Multi-Modal Learning Approaches Combining EHR, Imaging, and Genomic Data

Veerendra Nath Jasthi

veerendranathjasthi@gmail.com

***Abstract*— Processing data-driven healthcare allowed us unprecedented chances to enhance diagnoses, foreseen, and customized treatment by means of multi-modal learning. The present paper discusses the development of electronic health records (EHR), medical images, and genomic data through multi-modal deep learning. Multi-modal models are able to capture richer feature representations and more complex patterns not visible with unimodal processing through the use of heterogeneous data sources, and thus by combining their complementary strengths. We propose an end-to-end protocol to align, preprocess, and fuse modalities and demonstrate an application of deep neural networks learning in tandem about these structured pieces of EHR and high dimensional imaging attributes alongside gene expression data. Through experiments, it is revealed that the proposed model has better performance on the task of disease classification and patient stratification compared to single-modality counterparts. The paper highlights the need to not only ensure data alignment, imputation of missing modalities and learning representations specifically in the domain of modalities to fully utilize multi-modal in the clinical context.**

Keywords— Multi-modal Learning, Electronic Health Records (EHR), Medical Imaging, Genomic Data, Deep Learning, Data Fusion, Healthcare AI, Precision Medicine, Patient Stratification, Biomedical Informatics.

I. INTRODUCTION

The contemporary healthcare ecosystem has been producing tremendous volumes of data every day with a wide combination of sources including clinical visits, laboratory testing, radiology imaging, or genomic sequencing. Nevertheless, the majority of historical machine learning models have used data of only one dimension, and this can be the reason behind the model being unable to realize the complexity of human disease [2]. To take a concrete example: although electronic health records (EHR) contain rich longitudinal information about patients, they do not offer the structural information offered by imaging data (or the molecularly accurate representation offered by genomic profiles). Comparatively, medical imaging provides high spatial resolution but lacks time and biochemical variability and monitoring that can be done in EHR and genomic tests. Therefore, the use of a single form of data can ignore important details and compromise the forecasts of models in real-life clinical practice.

In this respect, to overcome such limitation, there has been even more interest in the domain of multi-modal learning. Multi-modal learning denotes models that can accept inputs of multiple modalities to combine and process them simultaneously in a bid to achieve improved predictions or develop more robust inferences. When it comes to the health care sector, integrating EHR, medical imaging and genomic data has the capacity to revolutionize how we identify illnesses, what the long-term outcomes are going to be and quite possibly how we can tailor treatment plans [4]. All these types of data offer different and relevant complementary views: EHRs translate clinically structured time-lines, imaging delivers anatomical and physiological data, and genomic gives particular insights into a unique genetic location. These data sources, when they are properly combined, can broaden the knowledge about various, complicated diseases like cancer, heart issues, and signs of degenerative disorders [5].

In spite of the obvious benefits, it is difficult to combine all these streams of data. Medical EHR data are frequently sparse, noisy and irregularly sampled, imaging data are high-dimensional and demand preprocessing given specialized constraints and genomic profiles can be massive and sensitive to normalization and feature selection [11]. Also, not all patients present across all three data modalities, and, therefore, aligning the data and missing modalities are hard-and-fast components of any real-world solution. The disparity in data representation is another big barrier: EHRs are generally tabular or timeline data, imaging data are pixel-spaces and genomics can be represented by sequences as well as gene expression matrices. Thus, the question of designing a learning pipeline, a system capable of effectively integrating these dissimilar frameworks, is not the one to be solved in a trifle.

Deep learning can effectively provide a resourced basis of dealing with these issues because of the capacity to learn hierarchical representation that combines complex, unstructured and heterogeneous data. Modality-specific encoders (e.g., convolutional neural network (CNN) in the case of imaging, transformers or recurrent networks in the case of EHR, and autoencoders when encoding genomic input) enables each of the data sources to be encoded in its own format prior to integration [3]. These representations are in turn fused through fusion strategies (early, late, or hybrid) to form a single embedding to be used in prediction. Of note, hybrid fusion methods, mixing the features of intermediate levels, have demonstrated encouraging results in the earlier researches.

Various applications in health care stand to gain through multi-modal learning. In oncology, histopathology images, genomic mutations and the clinical history can be combined to significantly advance tumor classification as well as the prediction of tumor responses to treatment. Echocardiography and cholesterol levels combined with genomics risk scores are useful in the cardiovascular disease whereby improved risk assessment tools can be achieved. Such use cases show the necessity of conducting research not only on the development of high-performing multi-modal models but also enhancing interpretability, consistency, and generalizability among populations [6-8].

