

<Conference abbreviation>

<Conference Series name>

<Volume number and Year> <DOI Number>

Review of Deep Learning Transformer Architectures for Multi-Omics Integration in Rare Disease Diagnosis

Pallavi Dilip Waghmare¹, Dr. Vijay N Patil²

¹, ² Pimpri Chinchwad Education Trust's Pimpri Chinchwad University

Email ID: ¹pallavi.waghmare23@pcu.edu.in, ²vijay.patil@pcu.edu.in

Abstract. Rare diseases affect a small percentage of the population individually, but collectively they impact millions of people worldwide and often remain difficult to diagnose due to complex genetic and molecular mechanisms. The rapid advancement of high-throughput sequencing technologies has enabled the generation of large-scale multi-omics data, including genomics, transcriptomics, proteomics, and metabolomics, which provide deeper insights into disease mechanisms. Therefore, researchers try to combine these datasets to obtain a better understanding of disease. However, integration is difficult because the data are large, complex, and very different from one another. In recent years, transformer models have been increasingly used for multi-omics integration. These models are capable of learning relationships between different omics layers and usually give better prediction results than conventional machine learning methods. This review presents the important transformer-based methods used for disease prediction and discusses their advantages in handling complex biological data. Despite their better performance, transformer models still face some challenges. Limited sample size, high computational requirements and poor interpretability remain major issues. Hence, further research is required before these methods can be applied effectively in clinical practice.

Keywords. Deep Learning, Transformer Models, Multi-Omics Data Integration, Rare Disease Diagnosis, Biomedical Data Analysis.

1. INTRODUCTION

Rare diseases are difficult to diagnose because their symptoms often resemble common disorders, leading to delayed or incorrect diagnosis [1]. Since many rare diseases involve complex genetic interactions, multi-omics analysis [3] has become important. Transformer models can capture complex relationships among genes, proteins, and metabolic pathways [8]. With larger datasets and improved interpretability, transformer-based multi-omics approaches could support earlier and more accurate rare-disease diagnosis.

2. PROBLEM STATEMENT

Diagnosing rare diseases is often difficult because these disorders usually involve several biological changes at the same time. Modern sequencing technologies can now generate large amounts of biological data. Researchers are able to collect genomic, transcriptomic, proteomic and metabolomic information from the same patient. This problem becomes bigger in rare diseases because only a small amount of patient data is available.

To solve this problem, researchers have started using transformer-based deep learning models. These models can study many datasets together and can find patterns that normal methods may not notice. They are especially useful for finding connections between genes, proteins and other biological factors.

3. OBJECTIVES

1. To understand how multi-omics data helps in better diagnosis and understanding of rare diseases.
2. To study the deep learning methods used for combining multi-omics data in biomedical research.
3. To study transformer models and understand how they work with complex biological data.
4. To understand how transformer models improve multi-omics data integration and disease prediction.
5. To study the problems, limitations and future scope of transformer-based multi-omics methods.

4. LITERATURE SURVEY

Paper 1: This paper was written by Shaza B. Zaghlool and Omneya Attallah in 2022. The authors said that if only one dataset is studied, then some important part of the disease may be missed. Because of that, they suggested deep learning. According to them, deep learning can find patterns which are hard to see normally.

Paper 2: Jael Sanyanda Wekesa and Michael Kimwele wrote this paper in 2023. In this paper, they talked about CNN, RNN and graph neural network. They felt these methods are useful because they understand difficult data better than normal computer methods.

Paper 3: Flavio Sartori and his group wrote about cancer research in 2025. The authors said that if doctors study gene data and clinical data together, then they can understand cancer in a much better way.

Paper 4: This paper was written by Akshay Krishnan Pushparaj and Malarmathi Muthukumar in 2026. The paper discussed variational autoencoders and graph neural networks. The authors said that these methods can help in finding biomarkers and understanding disease.

Paper 5: Jael Sanyanda Wekesa and Michael Kimwele again explained one important thing. In the paper, they said that deep learning can help in disease prediction and treatment planning. According to them, this method may become more useful in the future.

5. PROPOSED SYSTEM

The system collects gene, protein, and clinical data from medical databases and preprocesses it to identify important features. The data is then integrated to capture relationships among genes, proteins, and metabolic pathways. This combined data is analyzed by a transformer model to identify hidden patterns and predict possible rare diseases from new patient data.

SYSTEM DESIGN

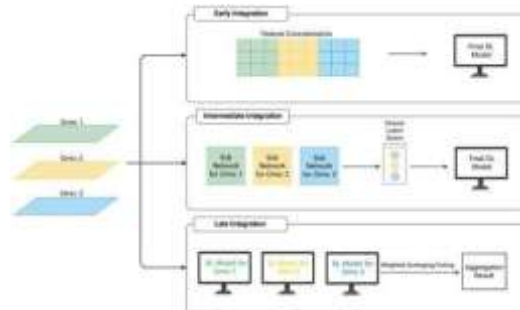


Fig 5.1. System Architecture

The diagram in Fig. 1 shows three ways therefore, deep learning uses three methods: early integration, intermediate integration and late integration.

