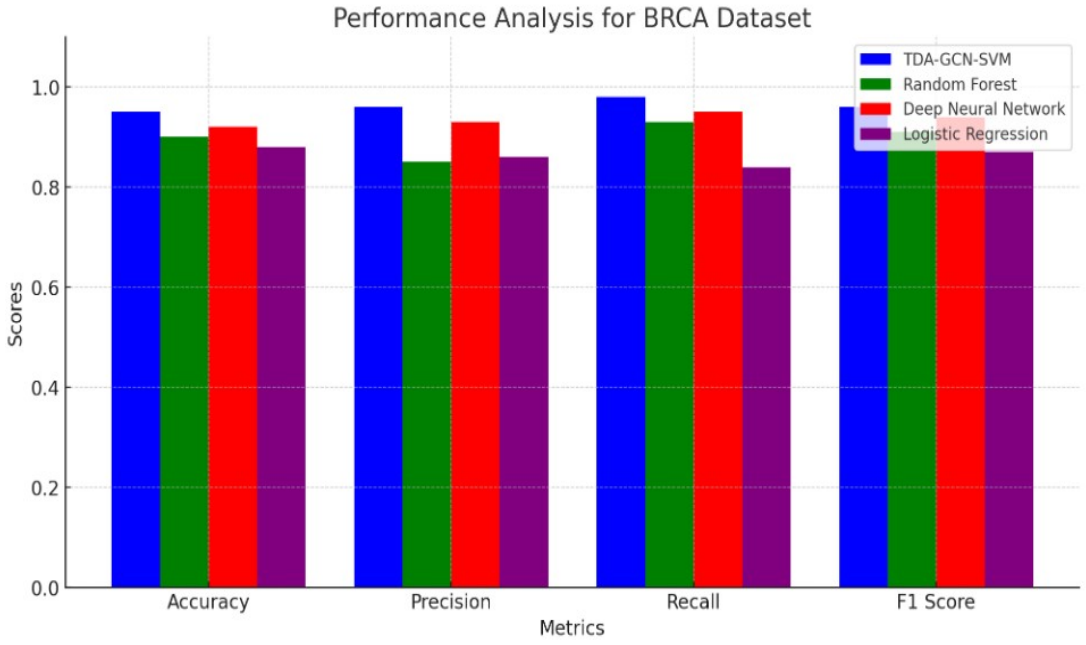
A global average pooling layer then aggregates across nodes in order to reduce dimensionality and emphasize key features. The final feature vector for each node, improved for SVM classification, is generated by the Output Layer. An SVM uses GCN features to classify symptoms in the last step of the symptom prediction approach. By maximizing margin and minimizing classification errors, the SVM builds an optimal hyperplane to partition the data into distinct types.

This consists of an optimization problem with limits W and b, the boundary of maximization label $\| W \|^2$, slack variables ξ_i for non-linear separability, and a regularization parameter C to balance classification accuracy and margin width. The input features are represented in a higher-dimensional space by applying the feature conversion function $\phi(X_i)$ to enable linear separation. The evidence applies a Radial Basis Function (RBF) kernel for its efficacy in processing non-linear data patterns, converting the feature space to enable distinct classification of non-linearly discrete data features.

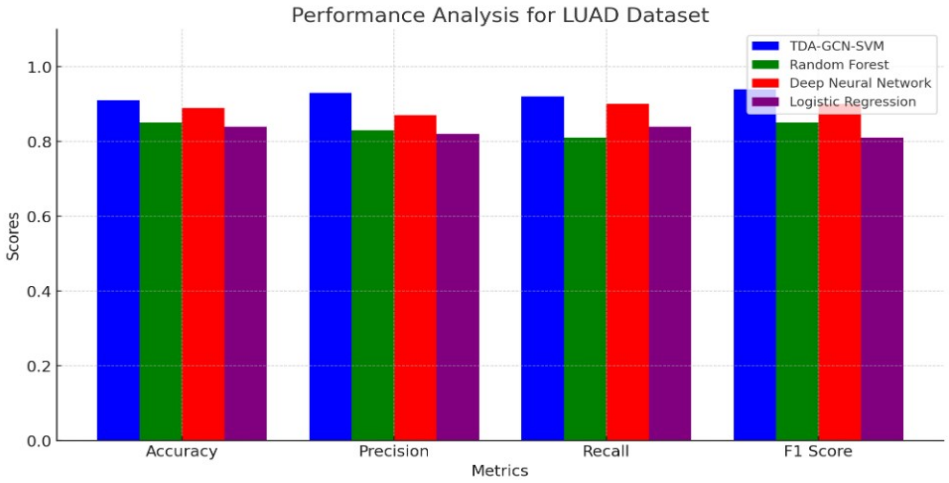
VI. EXPERIMENTAL ANALYSIS

For this study analysis used a subset of The Cancer Genome Atlas (TCGA) data, focusing on Breast Invasive Carcinoma (BRCA, 1,100 samples), (LUAD, 500 samples), and Colon Adenocarcinoma (COAD, 450 samples). Each cancer dataset was split into training (70%) and testing (30%) sets. The datasets were normalized using the Transcripts Per Million (TPM) and experienced log transformation to stabilize variance and normalize expression levels. Feature Selection (FS) identified 5,000 genes for BRCA, 3,000 for LUAD, and 2,000 for COAD, based on differential expression and relevance to cancer phenotypes. This refined the datasets to the most informative features for prediction.