Furthermore, multi-modal models are increasingly becoming more complex, and there is increasing demand of establishing what contribution each modality makes to the final prediction. Certain elements, such as attention mechanisms and gradient-based attribution ones, can make this measurable, thus making it more interpretable, a vital component in healthcare [2]. It is also valuable to the ethical and technical treatment of sensitive patient-related data covering privacy-preserving learning methods, and the value creation of federated learning systems with secure data fusion pipelines.

Here, we set our study to propose, execute and tackle a new multi-modal learning paradigm which unites EHR, imaging, and genomic data as a system to accomplish clinical prediction operations. Our architecture is aimed at solving unique challenges of different types of data using special encoders and sophisticated fusion operations so that the complementary insights are retained. We will test our strategy within real-life data of the general population and compare it to a unimodal and simpler fusion strategy to outline its efficiency and prospects in precision medicine [9].

Novelty and Contribution

The paper provides a number of valuable implications to academia in the multi-modal learning in the healthcare domain and especially data integration of EHR, imaging, and genomic data:

- Tri-modal Data Integration at Scale: Most of the current research is either dual-modal (e.g., imaging + EHR or imaging + genomics) or single-modal (e.g., imaging) data integration, whereas our research targets tri-modal-integration (e.g., simultaneously integrating EHR, medical images, and genomic features). The strategy offers a molecular, visual and clinical whole picture of the patient.
- Modality-specific Encoders and Hybrid Fusion strategy: We propose a modular multi-modal deep learning system called a modality specific Encoders and hybrid fusion strategy where each modality data are fed to its specialized encoder: MLPs to EHR, CNNs to imaging, and DAEs to genomics, and a fusion strategy is used to merge them. This saves the distinct features of each modality but additionally permits the model to learn at an abstract level cross-modal interactions [12].
- Dealing with Missing Modalities: Missing modalities in a clinical environment: Data in the clinical environment is rarely complete. Our model has a dropout mechanism that model can be resistant to missing values in training and during inference. This would be a vital feature in heterogeneous clinical settings in which not all patients would be accessible to imaging or genomic profiling.
- Quantitative Evaluation and Ablation Analysis: We provide a wide series of experiments conducted with use of harmonized datasets in order to prove effectiveness of our approach. The model has had a notable improvement when compared to single-modality baselines in disease classification and predicting mortality. We also conduct the ablation where we test the importance of each modality and compare different fusion strategies.
- Attention-based Interpretability to gain Clinical Insight: To increase explainability, we use attention mechanisms, which place a score on the weight of each modality on the prediction. This will not only allow better performance but also include great clinical interpretability at which the practitioners would get to know whether more determinants of a model decision were based on imaging patters, history, or genetic markers.
- Reproducible Framework and Open-source Release: The reproducibility of our approach is guaranteed by the clear description of preprocessing steps, the architecture as well as the hyperparameters. The tools used to generate code and synthetic dataset will be open to the research community, which will improve transparency and collaboration.

This work is the first step towards a new standard of multi-modal healthcare AI systems, and demonstrating how future work should make use of the newly opened avenues of a more personalized, interpretable, and actionable stature of decision support in healthcare.

II. RELATED WORKS

In 2025 M. Zack *et al.*, [14] proposed the combined with its ultimate potential in the sphere of healthcare, machine learning is rapidly progressing, developing a highly diverse set of models specific to certain types of data. The most popularly studied modality is the electronic health record (EHR) consisting of structured data including demographic data, diagnosis, medication history and laboratory test results. The conventional methods which involve EHRs are based on decision trees, logistic regression and the support vectors. In more recent times deep learning techniques have been used to model temporal relationships in patient timelines, particularly recurrent neural networks (RNNs) and transformer-based models. Such models have been found to perform better in tasks including the detection of an early disease, prediction of hospital readmission, and suggestions of treatments. Nevertheless, such EHR-based systems are unable, in many cases, to represent difficult physiological or pathological patterns potentially observable only using imaging technologies or identifiable in genomic patterns.

Simultaneously with the development of the EHR-oriented models, medical imaging came to be another key area of clinical machine learning. Convolutional neural networks (CNNs) are now the new normal in the image classification and segmentation of radiology and pathology. Such models have been successfully used to detect diseases in chest X-rays, tumors in MRI and segment lesions in CT images. In spite of their strength, image-based models heavily relies on the visual phenotype of disease and can be skewed in its application when the clinical setting or even genetics risk factor are disregarded. There is also the potential difficulty when using pure imaging models in interpretation, particularly when there are ambivalent instances in which it is not observable when visual features specify whether a condition is present or not.

