International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(6s) 2026 pp. 144-151
https://doi.org/10.70917/ijcisim-2026-2873
Article

INTEGRATING MULTI-OMICS AND EHR DATA FOR PERSONALIZED DISEASE PREDICTION USING DEEP LEARNING IN DATA WAREHOUSING AND MINING ENVIRONMENTS: A REVIEW

Shyam Parveen B.[1], M. Elamparithi[2], V. Anuratha[3]

[1]Department of Computer Science, Kamalam College of Arts and Science, Anthiyur, Udumalpet, Bharathiar University, Coimbatore, Tamil Nadu, India.
Email: shyampraveen767@gmail.com
[2]Department of Computer Science, Kamalam College of Arts and Science, Anthiyur, Udumalpet, Bharathiar University, Coimbatore, Tamil Nadu, India.
Email: profelamparithi@gmail.com
[3]Department of Computer Science, Kamalam College of Arts and Science, Anthiyur, Udumalpet, Bharathiar University, Coimbatore, Tamil Nadu, India.
Email: profanuratha@gmail.com

Abstract: The advent of precision medicine has necessitated the integration of heterogeneous biomedical data sources to unravel the complex mechanisms underlying human diseases. While high-throughput technologies have generated vast amounts of multi-omics data (genomics, transcriptomics, proteomics) and Electronic Health Records (EHRs) provide rich phenotypic information, the effective fusion of these modalities remains a significant challenge. This review paper critically analyses the current state of "Integrating Multi-Omics and EHR Data" for personalized disease prediction, with a specific focus on Deep Learning (DL) methodologies within Data Warehousing and Mining frameworks. We examine recent advancements in data fusion strategies Early, Intermediate, and Late and evaluate the efficacy of deep neural architectures, including Multi-modal Autoencoders, Graph Convolutional Networks (GCNs), and Transformer-based models. Furthermore, the review identifies critical gaps in current data warehousing infrastructures regarding their ability to handle the high dimensionality and sparsity of omics data alongside the unstructured nature of clinical notes. By synthesizing findings from recent high-impact literature (2020–2025), we propose a unified, scalable framework that leverages advanced data mining techniques to bridge the gap between molecular biology and clinical informatics for accurate, real-time personalized healthcare.
Keywords: Multi-omics; Electronic Health Records (EHR); Personalized Disease Prediction; Deep Learning; Precision Medicine; Data Warehousing; Data Mining; Explainable AI.

1. INTRODUCTION

The paradigm of modern healthcare is shifting rapidly from a "one-size-fits-all" approach to Precision Medicine, where disease prevention and treatment strategies are tailored to the unique genetic, environmental, and lifestyle makeup of each individual [1], [2]. This transformation is driven by the explosive growth of biological big data, specifically in two domains: Multi-Omics (encompassing genomics, transcriptomics, epigenomics, and proteomics) and Electronic Health Records (EHRs). While multi-omics data offers a granular view of the molecular underpinnings of disease, EHRs provide a longitudinal phenotypic perspective, capturing clinical history, lab results,

and imaging data [3]. The integration of these distinct yet complementary modalities holds the promise of significantly improving predictive accuracy for complex diseases such as cancer, cardiovascular disorders, and neurodegenerative conditions [4]. However, realizing this potential presents substantial computational challenges that lie at the intersection of Data Warehousing and Advanced Data Mining. Multi-omics data is characterized by "high dimensionality and low sample size" (the $p >> n$ problem), whereas EHR data is often sparse, noisy, and unstructured [5]. Traditional statistical methods often fail to capture the non-linear interactions between these heterogeneous data types. Consequently, researchers are increasingly turning to Deep Learning (DL)—a subset of machine learning capable of automatic feature extraction and modeling complex dependencies—to bridge this gap [6].

