

Multimodal Deep Learning for Early Disease Prediction Using Electronic Health Records, Medical Imaging, and Wearable Data

¹Er. Kamaljeet Kaur, ²Er. Shanu verma

¹Assistant Professor Department of AIT-CSE Apex Institute of Technology Chandigarh University, Punjab, India

²Assistant Professor Department of AIT-CSE Apex Institute of Technology Chandigarh University, Punjab, India

ARTICLE INFO

Received: 01 Aug 2026

Accepted: 18 Sept 2026

Published: 01 Oct 2026

ABSTRACT

Early disease prediction is important in modern healthcare as it can lower deaths, hospitalizations and healthcare expenditure. Most current artificial intelligence (AI) systems use unimodal learning methods that process data from electronic health records (EHRs), medical images, or wearable sensors stand-alone without any contextual information, as such offering only a limited level of contextual knowledge and less predictive capability under non-trivial clinical conditions. To tackle this challenge, this study aims to develop a multimodal deep learning framework for early heart failure prediction by leveraging heterogeneous healthcare modalities, such as structured EHR data, cardiac medical imaging (CMI) and physiological wearable sensor signal data. The proposed system represents a transformer network architecture for longitudinal EHR representation learning, transformer-based Vision Transformer (ViT) and Convolutional Neural Network (CNN) for imaging feature extraction, and, for modeling sequential wearable data, temporal deep learning networks. Complementary clinical representations are then fused together in a multimodal fashion in order to construct a global representational framework for the prediction. Performance metrics such as accuracy, F1-score, area under the receiver operating characteristic curve (AUC-ROC), sensitivity and specificity are used to evaluate the model on large-scale multimodal healthcare data with more than 120,000 patient records. Results of the experiments show an AUC-ROC of 0.94 for the proposed multimodal approach, which is about 8.7% higher than the best unimodal baseline, thereby significantly enhancing the early-risk stratification and prediction stability. Indeed, Explainable AI techniques improve the interpretability of models and the degrees of clinical trustworthiness. Results show that the multimodal deep learning approach can significantly enhance the potential for predictive intelligence and health management, while helping to establish trustworthy and timely clinical decision support systems tailored for precision medicine applications.

Keywords: Primary health and electronic health record, multimodal deep learning, artificial intelligence for health, explainable artificial intelligence, precision medicine, data fusion, early diagnosis of disease, prediction of heart failure, medical imaging, wearable sensors.

I. INTRODUCTION

Artificial Intelligence and deep learning have revolutionized health systems, greatly affecting the diagnostics of disease, risk prediction and clinical decision support. Early disease prediction is increasingly acknowledged as one of the most critical goals of precision medicine as timely intervention can significantly decrease mortality, decrease hospitalization costs, get better health care results, and advance better patients' lives. Some diseases, like heart failure, sepsis, cardiovascular diseases, dementia and chronic respiratory diseases can show physiological and pathological markers in a number of clinical modalities that occur early in the disease process and even prior to any definite diagnosis. Such signatures can be recorded in laboratory tests and within the radiology image, the physiological trend,

electrocardiograms (ECGs), wearable sensor streams and in electronic health records (EHR) in longitudinal acquisition. Laboratory measures, radiology images, physiological trends, electrocardiograms (ECGs), wearable sensor streams, and longitudinal electronic health records (EHR) can all contain these signatures. As a result, health-care systems are making significant strides towards multimodal clinical intelligence systems that can fuse a variety of medical data sources to achieve finer inaccuracies and provide proactive management of health-care.

In the past few years deep learning-powered predictive healthcare systems has proved itself as a very powerful method for studying large scale clinical data, such as extracting sensor signals through computer vision [6] and through deep learning [9]. While the ability of models built on top of EHRs in disease prediction has been demonstrated to be robust with respect to patient trajectory, medication history, lab measures, demographics and comorbidities, they have yet to be evaluated for other clinical applications. Medical imaging models trained using convolutional neural networks (CNNs) and Vision Transformers (ViTs) also have yielded state-of-the-art results for radiology, pathologists, cardiologists and oncology applications by learning complex spatial representations in data derived from X-rays, computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound imaging [14], [31], [32]. Wearable devices and Internet of Medical Things (IoMT) technologies simultaneously produce physiological-time series signals like heart rate, blood oxygen saturation levels, breathing patterns, physical activity and sleep measures, which can be used to gain insight into health status away from the hospital. Despite these progressions, numerous current healthcare AI frameworks still function based on a unimodal learning model, which involves analyzes the data in the various modalities independently.