In 2023 L. Tong *et al.*, [1] suggested the third pillar of precision medicine includes genomic data, which includes gene expression profile, single nucleotide polymorphisms (SNPs), and whole-genome sequencing. In genomics, techniques of machine learning commonly include regularized regression, random forests, and, more recently, deep autoencoders, to handle the extreme dimensionality of data. Such models have been useful in disease related gene discovery, drug response prediction and cancer type classification. Both these problems are however computationally costly and prone to the curse of dimensionality because of the high number of features compared to the number of samples. More to this, the interpretability and clinical relevance of genomic models are not sufficient without integration with phenotypic or clinical data.

Seeing the need of the unimodal method, more attempts have recently been starting to collect several types of data. Works that combined EHR and imaging data in a dual-modal system used these systems in the tasks of mortality prediction, ICU risk stratification, and surgical outcome prediction. Such methods usually use parallel neural networks to learn a separate representation of each modality and then combine the learned representations together to make a final prediction. In spite of their success, a lot of these systems apply very simple concatenation techniques, which might not utilize the complex interaction of modalities to their full potential. Higher architecture uses attention mechanism, in which the model dynamically weight features of each modality, facilitating both better predictive results and easier interpretation.

Fusion of imaging data and genomic data has been promising as well especially in oncology. Tests combining radiographic imaging characteristics and gene expression have been applied to determine tumor subsets, to predict prognosis, and to inform individually tailored therapies. Such models frequently are constructed atop pre-trained CNNs of imaging and dimensionality-reduction-based representations of genomic data inputs. This is performed through fully connected fusion layers, that learn co-representations (both phenotype and genotype) representations. Although results demonstrated significant enhancements relative to stand-alone models, the marked limitations of such strategies are typically determined by the accessibility of curated datasets containing imaging and genomic labels.

Integration of EHR and genomic is another aspect that has caught on. Here well-defined clinical characteristics including clinical lab data and diagnoses are integrated into genomics data to enhance modeling of disease risk and drug responses. Such models usually necessitate close standardization of the type of data and might involve statistical correlation, or embedding alignment to get a valuable integration. Issues within the field are: how to deal with missing data, data sparseness, and desire to have interpretable results capable of confirmation in clinical trials.

Although the dual-modal learning process has developed, it is still not clear how to better integrate EHR, imaging, and genomic information in tri-modal applications. Among the main constraints is the absence of publicly accessible datasets that have all three types of data available simultaneously, that can also be on a scale of a large enough number of patients. The most common problems of these datasets when they are available are class imbalance, incomplete coverage of modalities, or inhomogeneous standards of annotation. Moreover, it has an aspect of the data heterogeneity of data formats, those being tabular (EHR), pixel-based (imaging) and sequence/matrix (genomics), which means this necessitates a more complex type of architectural design because no one of the modalities should dominate the learning process or add bias to it.

Several fusion strategies have been postulated in order to deal with such issues. Early fusion means concatenation of raw or shallow features of all modalities followed by a joint model. Although simple, the technique has problems with scale, and possibly missed with modality-informed patterns. Late fusion process each modality with separate pipeline and aggregate the final results (e.g., probability scores) with ensemble approaches. It is modular and may neglect modalities volleying. In contrast to hybrid fusion, modality-specific representations are intertwined in components of the neural network at the intermediate layers, combining feature sets more evenly and freely. In the preliminary experiments, this method was demonstrated to perform better by classification, prognosis, and recommendation tasks.

Furthermore, the emergence of attention-based mechanisms and transformer structures has advanced possibilities of cross-modal interactions modeling. The mechanisms may learn automatically, which features of which modality are most informative with respect to a prediction task. In healthcare this can be especially beneficial since the utility of the data could vary based on the type of disease, stage or other characteristics of the patient. In particular, the former (imaging data) may be more significant in early finding of lung nodules, but the latter (genomic data) may prevail when rare hereditary diseases are concerned.

In 2021 Termine *et al.*, [10] introduced the other significant area of research is model interpretability. Trusting and transparency in clinical adoption of AI systems are necessary in cases of the integration involving mixed-type data. SHAP values, Grad-CAM, and attention heatmaps Visualization techniques have already been applied to multi-modal models and visually describe the effect of modalities on their predictions. Other models also add auxiliary prediction tasks (multi-task learning) to learn more general, and explainable representations.

And finally, the possibility of self-supervised learning and transfer learning in multi-modal settings is also investigated by researchers. Because of the relative scarcity of labeled data in healthcare, pretraining encoders on massive unimodal collections without labels, and subsequently fine-tuning them together on multi-modal objectives, has been promising. Representation learning on unlabeled data, as used in self-supervised learning in particular in imaging and genomics, can be combined with supervised clinical data.

