# Towards a robust healthcare prediction model using an adaptive multimodal fusion based on hierarchical transformers

*Lamiae Hachad*[1], *Tarik Hachad*[1]

[1]Computer Science Research Laboratory(LRI), Faculty of Sciences, Ibn Tofail University, Morocco

**Abstract.** The Healthcare field has applied multimodal learning that combines diverse data types to improve the quality of predictions in clinical settings in terms of precision. The limitations of such multimodal approaches are highly related to heterogeneous structure, modality relevance, noise, and data size. In this paper, we present an adaptive multimodal fusion model that assigns distinct importance weights to modalities via attention-based pooling within a hierarchical transformer architecture. Indeed, the proposed model derives features for each modality independently and then aggregates cross-modal features using a hierarchical attention mechanism. We evaluate our architecture in a controlled setting across multiple simulations, and we demonstrate that our model performs effectively across four data types for predicting health. Besides, our experiments revealed clear improvements in validation AUC and F1 scores, especially when the data is limited or noisy, which supports the strength and reliability of our hierarchical transformer-based fusion approach.

***Keywords*-- Multimodal Fusion, Hierarchical Transformers, Healthcare Prediction, Generalization and Robustness, Adaptive Attention Mechanisms.**

## 1 Introduction

The healthcare domain and medical technology have witnessed a drastic transition driven by the generation of large volumes of heterogeneous data from diverse sources, including electronic health records, genomic sequencing, radiological imaging, sensor monitoring, and patient-reported outcomes [1]. The data gathered may come from different sources, systems, norms, and file formats, and the detail levels are not standardized across them; hence, inconsistent integrations occur. The reason for the complex integration is the diversity, missing variables, and the complexity of feature interdependency issues. Recent studies indicate that multimodal fusion techniques can take advantage of the datasets to raise the diagnostic level, improve the clinical results, and provide a better holistic and personalized view of the patient's health condition compared with unimodal analysis [2], [3]. However, due to ongoing difficulties in managing dimensionality imbalance, integrating various data types, and ensuring model interpretability, the use of multimodal data fusion in healthcare remains limited [4], [5]. Most existing fusion methods use static concatenation of modality features, assuming equal importance across inputs and ignoring contextual relevance [5]. As a result, standard multimodal learning methods lack the generalizability across diverse patient populations and do not have adaptive mechanisms to consider the contributions of each modality, limiting their ability for modelling more complicated non-linear relationships and high vulnerability to noise, missing data, and quality variations typically seen in a real-world clinical setup. In this paper, we present an Adaptive Multi-Modal Fusion with Hierarchical Transformers through a 2-level framework that jointly models intra-modality structures and inter-modality dependencies. This is achieved in two stages. The first one is intra-modality level to model patterns for each modality. The second one is at the inter-modality level by incorporating cross-attention for context-propagated fusion. Unlike fixed fusion methods, our method introduces an adaptive clustering method to reweight the contributions of each modality to model the distribution of mixed-modalities by considering the reality of incomplete data. This is experimentally shown to be more interpretable, provide superior performance, and better generalization compared to any baseline fusion method used in this study.

The rest of the paper is organized as follows. In Section 2, we review related work on multimodal fusion in healthcare. In Section 3, we present the proposed Adaptive Multimodal Fusion Framework based on Hierarchical Transformers, detailing its architecture and fusion mechanism. In Section 4, we set up the experiment and present the results with comparisons to other methods. Finally, we close the paper in Section 5, and suggest possible future trends.

https://doi.org/10.1051/bioconf/202520001022

## 2 Related Work

The future of Precision medicine is driven by integrating heterogeneous medical data modalities such as clinical records, medical imaging, and omics data. Multimodal fusion approaches mainly rely on feature concatenation or late fusion, which integrate modality-specific features at the decision level. While previous studies [6], [7], [8] combine electronic health record-derived features with medical imaging and genomic information, they often assume that modalities are equally informative and fail to model complex interdependencies. As a result, it limits their adaptability to variable data quality and missing modalities. More researchers have also dived into deep learning, which learns joint representations across modalities. Medical imaging and clinical data are processed through convolutional and recurrent architectures, incorporating attention mechanisms to emphasize noticeable features within or across modalities [9], [10]. However, most existing models rely on fixed attention mechanisms or unimodal encoders, limiting their capacity to dynamically adjust the importance of each modality based on contextual relevance. Recently, the exploitation of transformer architectures' capabilities has enabled enhanced modelling of long-range dependencies and cross-modal interactions [11]. These architectures still have limited application in the healthcare domain due to data heterogeneity, missing modalities, and the need for interpretability. In this paper, we mitigate these challenges through hierarchical attention mechanism. Our approach introduces a Multimodal Adaptive Fusion Network that directly learns intra- and inter-modality relationships in an end-to-end fashion leading to context-aware, interpretable fusion applicable for real-world healthcare scenarios.

