LITERATURE REVIEW OF PREDICTING DISEASE GENE ASSOCIATION

Nihala K P

[1] Nihala K P,Assistant Professor, Computer Science and Engineering,Vimal Jyothi Engineering College,Kerala,India

ABSTRACT

Disease-gene interaction prediction plays an important role in understanding diseases and supporting drug discovery. This project uses Graph Autoencoders (GAEs) to model these interactions by analysing biological networks. GAEs are effective for this task because they can capture complex graph structures and learn meaningful representations of relationships between genes and diseases.

In this approach, a graph is constructed where nodes represent genes and diseases, and edges represent known interactions between them. A Graph Convolutional Network (GCN)-based encoder is used to learn structural features from the graph, while a decoder reconstructs the graph to identify potential new interactions. The model is trained by minimizing reconstruction loss, making it suitable for handling large-scale datasets.

Future work will focus on improving model interpretability and incorporating domain-specific knowledge to enhance prediction performance.

Keyword : -*Graph Convolution Network, Graph Auto Encoder, Disease-gene interaction.*

1. INTRODUCTION

In the field of bioinformatics, predicting disease-gene interactions plays a crucial role in understanding the underlying causes of diseases and identifying potential treatment strategies. Discovering new relationships between genes and diseases can support drug development, biomarker identification, and personalized medicine. However, this task is challenging due to the complexity of biological systems and the large volume of genomic data.

Traditional approaches for predicting disease-gene interactions mainly rely on manually curated databases or basic statistical methods. These methods are often limited in their ability to capture complex and hidden relationships within large biological networks.

To address these limitations, graph-based models have emerged as an effective solution. In this approach, genes and diseases are represented as nodes, and their interactions are represented as edges in a graph structure. Graph Autoencoders (GAE) are used to learn meaningful representations of this graph data by capturing the structural relationships between nodes. These models are further enhanced using Graph Convolutional Networks (GCN), which help in extracting important features from the graph.

In this project, a GAE-based model is used to predict disease-gene interactions by learning patterns from existing known relationships. The model analyzes the graph structure and identifies potential new links between genes and diseases. This approach helps in better understanding biological connections and supports further research in the biomedical field.

Future improvements can focus on enhancing the interpretability of the model and incorporating domain-specific biological knowledge to improve prediction performance and reliability.

2. LITERATURE REVIEW

Study [1] introduces an advanced method called Random Walk with Restart on Multiplex and Heterogeneous Networks (RWRMH) for identifying disease-associated genes. This method combines multiple biological networks such as protein-protein interaction (PPI), pathway interaction, and gene co-expression into a multiplex structure. It also includes a disease-disease similarity network. By performing random walks across these interconnected layers, the model captures complex relationships between genes and diseases. This helps improve the accuracy of identifying important genes related to diseases. The main advantage of this approach is that it integrates different types of biological data, providing a more complete understanding of disease mechanisms.

However, the method has some limitations. It depends heavily on existing biological data, which makes it difficult to identify genes related to rare or less-studied diseases. Additionally, handling multiple large networks increases computational complexity, making it less suitable for real-time applications.

In study [2], the authors propose a random walk-based approach that integrates multiple omics datasets to improve disease-gene prediction. The model combines different sources of data, including protein-protein interactions, gene expression data, disease-term relationships, and drug-target information. By using biological random walks (BRW), the method ranks genes based on their likelihood of being associated with a disease. This integration of multiple datasets helps the model capture more detailed and meaningful biological relationships, leading to better prediction accuracy. However, the method relies on existing databases, which may introduce bias toward already known gene-disease relationships. This can limit the discovery of new or less-represented genes.

Study [3] presents a method called Random Walk on Multigraphs (RWM), which combines different biological networks such as phenotype networks, protein-protein interaction networks, and functional similarity networks into a single multigraph. This integration improves the robustness of the model and helps in better identification of disease-associated genes. The approach allows information to flow across multiple types of biological data, enhancing prediction performance. However, one drawback is that all data sources are treated equally, even though some sources may be more important than others. Also, the transitions between layers are handled independently, which may not fully capture complex biological dependencies.

In paper [4], a deep learning-based model called ModulePred is introduced for predicting disease-gene associations. This method uses graph-based techniques along with functional modules to improve prediction performance. It mainly relies on protein-protein interaction networks and applies graph augmentation to enhance learning. The model shows strong performance in terms of prediction metrics such as precision and recall. It is also capable of identifying new candidate genes. However, the model has some limitations. It requires high computational resources and performs poorly when tested on external datasets, indicating limited generalization ability.

Study [5] focuses on the use of Graph Convolutional Networks (GCNs) for predicting gene-disease interactions. This method uses datasets such as DisGeNET and DGAssocMiner to model complex biological relationships. GCNs are effective in capturing structural information from graph data, making them suitable for large and complex datasets. The model shows good accuracy and scalability. However, it requires a large amount of high-quality data, and preprocessing such data can be time-consuming. Additionally, the computational cost is high, which may limit its use in environments with limited resources.

In study [6], a Graph Neural Network (GNN)-based framework called gene-DRAGNN is proposed for prioritizing disease-related genes. The model integrates various biological data sources such as gene ontology, gene-gene interaction networks, and gene-disease associations. It is capable of capturing complex relationships and shows strong predictive performance. The method is also scalable and can handle large datasets efficiently. However, it lacks diverse edge features, which limits its ability to fully represent all types of biological interactions. This can affect the completeness of predictions.