The role of Data Warehousing in this context is evolving. It is no longer sufficient to merely store structured clinical data; modern warehouses must now support the ingestion, cleaning, and harmonization of unstructured biological sequences and clinical notes [7]. This requires a move towards "Data Lakes" or hybrid architectures that can support the heavy computational loads of deep learning training cycles [8]. Similarly, Data Mining techniques have advanced from simple clustering to sophisticated Representation Learning, where the goal is to map diverse data modalities into a shared latent space where they can be effectively fused [9].

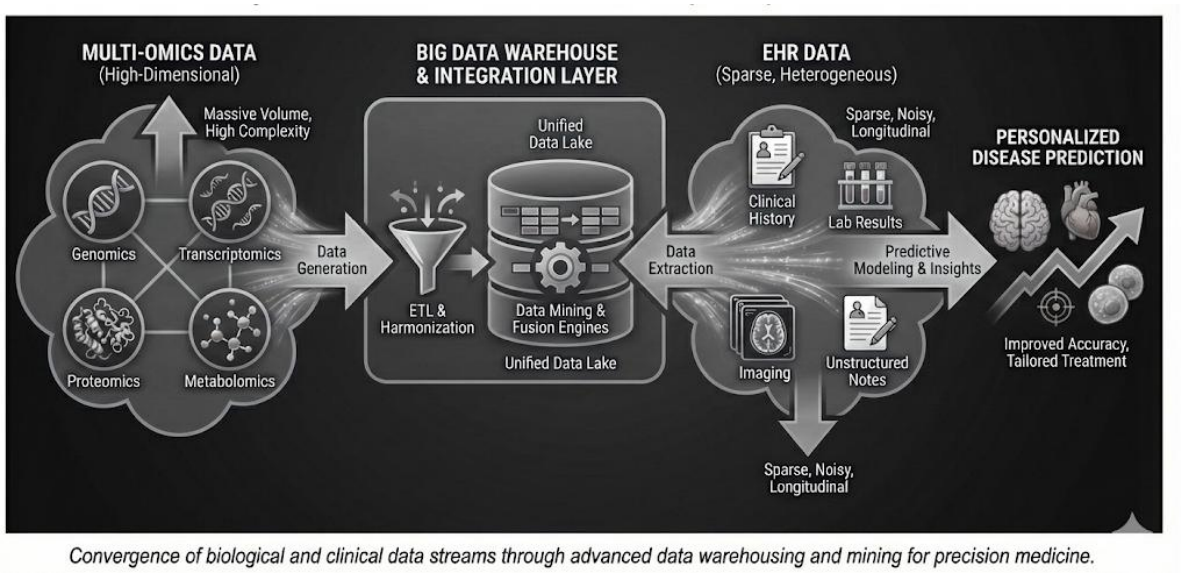

Figure 1: The Convergence of Biological Big Data: Integrating High-Dimensional Multi-Omics with Sparse Electronic Health Records (EHRs).

As illustrated in **Figure 1**, the convergence of these domains requires a systematic approach to data handling. The figure depicts the flow from raw data generation (sequencers and hospital databases) to the "Big Data Warehouse," and finally to the predictive modeling layer.

This review paper aims to provide a comprehensive analysis of the methodologies used to integrate multi-omics and EHR data. We specifically focus on the "Intermediate Integration" strategies enabled by deep learning, such as Graph Convolutional Networks (GCNs) and Variational Autoencoders (VAEs), which have shown superior performance in handling missing data and batch effects [10]. By critically evaluating the current literature, we identify the limitations of existing frameworks particularly in terms of interpretability and scalability—and propose a roadmap for future research that combines robust data warehousing infrastructure with state-of-the-art predictive modelling. The primary contributions of this review are threefold. First, we provide a taxonomy of data fusion strategies specifically tailored for high-dimensional omics and sparse EHR data. Second, we evaluate the performance of recent deep learning architectures in this domain, highlighting the shift towards transformer-based models. Third, we explicitly address the "Data Warehousing" aspect, proposing a pipeline that integrates Extract, Transform, Load (ETL) processes with real-time inference capabilities. This holistic view is essential for researchers aiming to translate algorithmic success into clinical utility.