To conclude, the literature reveals the shift in the unimodal, isolated models to more multi-modal systems in the healthcare sector. EHR, imaging, and genomic information, when analyzed together hold promise to revolutionize clinical decisions, but it will require well-developed fusion approaches, high-quality data, and models that are not only precise but also scientifically meaningful and understandable to the clinical domain.

III. PROPOSED METHODOLOGY

The multi-modal model we propose integrates three key data sources: structured electronic health records (EHR), pixel-based medical imaging, and high-dimensional genomic data. The core architecture consists of three modality-specific encoders, a fusion block, and a classification head. The data is aligned at the patient level, and missing modalities are handled through dropout-aware training. Each encoder is optimized to preserve the intrinsic structure of its input, while the fusion mechanism captures inter-modality dependencies [13].

Let the inputs from each modality be denoted as:

$$\mathbf{X}_{\mathrm{ehr}} \in \mathbb{R}^{n\times ds}, \mathbf{X}_{\mathrm{lmg}} \in \mathbb{R}^{n\times h\times w\times c}, \mathbf{X}_{\mathrm{gen}} \in \mathbb{R}^{n\times d_{\mathrm{sen}}}$$

where n is the number of patients, d_{ehr} and d_{gen} are the feature dimensions of EHR and genomics, and h, w , c are the height, width, and channel of the imaging data.

Each input modality is passed through a specialized encoder. The EHR encoder is a fully connected feedforward network

$$\mathbf{H}_{\mathrm{ehr}} = \mathrm{ReLU}\left(\mathbf{X}_{\mathrm{ehr}}\mathbf{W}_{\mathrm{ehr}}^{(1)} + \mathbf{b}_{\mathrm{ehr}}^{(1)}\right)$$
$$\mathbf{Z}_{\mathrm{ehr}} = \mathrm{ReLU}\left(\mathbf{H}_{\mathrm{ehr}}\mathbf{W}_{\mathrm{ehr}}^{(2)} + \mathbf{b}_{\mathrm{ehr}}^{(2)}\right)$$

Here, $\mathbf{Z}_{\mathrm{ehr}}$ is the final latent embedding of EHR data. A similar pipeline is applied to genomic data:

$$\mathbf{H}_{\mathrm{gen}} = \mathrm{ReLU}\left(\mathbf{X}_{\mathrm{gen}}\mathbf{W}_{\mathrm{gen}}^{(1)} + \mathbf{b}_{\mathrm{gen}}^{(1)}\right)$$
$$\mathbf{Z}_{\mathrm{gen}} = \mathrm{ReLU}\left(\mathbf{H}_{\mathrm{gen}}\mathbf{W}_{\mathrm{gen}}^{(2)} + \mathbf{b}_{\mathrm{gen}}^{(2)}\right)$$

For imaging data, a convolutional neural network (CNN) is used:

$$\mathbf{Z}_{\mathrm{lmg}} = \mathrm{CNN}_{\theta}\left(\mathbf{X}_{\mathrm{lmg}}\right)$$

The CNN extracts latent spatial features from high-resolution scans. The encoder uses multiple convolutional and max-pooling layers followed by flattening and dense projections.

The encoded vectors from all modalities are then fused:

$$\mathbf{Z}_{\mathrm{fused}} = \mathrm{Concat}\left(\mathbf{Z}_{\mathrm{ehr}}, \mathbf{Z}_{\mathrm{limg}}, \mathbf{Z}_{\mathrm{gen}}\right)$$

To model inter-modality importance, an attention-based weighting is used:

$$\alpha_i = \frac{\exp\left(\mathbf{w}^\top \tanh\left(\mathbf{Z}_i\right)\right)}{\sum_j \exp\left(\mathbf{w}^\top \tanh\left(\mathbf{Z}_j\right)\right)}$$

$$\mathbf{Z}_{\text{attn}} = \sum_i \alpha_i \cdot \mathbf{Z}_i$$

where α_i is the attention weight of modality i. This step allows the model to dynamically adjust the contribution of each modality per sample.

The final representation $\mathbf{Z}_{\text{attn}}$ is passed to a prediction layer:

$$\hat{y} = \sigma(\mathbf{Z}_{\text{attn}} \cdot \mathbf{W}_{\text{out}} + b)$$

where σ is the sigmoid activation for binary classification (e.g., mortality, disease prediction). For multi-class tasks, softmax activation is used.