## 3 Proposed Method

Adaptive Multimodal Fusion Framework based on Hierarchical Transformers (AMFF-HT) to integrate heterogeneous healthcare data sources – clinical tabular records, textual notes, medical images, and genomic profiles altogether in the same end-to-end architecture(**Fig 1**). This model becomes general by learning both intra-modality representations and inter-modality dependencies, thus being applicable to modality routing with dynamic and interpretive capabilities.

### 3.1 Intra-Modality Encoding

At the first level of hierarchy, we independently encode each modality to capture its individual properties:

- **Clinical tabular data** are processed by a multi-layer perceptron (MLP) to obtain structured physiological patterns.
- **Clinical text** is tokenized, embedded and modeled using a transformer encoder to adapt the contextual and temporal semantics.
- **Medical images** are then passed through a convolutional backbone (e.g., ResNet-18) to capture spatial and morphological features.
- **Genomic data (RNA-seq)** are encoded via a feed-forward network of gene-expression signatures.

These encoders generate latent embeddings that are specific to each modality.

### 3.2 Inter-Modality Fusion

At the second level of our hierarchy, we perform a cross-attention to learn their inter-modality relationships such that each modality acts as query and allows for context provided by other modalities. This architecture permits the network to promote clinically significant signals while downplaying noise or missing data. This flexible attention mechanism ensures that the model can learn to scale between the contribution ratios of two modalities for a clinical prediction task, based on its importance.

### 3.3 Prediction and Optimization

The classification head of the concatenated multimodal representation for outcome prediction (e.g., mortality, diagnosis) is the final stage. The model is trained end-to-end with binary cross-entropy loss by Adam, and tested by the AUC, F1-score, and accuracy metrics. Dropout and modality-masking strategies were implemented to enhance the generalization and robustness of the model for incomplete data

### 3.4 Interpretability

Interpretability of the model is obtained through attention weights given to each modality. They give an interpretable view about how each modality affects the prediction of the model, allowing clinicians to understand and trust the decisions of such models

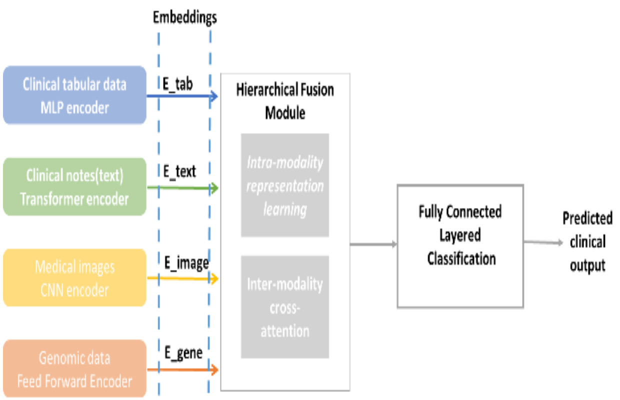


**Fig 1:**Overview of the proposed Adaptive Multimodal Fusion Framework

## 4 Results and Discussion

To assess the performance of the proposed AMFF-HT, we have performed several simulation scenarios. We start by looking at how AUC, F1 and accuracy evolve over training epochs, as well as the trends followed by training and validation loss (**Fig 2**). Indeed, the model AMFF-HT presents a stable and monotonous convergence; as well as in the early stage of stabilizing AUC value before full convergence of F1-score, which further illustrates the refinement of the decision boundary. In addition, the difference between the training and validation losses was small, indicating mild overfitting and good stability of optimization, which is favorable in clinical predictive modeling pipelines.

For the second simulation, we turn on only one modality branch at test time and assess the impact of each modality individually using validation metrics (**Fig 3**). We see that the AUC for imaging features is higher, indicating their discriminative power, and F1-scores are lower because borderline clinical cases will be misclassified. Clinical descriptors in text and table format facilitated the threshold calibration, increasing F1-score by minimizing false negatives. While embedding RNA-Seq results (albeit noisy) improved tail performance of rare patterns to help increase minority-class sensitivity.