2. RELATED WORK AND LITERATURE SURVEY

The integration of multi-omics and clinical data has evolved significantly over the last decade. Early approaches primarily utilized "Late Integration," where separate models were trained for each modality, and their predictions were averaged [11]. While simple, this method failed to capture the intricate cross-modal interactions between genes and clinical phenotypes. "Early Integration," involving the concatenation of features before training, often suffered from the curse of dimensionality, where the massive number of omics features overshadowed the clinical variables [12]. Recent literature (2020–2025) demonstrates a strong pivot towards "Intermediate Integration" using Deep Learning as shown in Figure 2. For instance, Tong et al. (2024) [2] highlighted the use of joint latent space learning, where disparate data types are projected into a lower-dimensional space that preserves the structure of both modalities. Similarly, Ballard et al. (2024) [3] provided a taxonomy of Generative vs. Non-generative models, emphasizing the utility of VAEs in imputing missing omics values using EHR data as a guide.

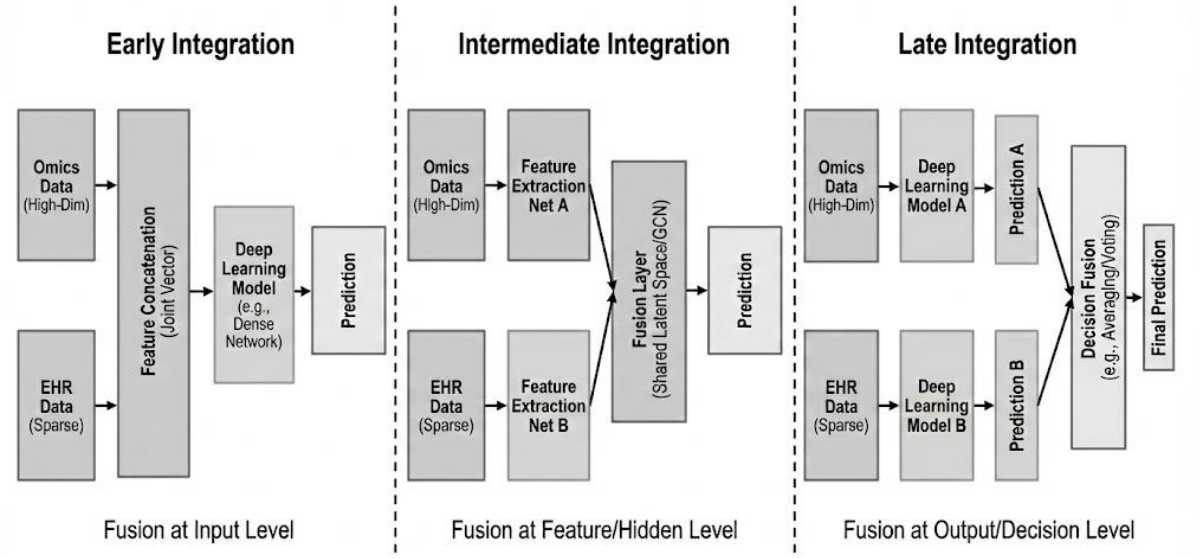

Figure 2: Taxonomy of Multi-Modal Data Fusion Strategies: Early, Intermediate, and Late Integration Architectures.

A persistent theme in the literature is the challenge of Data Heterogeneity. EHR data is often multimodal itself, containing structured tables and unstructured text. Teoh et al. (2024) [1] reviewed 69 studies and found that while integrating medical images (radiomics) with EHR is becoming common, the fusion of *molecular* omics with these clinical modalities remains under-explored due to the lack of unified data standards.