Paper [7] introduces XGDAG, a framework that combines Graph Neural Networks with explainability techniques. The main goal of this approach is to make model predictions more transparent and understandable. It uses Positive-Unlabeled (PU) learning, where only known positive data is available, and assigns labels to unknown data using propagation methods. The model uses GraphSAGE for prediction and applies explainability methods like GNNExplainer to identify important genes and subgraphs. This improves trust in the model's predictions. Although the method performs well and provides useful insights, it is more complex and requires careful implementation.

Study [8] is a survey paper that provides an overview of different Graph Neural Network architectures. It categorizes GNNs into spectral and spatial methods and explains their working principles. The paper also discusses various applications of GNNs, including bioinformatics, social networks, and image processing. It highlights the strengths of GNNs in handling structured and relational data. Additionally, it identifies challenges such as scalability, over-smoothing, and lack of theoretical understanding. However, the paper focuses more on concepts and does not provide implementation details or direct comparisons of models.

In study [9], a method using privileged information and heteroscedastic dropout (PIHD) is proposed for disease-gene prediction. This approach uses additional biological information such as Gene Ontology and Disease Ontology to improve prediction accuracy. It also handles missing data effectively using dropout techniques. While the method shows good performance, it has some drawbacks. The use of privileged information may lead to data leakage, affecting the reliability of results. It also requires more computational resources and may struggle with sparse datasets.

Finally, paper [10] introduces GenePredictKG, a multi-relational Graph Convolutional Network model that uses knowledge graphs for predicting disease-gene associations. It integrates different biological data sources such as Human Phenotype Ontology and Gene Ontology. The model shows very high accuracy and can predict new gene-disease relationships effectively. It is also scalable and suitable for large datasets. However, it treats all

relationships equally, which may reduce accuracy when some connections are more important. Additionally, the model faces challenges related to data imbalance, which can affect prediction quality.

4. CONCLUSIONS

From the above studies, it is clear that predicting disease-gene interactions is an important and challenging task in bioinformatics. Many traditional approaches, such as random walk-based methods, have been widely used and show good performance by utilizing biological networks like protein-protein interactions, gene expression data, and phenotype information. These methods are effective in capturing relationships between genes and diseases, but they mainly depend on existing datasets and often struggle to identify new or less-known associations.

With the advancement of deep learning, Graph Neural Networks (GNNs), Graph Convolutional Networks (GCNs), and related models have become more popular. These models are capable of handling complex and large-scale biological data by learning hidden patterns and structural relationships in graph-based representations. Some approaches also integrate multiple data sources or use knowledge graphs, which improves prediction accuracy and provides a more complete understanding of biological systems.

However, several limitations are still observed across the existing methods. Many models require large and high-quality datasets, making data collection and preprocessing difficult. High computational cost and scalability issues are also common challenges. In addition, some models lack interpretability, making it difficult to understand how predictions are made. There are also issues such as data imbalance, bias toward known data, and limited ability to generalize to new datasets.

Therefore, there is a need for more efficient and scalable models that can handle complex biological data while also improving interpretability and reducing dependency on existing knowledge. These gaps highlight the importance of developing advanced graph-based approaches, such as Graph Autoencoders, which can better learn hidden relationships and predict novel disease-gene interactions.

6. REFERENCES

[1] A. Valdeolivas, ´E. Remy, L. Tichit, G. Odelin, C. Navarro, S. Perrin,P. Cau, N. Levy, and A. Baudot, “Prediction of Disease-associatedGenes by advanced Random Walk with Restart on Multiplex and Heterogeneous Biological Networks,” in JOBIM 2017, (Lille, France),July 2017.

[2] M. Gentili, L. Martini, M. Sponziello, and L. Becchetti, “Biological random walks: multi-omics integration for disease gene prioritization,”vol. 38, pp. 4145–4152, 07 2022.

[3] Y. W. Chandra and S. Suyanto, “Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data,” 2012.

[4] L. J. X. J. S. H. W. S. S. X. Jia X, Luo W, “A deep learning framework for predicting disease-gene associations with functional modules and graph augmentation,” 2024.

[5] C. M. Cinaglia P, “Identifying candidate gene-disease associations via graph neural networks,” 2023.

[6] A. Altabaa, D. Huang, C. Byles-Ho, H. Khatib, F. Sosa, and T. Hu,“genedragnn: Gene disease prioritization using graph neural networks,” in 2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2022.

[7] A. Mastropietro, G. De Carlo, and A. Anagnostopoulos, “XGDAG: explainable gene–disease associations via graph neural networks,” vol. 39,08 2023.

[8] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, “Graph neural networks: A review of methods and applications,” AI Open, vol. 1, pp. 57–81, 2020.

[9] J. Shu, Y. Li, S. Wang, B. Xi, and J. Ma, “Disease gene prediction with privileged information and heteroscedastic dropout,” Bioinformatics, vol. 37, pp. i410–i417, 07 2021.

[10] D. P. X. R. Gao Z, Pan Y, “A knowledge graph-based disease-gene prediction system using multi-relational graph convolution networks.,” 2023.

[11] M. Chatzianastasis, M. Vazirgiannis, and Z. Zhang, “Explainable Multi- layer Graph Neural Network for cancer gene prediction,” Bioinformatics, vol. 39, p. btad643, 10 2023.