5.1 Early Integration (Feature-Level Integration)

Early integration is the easiest method. In this approach, all the data from Omic 1, Omic 2 and Omic 3 is joined at the beginning itself.

5.2 Intermediate Integration (Representation-Level Integration)

Intermediate integration is a better and more advanced method. Here, each omics dataset is first studied separately.

5.3 Late Integration (Decision-Level Integration)

Late integration works in a different way. In this method, every omics dataset is given to a separate deep learning model.

6. EXPECTED OUTCOME

The proposed system is expected to help doctors diagnose rare diseases in a better way. Rare diseases are difficult to identify because they involve many changes inside the body. If only one type of data is used, such as genes or proteins, then some important information may be missed. Therefore, the system studies genomics, transcriptomics, proteomics and metabolomics together. This gives a more complete idea about the disease.

6.1 Improved Diagnostic Accuracy

The proposed system uses different types of biological data together, so it can find disease patterns more clearly. Because of this, the chances of wrong diagnosis become less.

6.2 Effective Multi-Omics Data Integration

The proposed model joins all these datasets into one system. This helps in understanding how different parts of the body are connected during a disease.

6.3 Identification of Important Biomarkers

By studying all the omics data together, the system may find which biomarkers are strongly related to a rare disease. This can help doctors detect the disease earlier.

6.4 Better Prediction Performance

It can focus on the most important information and ignore less useful data. Because of this, the model may give more accurate predictions. During testing, the system is expected to show better accuracy, precision, recall and F1-score.

6.5 Support for Precision Medicine

The system may also help in precision medicine. Precision medicine means giving treatment according to the condition of a particular patient. Since the system studies many biological details of the patient, it may help doctors choose the best treatment for that person.

7. CONCLUSION

The proposed system uses many kinds of biological data together. It studies genomics, transcriptomics, proteomics and metabolomics in one place. This helps researchers and doctors understand the disease in a better way. The transformer model is useful because it can find connections between genes, proteins and other changes inside the body. These connections are often difficult to find with normal methods. Because of this, the system may help doctors detect rare diseases faster and more accurately. So, this method may become useful in future healthcare. It can help doctors, researchers and hospitals understand rare diseases in a better way.

8. FUTURE SCOPE

In future, the system can include more types of data. Along with gene and protein information, researchers can also add medical images, patient history and other biological data to improve diagnosis.

The system may also become better if more patient data is collected. Rare disease data is usually less, so larger datasets from different hospitals and countries may help improve the model.

9. REFERENCES

- [1] M. Ashburner et al., "Gene ontology: Tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.
- [2] S. B. Zaghlool and O. Attallah, "A review of deep learning methods for multi-omics integration in precision medicine," in *Proc. IEEE Int. Conf. Big Data*, 2022.
- [3] J. S. Wekesa and M. Kimwele, "A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment," *Frontiers in Genetics*, vol. 14, p. 1199087, 2023.
- [4] F. Sartori et al., "A comprehensive review of deep learning applications with multi-omics data in cancer research," *Genes*, vol. 16, no. 6, p. 648, 2025.
- [5] A. K. Pushparaj and M. Muthukumar, "Deep learning architectures for multi-omics data integration: Bridging biomarker discovery and clinical translation," *Preprints*, 2026.
- [6] I. Subramanian et al., "Multi-omics data integration, interpretation, and its application," *Bioinformatics and Biology Insights*, vol. 14, 2020.

- [7] M. D. Ritchie et al., "Methods of integrating data to uncover genotype–phenotype interactions," *Nature Reviews Genetics*, vol. 16, no. 2, pp. 85–97, 2015.
- [8] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biology*, vol. 18, no. 1, p. 83, 2017.
- [9] K. Chaudhary et al., "Deep learning–based multi-omics integration for cancer survival prediction," *Bioinformatics*, vol. 34, no. 14, pp. i446–i454, 2018.
- [10] R. Argelaguet et al., "Multi-omics factor analysis—A framework for unsupervised integration of multi-omics data sets," *Molecular Systems Biology*, vol. 14, no. 6, 2018.

Biographies



Pallavi Dilip Waghmare received the Bachelor's Degree in Computer Engineering from Sanjivani College of Engineering Kopergaon in 2020 and the Master's Degree in Computer Engineering from Amrutvahini College of Engineering in 2022. She is currently pursuing PhD from PCET's Pimpri Chinchwad University, Pune and working as an Assistant Professor at the Department of Computer Engineering, Faculty of Engineering, SVIT, Chincholi.



Mr. Vijay N. Patil received the Bachelor's Degree in Electronics Engineering from Shivaji University W.I.T. Solapur in 1992, the Master's Degree in Electronics Engineering from B.A. Marathwada University (Govt. College, Aurangabad) in 2000 and the PhD (Electronics Engineering) from North Maharashtra University, Jalgaon. He is currently working as Dean, School of Engineering and Technology, and PhD Program Director, PCET's Pimpri Chinchwad University, Pune