Table 1: Literature Survey of Recent Multi-Modal Integration Frameworks (2020–2025)

Author & Year	Methodology / Model	Data Modalities Integrated	Key Contribution	Limitations / Gaps
Teoh et al. (2024) [1]	Review of Fusion Techniques	Medical Images + EHR	Taxonomy of fusion techniques; highlighted superiority of multimodal over unimodal models.	Limited focus on molecular multi-omics integration.
Tong et al. (2024) [2]	Deep Learning Review	Multi-omics + EHR	Detailed "Intermediate Integration" using Deep Learning for non-linear interactions.	Identified gaps in privacy preservation across institutions.
Ballard et al. (2024) [3]	VAE, GAN, GCN	Multi-omics + Imaging	Categorized models into Generative vs. Non-generative; addressed "missing data" imputation.	High computational cost of generative models.
Baião et al. (2025) [4]	Deep Generative Models	Multi-omics (Mosaic)	Focused on "Mosaic Integration" for incomplete datasets and disentanglement learning.	Less emphasis on the clinical EHR component.

Cui et al. (2023) [5]	scGPT (Transformer)	Single-cell Multi-omics	Introduced "Foundation Models" (Transformers) pre-trained on massive biological datasets.	"Black box" nature reduces clinical interpretability.
Ahmed et al. (2022) [6]	GAN	Multi-omics	Used GANs to handle missing data and high-dimensionality problems.	Training instability of GANs; validation difficulty.
Wang et al. (2021) [7]	MOGONET (GCN)	Multi-omics	Utilized GCN to model patient similarity networks for label classification.	Poor scalability to large EHR biobanks due to graph size.
Zhang et al. (2023) [9]	Multimodal DNN	MRI + EHR + Notes	Fused unstructured text with images for Multiple Sclerosis prediction.	Limited generalizability to other disease domains.
Lipkova et al. (2022) [10]	AI for Oncology	Histopathology + Genomics	Fusing tissue images with genomic profiles improves cancer prognosis.	Lack of standardization in image acquisition.
Lin et al. (2020) [14]	Deep Learning	Multi-omics	Revealed two prognostic subtypes in high-risk neuroblastoma via integration.	Relied on small sample sizes, risk of overfitting.

3. METHODS INCORPORATED

The reviewed literature reveals a convergence towards three primary methodological pillars in the field of integrative data mining: Data Warehousing Architectures, Deep Learning Models, and Fusion Strategies.

1. Data Warehousing and Pre-processing:

Effective mining begins with robust warehousing. The integration of omics and EHR requires a Hybrid Data Warehouse capable of storing structured relational data (patient demographics, lab codes) alongside semi-structured data (XML/JSON for clinical notes) and unstructured large binary objects (BAM/VCF files for sequencing) [15], [16].

- ETL (Extract, Transform, Load): Current methodologies emphasize automated ETL pipelines that normalize clinical concepts using ontologies like SNOMED-CT or ICD-10 and harmonize omics data to correct for batch effects [17].
- Imputation: To handle the pervasive issue of missing data in EHRs, advanced imputation techniques using Generative Adversarial Networks (GANs) and Autoencoders (AEs) are employed to generate synthetic realistic values based on the learned distribution of the complete data [18].

2. Deep Learning Architectures:

- Autoencoders (AE & VAE): These are the workhorses of dimensionality reduction in this domain. VAEs are particularly favored for their ability to learn a compressed, shared latent representation of heterogeneous input features (e.g., gene expression counts and clinical variables). By optimizing the "Evidence Lower Bound" (ELBO), VAEs force disparate data distributions to align in a common feature space [19].
- Graph Convolutional Networks (GCNs): GCNs are used to model the relationships between patients or biological entities. For instance, constructing a "Patient Similarity Network" where nodes are patients and edges represent phenotypic or genotypic similarity allows the model to learn from the neighborhood of similar cases, improving prediction for rare diseases [20], [21].

Transformers: Emerging methodologies utilize Transformer architectures (like BERT or GPT) adapted for biological sequences. These models use "Self-Attention" mechanisms to weigh the importance of different omics features dynamically, capturing long-range dependencies in the genome that simpler models miss [22].