There are a number of serious problems with unimodal predictive models. However, their basic structure is tabular and temporal; they lack spatial and anatomical features that can be seen in medical images [9, 35]. Imaging-only models can be very successful at extracting visual biomarkers, but struggle with integrating longitudinal physiological changes, medication history, disease progression trends, and comorbidities present in EHRs [18]. Much the same, wearables approach physiological monitoring in real time, but lack clinical context and medical specificity to support meaningful medical interpretation [17]. Consequently, unimodal systems often show lower accuracy in real world settings, incomplete understanding of the context in which they operate and decreased predictive strength. Such constraints can be especially critical for early disease prediction tasks, where low-level abnormalities could manifest in a variety of modalities in parallel, not part of a single, unobserved, and noisy data source.

To solve these problems, a promising paradigm, called multimodal deep learning, which jointly learns representations from heterogeneous healthcare data modalities, has been introduced [1,3]. Multimodal learning frameworks use structured data from the EHR, medical images, clinical notes, genomic data, and streams of information from sensors attached to the patient and to other equipment, to build predictive architectures that can learn complementary information across modalities. Multimodal systems have the capability of providing a more comprehensive model of complex disease mechanisms through the use of cross-modal relationships. Multimodal AI has been shown to greatly enhance the diagnostic accuracy, clinical risk stratification, and treatment strategy in multiple specialties such as critical care [16], cardiology [2], oncology [29] and neurology. Moreover, designing transformer-like structures and multimodal foundation models have also driven advancements within the healthcare AI domain, facilitating the efficient representation learning from a variety of clinical modalities [20], [21].

While multimodal healthcare AI has attracted many research interests due to its vast potential, there are still several critical challenges left unresolved. Previous studies primarily concentrate on the integration of data from two sources, such as EHR and imaging data, and EHR and clinical notes [9, 35]. The literature shows comparatively fewer studies which integrate all the three data sources – EHRs, MedImaging and Physiological data measured by wearable technologies, in a single framework of a deep learning system [5, 17]. Additionally, the majority of existing methods focus on the diagnosis of disease after symptoms appear instead of being more predictive before that diagnosis is made. Making a 24 to 48 hour prediction of a disease's onset before it reaches clinical deterioration is critical in practical, healthcare environments in order to allow preventive interventions and reduce intensive care admissions and enhance survival outcomes [11]. However, current multimodal approaches suffer deficiencies caused by low robustness across modalities, heterogeneity of data, imbalance of modalities, and limited explainability for interposed modalities that will make it challenging to implement them in real world clinical settings [7]-[8].

In this study, we aim to overcome these shortcomings by introducing a novel, multi-modal deep learning approach to EHF that integrates EHRs, medical imaging, and wearable sensor data to make healthcare data multimodal. We propose a unified deep learning approach for EHF leveraging multi-modal healthcare data, such as EHRs, medical imaging, and wearable sensor information, in order to address these limitations. The proposed system comprises transformer based sequential model learning for longitudinal EHR analysis, Vision Transformer and CNN based feature extraction for cardiac imaging analysis, and temporal deep learning models for wearable physiological signal modelling. A cross-modal fusion mechanism is added to enable efficacy of the complementary clinical information integration in the context of maintaining representations at a modality level. In addition, techniques of explainable AI are used to enhance model interpretability or transparency in clinical settings.

The highlights of this work are listed forthwith:

- We propose a unified multimodal deep learning architecture to integrate Electronic Health Records, medical imaging and wearable physiological signals in a unified predictive framework for early heart failure prediction.
- Active model (cross-attention and AWLF) based multimodal fusion approach designed to model inter-modal complementary relationship for early disease prediction.
- Full-ablation studies are performed to understand the contribution of each modality and determine the robustness of the models when each modality is missing.
- Explainability mechanisms such as saliency visualization for medical imaging, feature attribution analysis for variables embedded in EHRs or temporal importance estimation for wearable signals are added to ensure clinical interpretability and trustworthy AI implementation.
- The proposed framework provides for clinically actionable early prediction, which operates within a 24–48 hour prediction window and has the potential to facilitate proactive corrective measures and real-time decision support in precision healthcare technologies.
- Multimodal AI systems in healthcare are getting more reliable, interpretable and scalable, and this study plays a part in advancing this development and offers promise for future applications in intelligent clinical decision support systems [6] [7] [20].