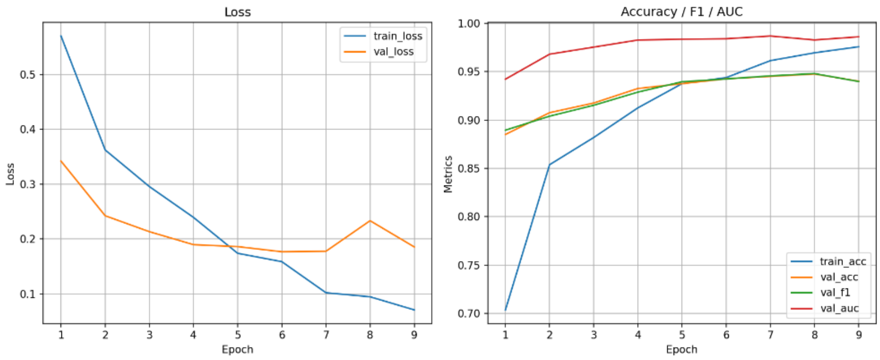


**Fig 2:**Training dynamics of the AMFF-HT

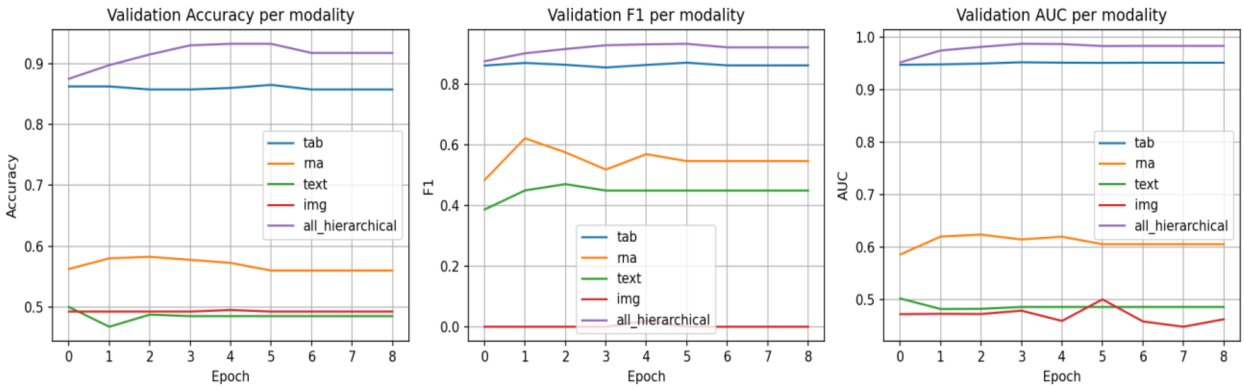


**Fig 3:**Per-Modality Generalization Ability

**Table 1:** Comparative Benchmark of Fusion Strategies

| | AUC | f1 | Accuracy |
|---|---|---|---|
| **Early Fusion** | 0.668 | 0.618 | 0.618 |
| **Late Fusion** | 0.654 | 0.591 | 0.605 |
| **Our Fusion (AMFF-HT)** | 0.721 | 0.675 | 0.680 |

As the third simulation scenario, we evaluated Early Fusion, Late Fusion, and AMFF-HT based on the common metrics (**Table *1***). In fact, the Early Fusion in our case obtained a high raw accuracy as well due to the direct concatenation of features, which, however, suffered from modality interference, reducing modality-specific signals. The Late fusion worked worse as well, supporting the fact that individual experts do not directly benefit from cross-modal synergies when no joint attention mechanism is present. Yet we found the balanced trade-off of AMFF-HT most competitive, providing F1 and AUC stability over sources with high multimodal edge cases. This does confirm our claim that the hierarchical attention pipeline serves as a better tool to align clinical semantics across heterogeneous data streams. These results validate the fact that each modality conveys complementary pieces of information and also substantiate the necessity to combine them hierarchically rather than simply concatenating.

Finally, we studied the effect of dataset size (N= 500, N=1200 and N=2000 patients) on generalization performance of the AMFF-HT (**Fig 4**).