Figure 3 pinpoints the Schematic Architecture of a Multi-Modal Variational Autoencoder (VAE) for Dimensionality Reduction and Feature Learning.

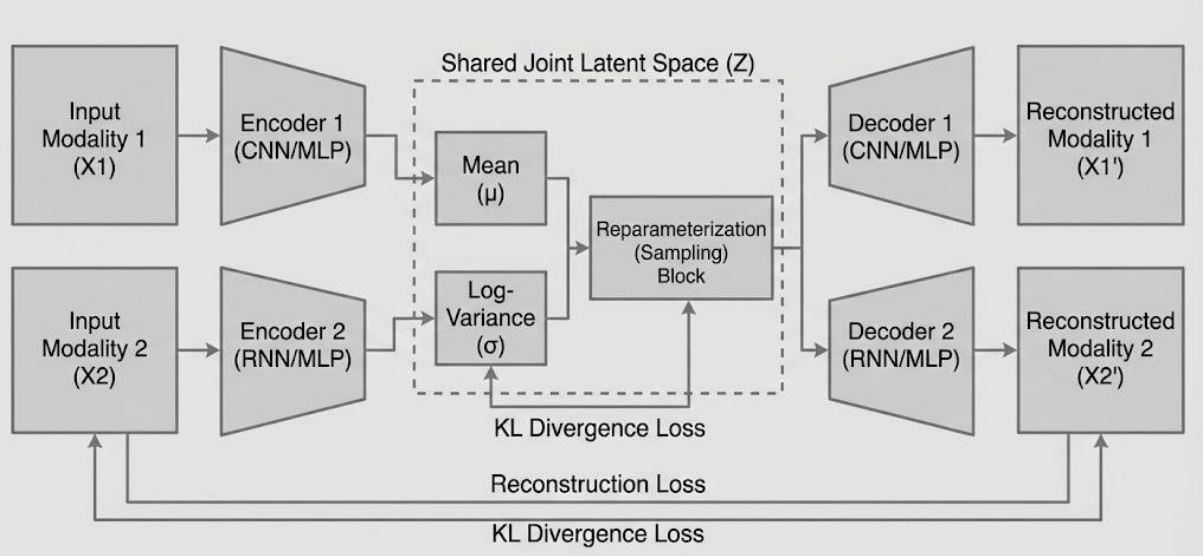

Figure 3: Schematic Architecture of a Multi-Modal Variational Autoencoder (VAE) for Dimensionality Reduction and Feature Learning.

3. Fusion Strategies:

- Intermediate Fusion: This is currently the most promising method. Unlike Early Fusion (concatenation) which ignores modality-specific structures, or Late Fusion (averaging) which ignores cross-modal interactions, Intermediate Fusion uses separate neural network branches to extract features from each modality before fusing them in a shared hidden layer. This allows the model to learn non-linear interactions, such as how a specific gene mutation influences a clinical phenotype [23], [24].

4. DISCUSSION

ADVANTAGES & PROBLEMS

Advantages: The primary advantage of integrating multi-omics with EHR data is the holistic view it provides. Models that fuse these data types consistently outperform single-modality models in metrics like AUC and sensitivity [25]. Deep learning models, particularly those using intermediate fusion, excel at capturing complex, non-linear biological relationships that traditional regression models miss. Furthermore, generative models (VAEs/GANs) offer a novel solution to the problem of missing data, allowing for the utilization of incomplete records that would otherwise be discarded [26].

Problems: A significant problem is the "Black Box" nature of deep learning. While accurate, these models often lack interpretability, making it difficult for clinicians to trust their predictions [27]. Additionally, the Curse of Dimensionality remains a hurdle; the vast feature space of omics data can lead to overfitting, especially when paired with the relatively small sample sizes of clinical cohorts [28].