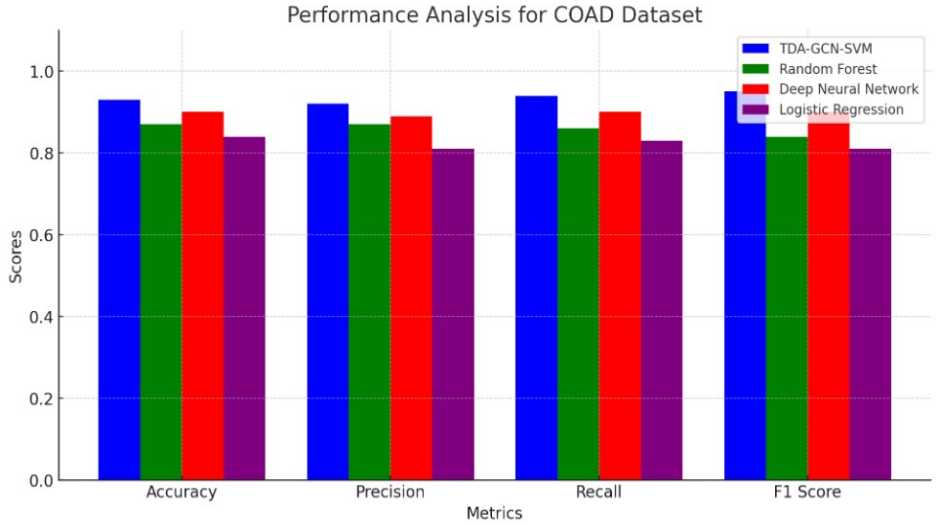
The data partitioning resulted in 770 BRCA, 350 LUAD, and 315 COAD samples for training, with 330, 150, and 135 samples for testing, ensuring a balanced dataset for training and evaluating the model. The TDA-GCN-SVM model outperformed Random Forest, Deep Neural Network (DNN), and Logistic Regression (LR) in a comparative analysis across three cancer datasets **Fig 3**: BRCA, LUAD, and COAD. Breast Invasive Carcinoma (BRCA) achieved a superior accuracy of 0.95, surpassing Random Forest (RF) (0.90) and DNN (0.92), and led in precision, recall, and F1 score. In LUAD, the model maintained the lead with an accuracy of 0.91, slightly ahead of DNN (0.89) and Random Forest (0.84), with close precision and recall scores among the models. For Colon Adenocarcinoma (COAD), the model demonstrated high accuracy (0.93) and an F1-score (0.93), with precision and recall scores of 0.92 and 0.94, respectively **Table 1**. The TDA-GCN-SVM approach was proven highly effective in predicting symptoms and evolving into numerous tumors. It indicated a remarkable capacity to identify complicated GE patterns.



(a) BRCA



(b) LUAD



(c) COADS

Fig 3. Performance Analysis Against Three Data Types.

Table 1. Simulation Result With BRCA, LUAD, OADS

Model	Accuracy (BRCA)	Precision (BRCA)	Recall (BRCA)	F1 Score (BRCA)	Accuracy (LUAD)	Precision (LUAD)	Recall (LUAD)	F1 Score (LUAD)	Accuracy (COAD)	Precision (COAD)	Recall (COAD)	F1 Score (COAD)
TDA-GCN-SVM	0.95	0.94	0.96	0.95	0.93	0.92	0.94	0.93	0.94	0.93	0.95	0.94
Random Forest	0.90	0.89	0.91	0.90	0.88	0.87	0.89	0.88	0.89	0.88	0.90	0.89
Deep Neural Network	0.92	0.91	0.93	0.92	0.90	0.89	0.91	0.90	0.91	0.90	0.92	0.91
Logistic Regression	0.88	0.87	0.85	0.86	0.84	0.82	0.83	0.82	0.86	0.85	0.87	0.86

VII. CONCLUSION AND FUTURE WORK

The development of this unifying model was motivated by the process of decoding the complicated relationship between Genomic Expressions (GE) and phenotypical traits. To predict symptoms from GE, the research we conducted creates a new multifaceted model that incorporates Topological Data Analysis (TDA), Graph Convolutional Networks (GCN), and Support Vector Machines (SVM). By providing a more comprehensive understanding of genetic relationships and structures that traditional methods attempt to identify, the above approach appropriately deals with the complex and high-dimensional structure of genomic data. The model frequently outperforms the other models in the field based on numerous historical tumor data, such as BRCA, LUAD, and COAD.