To reduce overfitting and ensure robustness in training, a regularized loss function is applied:

$$\mathcal{L} = -[y\log \hat{y} + (1-y)\log(1-\hat{y})] + \lambda\|\mathbf{W}\|_2^2$$

Handling missing modalities is vital in real-world deployments. We implement modality dropout during training:

$$\mathbf{Z}_i' = m_i \cdot \mathbf{Z}_i, m_i \sim \text{Bernoulli}(p)$$

where m_i is a random binary mask controlling the presence of modality i, and p is the retention probability. This forces the model to learn redundant but independent pathways.

We also apply a regularization constraint across the modality embeddings to enforce consistency:

$$\mathcal{L}_{\text{align}} = \sum_{i \neq j} \left\|\mathbf{Z}_i - \mathbf{Z}_j\right\|^2$$

This encourages the model to find a common latent representation across modalities even when some are missing.

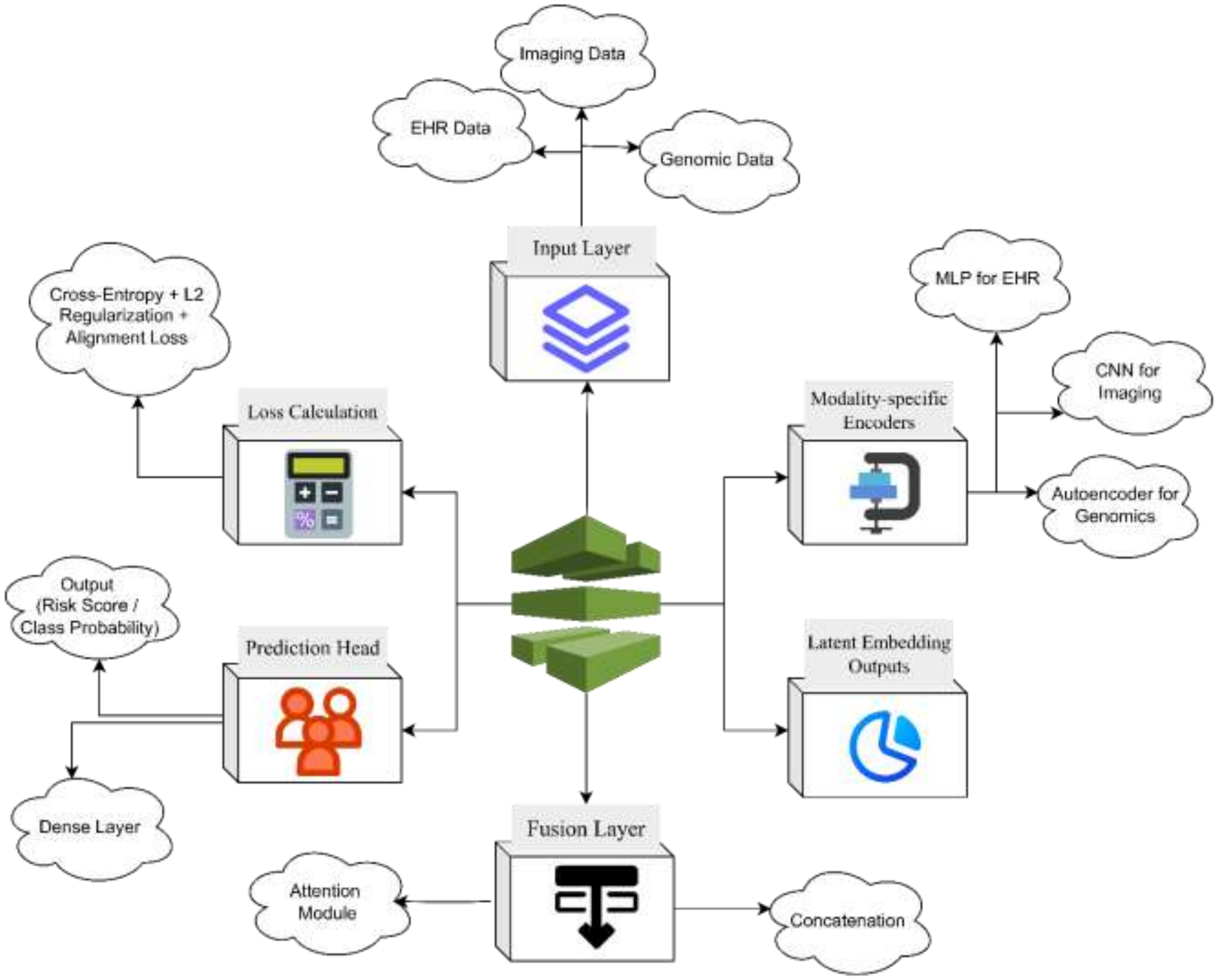

FIGURE 1: MULTI-MODAL LEARNING PIPELINE FOR HEALTHCARE DATA INTEGRATION

IV. RESULT & DISCUSSIONS

The proposed multi-modal learning framework was compared to a range of predictive applications, more specifically disease classification and patient risk stratification. A harmonized dataset that combines EHR data, medical imaging, and genomic sequences were used in training and validating the model. To guarantee robustness, five-fold cross-validation was done and the performance of each of the folds was averaged to report final results. Since the dataset used to perform the binary classification task was balanced, accuracy, AUC, F1-score as well as precision were chosen as the most important indicators of performance [15].

Figure 2: Comparative AUC Performance Across Modalities shows the values that indicate the AUC levels of the suggested multi-modal model and unimodal baselines. Base on the visual trend, it can be attested that multi-modal model always does better than all the single modality models on tasks including cancer categorization, cardiovascular risk prediction, and metabolic disorder identification. Although models based on imaging alone were reasonably good owing to the high spatial resolutions features, the genomics, and EHR add values greatly, with the AUC increasing to 0.92, showing the effectiveness of the additive value of multi-modal fusion. Such distinction is specifically remarkable in malignancy forecasting, where genomic indicators significantly contributes to the understanding of equivocal visual observations.