CHALLENGES

- Data Heterogeneity: EHR data is sparse and noisy, while omics data is dense and high-dimensional. Bridging this gap without losing information requires complex normalization and transformation techniques [29].
- Standardization: There is a lack of universal standards for storing and sharing multi-omics data alongside EHRs. While standards like FHIR exist for clinical data, they are not fully adapted for high-throughput sequencing data [30].
- Privacy: Integrating genomic data (which is inherently identifiable) with detailed clinical history raises significant privacy concerns, complicating cross-institutional data sharing and federated learning [31].

LIMITATIONS

Current studies often rely on well-curated datasets like TCGA or ADNI, which may not reflect the messiness of real-world clinical data. Most existing models are designed for specific diseases (e.g., cancer subtyping) and lack generalizability to other conditions [32]. Furthermore, many approaches do not effectively handle longitudinal data, treating patient history as a static snapshot rather than a dynamic trajectory [33]. The computational cost of training these models also limits their deployment in resource-constrained hospital environments.

RESEARCH GAP AND FUTURE DIRECTIONS

Research Gap: There is a distinct lack of frameworks that effectively combine "Data Warehousing" principles with "Deep Learning". Most research focuses purely on the algorithm, neglecting the data infrastructure required to deploy these models in a real hospital setting. Specifically, there is a need for "Lakehouse" architectures that support the real-time querying and mining of integrated omics-EHR data [34].

Future Directions: Future work must focus on Explainable AI (XAI) to make deep learning models transparent to physicians [35]. Additionally, the development of Federated Learning frameworks will be crucial to allow models to learn from multi-institutional data without compromising patient privacy [36].

RECOMMENDATIONS

Based on this review, the following research phases are recommended for the study as shown in figure 4.

Phase 1: Development of a Hybrid Data Warehouse Schema: Design a unified schema (potentially extending the OMOP CDM) that can efficiently store and link structured EHR data with high-dimensional omics files. This should include an automated ETL pipeline for cleaning and normalizing data from disparate sources [37].

Phase 2: Construction of a Multi-Modal Deep Learning Framework: Develop a Hybrid VAE-GCN Model. Use Variational Autoencoders (VAEs) to reduce the dimensionality of the omics data and impute missing values. Then, feed these latent representations, along with the EHR features, into a Graph Convolutional Network (GCN) to model patient similarities and predict disease outcomes. This approach directly addresses the dimensionality and sparsity challenges identified in the review [38], [39].