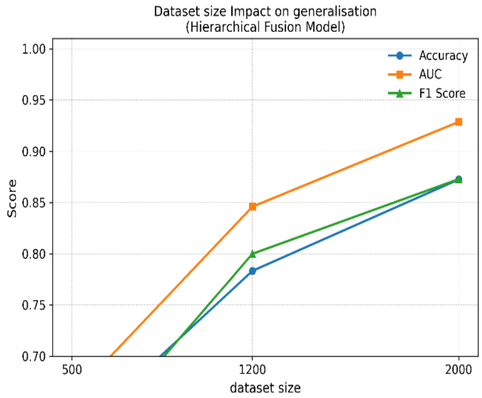


**Fig 4:**Generalization Capacity and Dataset Size Impacts

We note that for 1200 samples, the metrics converge rapidly to a performance plateau, and then show no further improvement while fluctuating with a larger variance as is customary in data-constrained environment. In addition to that, the learning process yielded more stable while AUC scores increased and F1 scores became more consistent when we used 2000 samples. Thus, AMFF- HT benefits from additional multimodal diversity, leading to better-calibrated representations and greater decision robustness. Consequently, we can prove that Hierarchical Fusion architectures are data-efficient, but they still draw significantly on population scaling, which is critical in real-world hospital deployment contexts.

## 5 Conclusion

In this paper, we have presented a hierarchical transformer system which uses multimodal fusion to improve healthcare prediction task performance while enhancing system stability and flexibility. The study has demonstrated that early fusion produces excellent results when data sources align well, but it performs poorly when data sizes are limited. On the other hand, the hierarchical fusion method achieves better results than early fusion because it combines modalities through attention-based refinement at different stages. The evaluation of individual modalities demonstrated that different clinical data types bring unique information to the analysis. As well, no single modality consistently dominates across metrics. Furthermore, the dataset size analysis has shown that the proposed hierarchical fusion maintains stable performance under reduced data availability. It shows less metric variance than early and late fusion baselines. As a perspective, we plan to scale up the framework on real-world clinical datasets. We will also consider dynamic modality selection to guarantee robustness for deployment.

## References

[1] A. Worku, F. G. Arage, and F. B. Kebede, “Introduction to health care data analytics- an overview,” *Discov. Health Syst.*, vol. 4, no. 1, p. 107, Sept. 2025, doi: 10.1007/s44250-025-00291-x.
[2] S. Kumar, S. Rani, S. Sharma, and H. Min, “Multimodality Fusion Aspects of Medical Diagnosis: A Comprehensive Review,” *Bioengineering*, vol. 11, no. 12, 2024, doi: 10.3390/bioengineering11121233.
[3] J. R. Teoh, J. Dong, X. Zuo, K. W. Lai, K. Hasikin, and X. Wu, “Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications.,” *PeerJ Comput. Sci.*, vol. 10, p. e2298, 2024, doi: 10.7717/peerj-cs.2298.
[4] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, “Multimodal affective analysis using hierarchical attention strategy with word-level alignment,” in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2018, p. 2225.
[5] S. Steyaert, Y. L. Qiu, Y. Zheng, P. Mukherjee, H. Vogel, and O. Gevaert, “Multimodal deep learning to predict prognosis in adult and pediatric brain tumors,” *Commun. Med.*, vol. 3, no. 1, p. 44, 2023.
[6] M. T. MacLean *et al.*, “Linking abdominal imaging traits to electronic health record phenotypes,” *medRxiv*, pp. 2020–09, 2020.
[7] A. J. Robertson, A. J. Mallett, Z. Stark, and C. Sullivan, “It is in our DNA: Bringing electronic health records and genomic data together for precision medicine,” *JMIR Bioinforma. Biotechnol.*, vol. 5, p. e55632, 2024.
[8] Y. Hao *et al.*, “Multimodal Integration in Health Care: Development With Applications in Disease Management,” *J. Med. Internet Res.*, vol. 27, p. e76557, 2025.
[9] F. Abdullakutty, A. Bouridane, S. Al-Maadeed, R. Al Saady, and H. M. Abu Rasheed, “Attention-Enhanced Mixture of Experts for Multimodal Breast Cancer Detection,” *Available SSRN 5375894*.
[10] L. Luo *et al.*, “Deep learning in breast cancer imaging: A decade of progress and future directions,” *IEEE Rev. Biomed. Eng.*, 2024.
[11] A. Vaswani *et al.*, “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017.