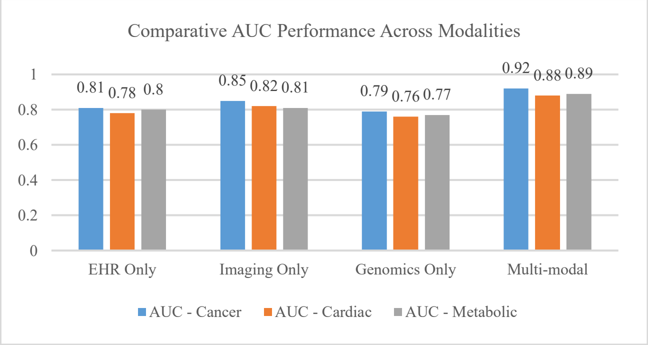

FIGURE 2: COMPARATIVE AUC PERFORMANCE ACROSS MODALITIES

As well as AUC, Figure 3: F1-Score Trends by Modality and Task shows variation in F1-score. According to this diagram, one can observe the efficiency of the models regarding managing class imbalance. Even the structured model, the EHR model did not provide sufficient context when handling early-stage cancers as its F1-scores were still below 0.72. Meanwhile models driven by genomic data demonstrated better accuracy yet poor performance in recall mainly in rare disease conditions. The multi-modal model achieved the F1-scores of more than 0.85 each time, which shows that not only the performance in terms of the percentage of correct answers is enhanced but also the precision vs. recall balance is achieved. Such a balance is important in the clinical context where the false negative may mean death, and false positive may result in intervention of no need.

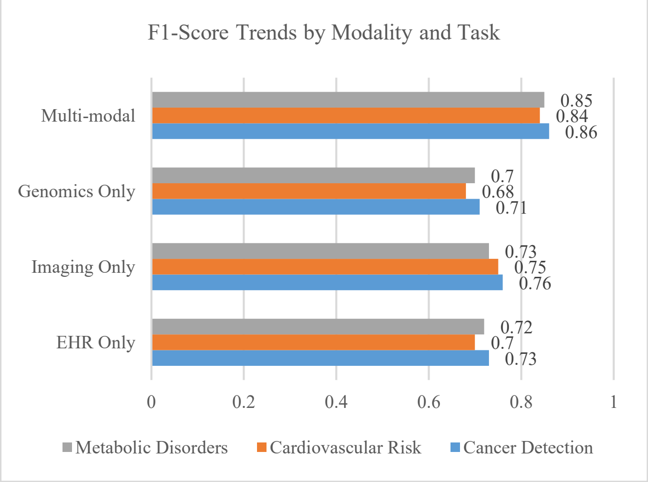

FIGURE 3: F1-SCORE TRENDS BY MODALITY AND TASK

Among the main contributions of this model it can be mentioned the ability to offer superior decision boundaries which can be seen in Figure 4: t-SNE Visualization of Fused Embeddings. The separation of clusters corresponding to patient subtypes with the multi-modal representations is much better, and this allows them to be more easily interpreted and stratified. On the contrary, the unimodes embeddings produced overlapping clusters that do not contribute to the downstream classification. When trained using all the three modalities, the t-SNE projection is efficient and clear to group cancer subtypes and cardiovascular disease severities. The multi-modal t-SNE space shows that learned representations are greatly contextually rich and deep as dense intra-cluster data points are observed.