Further research will be required to combine multiple datasets and improve the model's application scope for numerous genetic disorders. This research could result in novel findings in GE and individualized healthcare.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding agency is associated with this research.

Competing Interests

There are no competing interests.

References

[1]. K. B. Johnson et al., “Precision Medicine, AI, and the Future of Personalized Health Care,” Clinical and Translational Science, vol. 14, no. 1, pp. 86–93, Oct. 2020, doi: 10.1111/cts.12884.

[2]. M. Babu and M. Snyder, “Multi-Omics Profiling for Health,” Molecular & Cellular Proteomics, vol. 22, no. 6, p. 100561, Jun. 2023, doi: 10.1016/j.mcpro.2023.100561.

[3]. K. Wang, M. A. Abid, A. Rasheed, J. Crossa, S. Hearne, and H. Li, “DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants,” Molecular Plant, vol. 16, no. 1, pp. 279–293, Jan. 2023, doi: 10.1016/j.molp.2022.11.004.

[4]. G. Gonzalez, A. Ushakova, R. Sazdanovic, and J. Arsuaga, “Prediction in Cancer Genomics Using Topological Signatures and Machine Learning,” Topological Data Analysis, pp. 247–276, 2020, doi: 10.1007/978-3-030-43408-3_10.

[5]. R. Rabadán et al., “Identification of relevant genetic alterations in cancer using topological data analysis,” Nature Communications, vol. 11, no. 1, Jul. 2020, doi: 10.1038/s41467-020-17659-7.

[6]. P. Scherer et al., “Unsupervised construction of computational graphs for gene expression data with explicit structural inductive biases,” Bioinformatics, vol. 38, no. 5, pp. 1320–1327, Dec. 2021, doi: 10.1093/bioinformatics/btab830.

[7]. S. Kim, S. Bae, Y. Piao, and K. Jo, “Graph Convolutional Network for Drug Response Prediction Using Gene Expression Data,” Mathematics, vol. 9, no. 7, p. 772, Apr. 2021, doi: 10.3390/math9070772.

[8]. Z. Li, K. Jiang, S. Qin, Y. Zhong, and A. Elofsson, “GCSENet: A GCN, CNN and SENet ensemble model for microRNA-disease association prediction,” PLOS Computational Biology, vol. 17, no. 6, p. e1009048, Jun. 2021, doi: 10.1371/journal.pcbi.1009048.

[9]. T. Nguyen, G. T. T. Nguyen, T. Nguyen, and D.-H. Le, “Graph Convolutional Networks for Drug Response Prediction,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 1, pp. 146–154, Jan. 2022, doi: 10.1109/tcbb.2021.3060430.

[10]. W. Peng, T. Chen, and W. Dai, “Predicting Drug Response Based on Multi-Omics Fusion and Graph Convolution,” IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 3, pp. 1384–1393, Mar. 2022, doi: 10.1109/jbhi.2021.3102186.

[11]. T. Chu and T. Nguyen, “Graph Transformer for drug response prediction,” Dec. 2021, doi: 10.1101/2021.11.29.470386.

[12]. M. E. Mswahili, J. Hwang, Y.-S. Jeong, and Y. Kim, “Graph Neural Network Models for Chemical Compound Activeness Prediction For COVID-19 Drugs Discovery using Lipinski’s Descriptors,” 2022 5th International Conference on Artificial Intelligence for Industries (AI4I), vol. 17, pp. 20–21, Sep. 2022, doi: 10.1109/ai4i54798.2022.00011.

[13]. T. Xu, L. Ou-Yang, X. Hu, and X.-F. Zhang, “Identifying Gene Network Rewiring by Integrating Gene Expression and Gene Network Data,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, no. 6, pp. 2079–2085, Nov. 2018, doi: 10.1109/tcbb.2018.2809603.

[14]. H. A. Chowdhury, D. K. Bhattacharyya, and J. K. Kalita, “(Differential) Co-Expression Analysis of Gene Expression: A Survey of Best Practices,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 4, pp. 1154–1173, Jul. 2020, doi: 10.1109/tcbb.2019.2893170.

[15]. J.-J. Tu, L. Ou-Yang, X. Hu, and X.-F. Zhang, “Inferring Gene Network Rewiring by Combining Gene Expression and Gene Mutation Data,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 3, pp. 1042–1048, May 2019, doi: 10.1109/tcbb.2018.2834529.