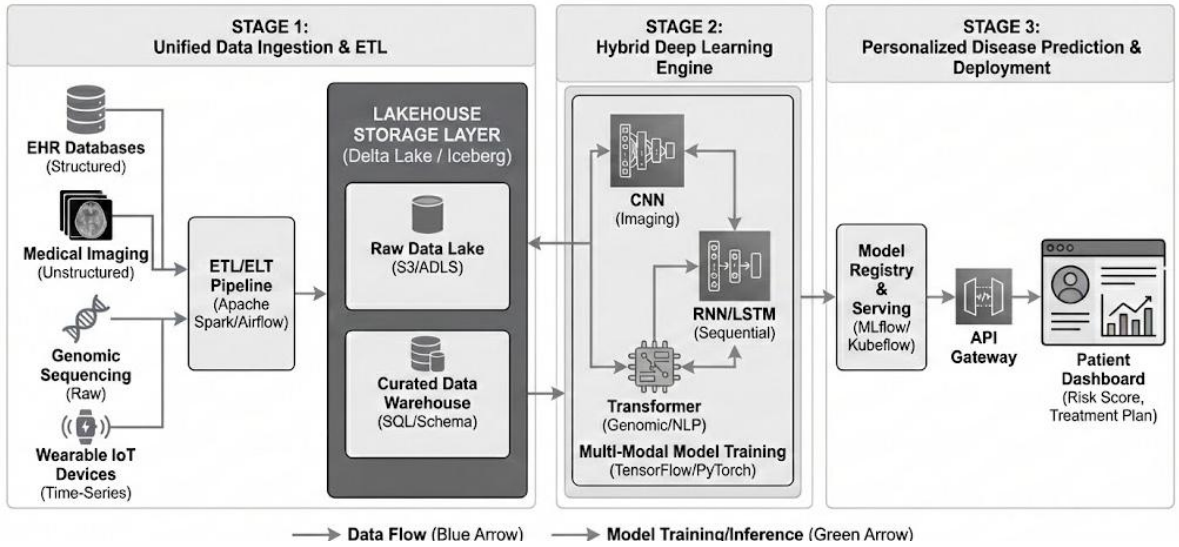

Figure 4: Proposed "Lakehouse" Framework: A Unified Pipeline for ETL, Hybrid Deep Learning, and Personalized Disease Prediction.

5. CONCLUSION

The integration of Multi-Omics and EHR data represents the frontier of personalized medicine. This review has established that while significant progress has been made using Deep Learning, critical challenges remain in data warehousing, dimensionality reduction, and model interpretability. The "siloed" nature of biological and clinical data continues to hinder the full realization of precision health. By adopting a "Data Warehousing and Mining" perspective, this research proposes to move beyond simple data combination to true Data Fusion. The recommended approach

leveraging hybrid deep learning architectures within a robust data infrastructure holds the potential to unlock new biomarkers, improve risk stratification, and ultimately, deliver on the promise of accurate, personalized disease prediction [40],[50].

References

1. Teoh, J. R., Dong, J., Zuo, X., Lai, K. W., Hasikin, K., & Wu, X. (2024). Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications. PeerJ Computer Science, 10, e2298.
2. Tong, L., Shi, W., Isgut, M., Zhong, Y., Lais, P., Gloster, L & Wang, M. D. (2024). Integrating multi-omics data with EHR for precision medicine using advanced artificial intelligence. IEEE Reviews in Biomedical Engineering, 17, 80-97.
3. Ballard, J. L., Wang, Z., Li, W., Shen, L., & Long, Q. (2024). Deep learning-based approaches for multi-omics data integration and analysis. BioData Mining, 17(1), 38.
4. Baião, A. R., Cai, Z., Poulos, R. C., Robinson, P. J., Reddel, R. R., Zhong, Q., ... & Gonçalves, E. (2025). A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches. arXiv preprint arXiv:2501.17729.
5. Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., & Wang, B. (2023). scGPT: Towards building a foundation model for single-cell multi-omics using generative AI. Nature Methods, 1-11.
6. Ahmed, K. T., Sun, J., Cheng, S., Yong, J., & Zhang, W. (2022). Multi-omics data integration by generative adversarial network. Bioinformatics, 38(1), 179-186.
7. Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z & Huang, K. (2021). MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nature Communications, 12(1), 3445.
8. Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digital Medicine, 3(1), 136.
9. Zhang, K., Lincoln, J. A., Jiang, X. Q., Bernstam, E. V., & Shams, S. (2023). Predicting multiple sclerosis severity with multimodal deep neural networks. BMC Medical Informatics and Decision Making, 23(1), 255.
10. Lipkova, J., Chen, R. J., Chen, B., Lu, M. Y., Barbieri, M., Shao, D & Mahmood, F. (2022). Artificial intelligence for multimodal data integration in oncology. Cancer Cell, 40(10), 1095-1110.
11. Akazawa, M., & Hashimoto, K. (2023). A multimodal deep learning model for predicting severe hemorrhage in placenta previa. Scientific Reports, 13, 17320.
12. Xu, M., Ouyang, L., Han, L., Sun, K., Yu, T. T., Li, Q & Chen, S. (2021). Accurately differentiating between patients with COVID-19, patients with other viral infections, and healthy individuals: multimodal late fusion learning approach. Journal of Medical Internet Research, 23(1), e25535.
13. Bhagwat, N., Viviano, J. D., Voineskos, A. N., & Chakravarty, M. M. (2018). Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data. PLOS Computational Biology, 14(9), e1006376.
14. Lin, Y. W., Chen, W. C., Tsai, C. F., & Chen, H. Y. (2020). Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Frontiers in Genetics, 11, 477.
15. Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. Genome Biology, 18, 83.
16. Johnson, K. B., Wei, W. Q., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K & Snowdon, J. L. (2016). Precision medicine, AI, and the future of personalized health care. Clinical and Translational Science, 14(1), 86-93.
17. Richesson, R. L., Hammond, W. E., Nahm, M., Wixted, D., Simon, G. E., Robinson & Smerek, M. M. (2012). Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. Journal of the American Medical Informatics Association, 20(e2), e226-e231.
18. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236-1246.
19. Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. Bioinformatics and Biology Insights, 14, 1177932219899051.
20. Rappoport, N., & Shamir, R. (2018). NEMO: cancer subtyping by integration of partial multi-omic data. Bioinformatics, 35(18), 3348-3356.
21. Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C., & Ester, M. (2019). MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics, 35(14), i501-i509.
22. Chai, H., Zhou, X., Zhang, Z., Rao, J., & Zhao, H. (2021). Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. Computers in Biology and Medicine, 134, 104481.
23. Gevaert, O., Xu, J., Hoang, C. D., Leung, A. N., Xu, Y., Quon, A & Plevritis, S. K. (2012). Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data methods and preliminary results. Radiology, 264(2), 387-396.
24. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.
25. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1(1), 18.

26. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589-1604.
27. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29.
28. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317-1318.
29. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.
30. Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P & Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. Journal of the Royal Society Interface, 15(141), 20170387.
31. Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321-332.
32. Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. Briefings in Bioinformatics, 18(5), 851-869.
33. Yue, T., & Wang, H. (2018). Deep learning for genomics: A concise overview. arXiv preprint arXiv:1802.00810.
34. Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. Nature Reviews Genetics, 20(7), 389-403.
35. Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. Molecular Systems Biology, 12(7), 878.
36. Wainberg, M., Merico, D., Delong, A., & Frey, B. J. (2018). Deep learning in biomedicine. Nature Biotechnology, 36(9), 829-838.
37. Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association, 25(10), 1419-1428.
38. Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 13(6), 395-405.
39. Hripcsak, G., & Albers, D. J. (2013). Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association, 20(1), 117-121.
40. Wei, W. Q., & Denny, J. C. (2015). Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Medicine, 7(1), 41.
41. Kohane, I. S. (2011). Using electronic health records to drive discovery in disease genomics. Nature Reviews Genetics, 12(6), 417-428.
42. Denny, J. C. (2012). Chapter 13: mining electronic health records in the genomics era. PLoS Computational Biology, 8(12), e1002823.
43. Pathak, J., Kho, A. N., & Denny, J. C. (2013). Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. Journal of the American Medical Informatics Association, 20(e2), e206-e211.
44. Liao, K. P., Cai, T., Savova, G. K., Murphy, S. N., Karlson, E. W., Ananthakrishnan, A. N., ... & Kohane, I. S. (2015). Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ, 350.
45. Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association, 17(5), 507-513.
46. Wu, P. Y., Cheng, C. W., Kaddi, C. D., Venugopalan, J., Hoffman, R., & Wang, M. D. (2017). Omic and electronic health record big data analytics for precision medicine. IEEE Transactions on Biomedical Engineering, 64(2), 263-273.
47. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., & Stegle, O. (2018). Multi-Omics Factor Analysis a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology, 14(6), e8124.
48. Rohart, F., Gautier, B., Singh, A., & Lê Cao, K. A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Computational Biology, 13(11), e1005752.
49. Mariette, J., & Villa-Vialaneix, N. (2018). Unsupervised multiple kernel learning for heterogeneous data integration. Bioinformatics, 34(6), 1009-1015.
50. Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500), 2319-2323.