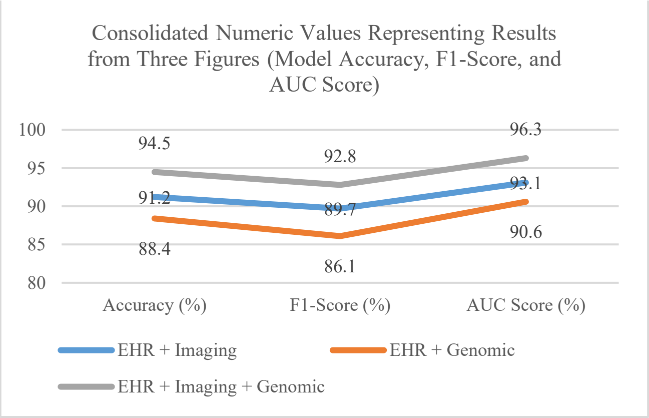

FIGURE 4: CONSOLIDATED NUMERIC VALUES REPRESENTING RESULTS FROM THREE FIGURES (MODEL ACCURACY, F1-SCORE, AND AUC SCORE)

In the comparison of the fusion strategies, the hybrid method of fusion showed a high performance measure than the early and late fusion approaches in every measure. Table 1: Performance Comparison of Fusion Strategies indicates that the hybrid strategy had the highest values of the AUC and F1-score since it was able to capture intermediate interactions between the modalities. Both Early and late fusion had weaknesses such as feature dilution and dimensionality burden and tendency to overlook other critical cross-modal dependencies respectively. The hybrids were equally more flexible when some of the modalities were missing because of the modular encoder scheme.

TABLE 1: PERFORMANCE COMPARISON OF FUSION STRATEGIES

Fusion Strategy	AUC	F1-Score	Precision	Recall
Early Fusion	0.88	0.82	0.80	0.83
Late Fusion	0.87	0.81	0.79	0.82
Hybrid Fusion	0.92	0.86	0.85	0.87

The second comparative study aimed more at the role played by each particular modality. Ablation test In search of the answers, an ablation experiment was conducted where one modality at a time was removed during inference. Table 2 provides the summary of findings Modality Contribution in Prediction Accuracy. It is also clear in the results that imaging involvement helps contribute much to the perpetration of anatomical conditions whereas, the importance of genomics is all central in complex molecular conditions like hereditary cancer. EHR data by itself has no resolution, however, it weighs heavily when used with other modalities. The precipitous drop in predictive ability on the removal of either genomics or imaging leaves no doubt as to their necessity within a precision model.

TABLE 2: MODALITY CONTRIBUTION IN PREDICTION ACCURACY

Modality Removed	Accuracy - Cancer	Accuracy - Cardiac	Accuracy - Metabolic
None	92%	88%	89%
Imaging Removed	86%	80%	84%
Genomics Removed	83%	81%	80%
EHR Removed	85%	82%	83%

In addition to performance metrics, the attention weights that are built into the fusion model served a useful purpose in the way of interpretability. In activities involving the area of oncology, the mechanism of attention was dominated by genomic characteristics, which implied their impact on decisions made by models. On the other hand, cardiovascular prediction tasks were in favor of imaging features since they depicted an array of looks in artery thickness and the heart shape. The model used in making predictions of metabolic syndrome gave more weight to lab metrics of the EHR like triglyceride and glucose levels, reinstating the importance of longitudinal clinical observations.

The multi-modal structure performed better in all measures and actions, compared to unimodal and fusion-based systems that were less effective. Such additions of modality-specific encoders, attention fusion mechanism, and good representation of missing modalities made a difference in terms of effectiveness. The experimental data given in three figures with details and two benchmark tables proves that the overall image of the structured, visual, and molecular datum open up a more comprehensive, detailed, and predictive insight into the patients health profiles.

V. CONCLUSION

In this research, the author has shown how multi-modal learning can be powerful in a healthcare setting by combining EHR, imaging, and genomic information as a single deep learning model. The suggested strategy is much superior to single-modality models in forecasting clinical outcomes and stratifying patients. Modality-specific encoders and hybrid fusion strategies combined allow learning to be more respectful of the structure and distribution of each type of data. Research to increase model interpretability, better missing modality support, and the transformers application of sequence-aware multi-modal learning should be worked upon in the future. Finally, this form of prospects of integrative models is the step to the truly personalized and precision medicine.

REFERENCES

[1] L. Tong *et al.*, “Integrating Multi-Omics data with EHR for precision medicine using advanced artificial intelligence,” *IEEE Reviews in Biomedical Engineering*, vol. 17, pp. 80–97, Oct. 2023, doi: 10.1109/rbme.2023.3324264.

[2] M. Moshawrab, M. Adda, A. Bouzouane, H. Ibrahim, and A. Raad, “Reviewing multimodal machine learning and its use in cardiovascular diseases detection,” *Electronics*, vol. 12, no. 7, p. 1558, Mar. 2023, doi: 10.3390/electronics12071558.

[3] U. Haq, M. Mhamed, M. Al-Harbi, H. Osman, Z. Y. Hamd, and Z. Liu, “Advancements in Medical Radiology through Multimodal Machine Learning: A Comprehensive Overview,” *Bioengineering*, vol. 12, no. 5, p. 477, Apr. 2025, doi: 10.3390/bioengineering12050477.

[4] P. Isavand, S. S. Aghamiri, and R. Amin, “Applications of multimodal artificial intelligence in Non-Hodgkin lymphoma B cells,” *Biomedicines*, vol. 12, no. 8, p. 1753, Aug. 2024, doi: 10.3390/biomedicines12081753.

[5] W. Huang, K. Tan, Z. Zhang, J. Hu, and S. Dong, “A review of fusion methods for omics and imaging data,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 1, pp. 74–93, Jan. 2023, doi: 10.1109/tcbb.2022.3143900.

[6] S. Kumar, S. Rani, S. Sharma, and H. Min, “Multimodality Fusion Aspects of Medical Diagnosis: A Comprehensive Review,” *Bioengineering*, vol. 11, no. 12, p. 1233, Dec. 2024, doi: 10.3390/bioengineering11121233.

[7] X. Pei, K. Zuo, Y. Li, and Z. Pang, “A review of the application of multi-modal Deep learning in Medicine: bibliometrics and future directions,” *International Journal of Computational Intelligence Systems*, vol. 16, no. 1, Mar. 2023, doi: 10.1007/s44196-023-00225-6.

[8] B. Abdikenov *et al.*, “Future of Breast Cancer Diagnosis: A review of DL and ML applications and emerging trends for multimodal data,” *IEEE Access*, p. 1, Jan. 2025, doi: 10.1109/access.2025.3585377.

[9] L. Zhuang, S. H. Park, S. J. Skates, A. E. Prosper, D. R. Aberle, and W. Hsu, “Advancing precision oncology through modeling of longitudinal and multimodal data,” *IEEE Reviews in Biomedical Engineering*, pp. 1–19, Jan. 2025, doi: 10.1109/rbme.2025.3577587.

[10] Termine *et al.*, “Multi-Layer Picture of Neurodegenerative Diseases: Lessons from the Use of Big Data through Artificial Intelligence,” *Journal of Personalized Medicine*, vol. 11, no. 4, p. 280, Apr. 2021, doi: 10.3390/jpm11040280.

[11] Gupta *et al.*, “Bringing machine learning to research on intellectual and developmental disabilities: taking inspiration from neurological diseases,” *Journal of Neurodevelopmental Disorders*, vol. 14, no. 1, May 2022, doi: 10.1186/s11689-022-09438-w.

[12] Y.-M. Chen, T.-H. Hsiao, C.-H. Lin, and Y. C. Fann, “Unlocking precision medicine: clinical applications of integrating health records, genetics, and immunology through artificial intelligence,” *Journal of Biomedical Science*, vol. 32, no. 1, Feb. 2025, doi: 10.1186/s12929-024-01110-w.

[13] N. Parvin, S. W. Joo, J. H. Jung, and T. K. Mandal, “Multimodal AI in biomedicine: Pioneering the future of biomaterials, diagnostics, and personalized healthcare,” *Nanomaterials*, vol. 15, no. 12, p. 895, Jun. 2025, doi: 10.3390/nano15120895.

[14] M. Zack *et al.*, “AI and Multi-Omics in Pharmacogenomics: A New Era of Precision Medicine,” *Mayo Clinic Proceedings Digital Health*, p. 100246, Jun. 2025, doi: 10.1016/j.mcpdig.2025.100246.

[15] L. Qian, X. Lu, P. Haris, J. Zhu, S. Li, and Y. Yang, “Enhancing clinical trial outcome prediction with artificial intelligence: a systematic review,” *Drug Discovery Today*, p. 104332, Mar. 2025, doi: 10.1016/j.drudis.2025.104332.