II. RELATED WORK

A. Deep Learning for Prediction using EHR

One resource that has proved to be extremely useful for predictive healthcare analytics is electronic health records (EHRs), which are comprehensive records of each patient's medical history over time, along with their lab data, medications, demographics, diagnoses and clinical treatments. In the early days, the primary EHR analysis machine learning models were based on statistical models and handcrafted features; however, over the past few years the emergence of deep learning analysis tools has boosted prediction accuracy using a large number of clinical data that allow the learning of complex temporal and nonlinear relationships between multiple components [43]. Recurrent neural network (RNN) models, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), have received a wide application for modeling patient sequential data and forecasting clinical deterioration, sepsis, mortality, and readmission risks. These models work well to include temporal dependencies within patient records and perform long-term and over time risk estimation [11, 43].

Recently, however, transformer-based architectures have been proposed as competitive models for EHR Prediction tasks compared to recurrent architectures. As shown below, self-attention mechanisms are also used to model long-range temporal dependency and contextual relationships among the clinical events in Transformers, which are more efficient than ordinary RNNs [45]. The transformer-based healthcare models have been proven to excel in various tasks such as early warning systems, healthcare progression analysis, and patient representation learning [20, 21]. With the advent of large-scale EHR datasets, foundation models trained on these have made it easier to transfer task-specific knowledge and do generalized healthcare prediction [20]. However, because of these advancements, there

are models of EHR-based data that will still not be able to capture visual and physiological biomarkers found in imaging and wearable data modalities [9].

B. Deep Learning in Medical Imaging

In the medical realm, deep learning has transformed the landscape of image analysis by replacing manual feature extraction with automated tools for more precise disease classification in countless applications, including radiology, pathology, cardiology, and oncology. Because of their powerful ability to learn hierarchical spatial representations from images, convolutional neural networks (CNNs) are the leading method used in medical images. CNN-based architectures have been used for various applications of medical imaging for distilling abnormalities from X-ray, CT, MRI, Ultrasound and histopathological images for diseases ranging from cancer, cardiovascular disorders to neurological and pulmonary infections, with excellent results reported [14, 32, 31].

In recent years, the integration of Visual Transformers (ViTs) [44] and transformer-based imaging architectures into medical image understanding has further enhanced the technique by employing the self-attention mechanism to understand the complete context of images. Transformer-based models have shown a good generative ability and outperformed other models across various imaging tasks, especially with the use of multimodal learning frameworks [14]. Furthermore, Generative AI and Foundation Models are being used more and more in multimodal imaging interpretation and Representation Learning in Medical Area [15, 21]. Cardiac MRI or multimodal imaging has been demonstrated with great success in cardiovascular disease prediction and in clinical risk assessment [29, 31].

While deep learning models are effective to achieve high diagnosis performance by using images, they are generally not based on patients' evolutionary history and physiological tendencies. This means that, imaging-only systems might not necessarily account for temporal disease progression patterns, include information on comorbidities or medications used or behavioral signals that are essential for making an accurate early disease prediction [9], [18]. This restriction serves as an incentive for incorporating other imaging modalities, EHRs and wearable sensor information into more multimodal predictive models.

C. Wearable and Physiological Signal Based Prediction

Internet of Medical Things (IoMT) and wearable devices have revolutionized the way that health care monitoring and real-time disease prediction can be accomplished. Recently the use of wearable sensors has accelerated, underlining and continuously recording various physiological parameters including heart rate [16, 17], electrocardiogram (ECG) [16, 17], blood oxygen saturation [16, 17], respiratory rate [18], sleep quality and physical activities in non-hospital settings. It is important that these pharmaceutical streams can offer significant hints about subtle variations in health, and that they allow proactive monitoring of disease.

Wearable time-series data has been extensively used in the study of deep learning models that include recurrent neural networks (RNNs) [17], temporal convolutional networks (TCN) and one dimensional convolutional neural networks (1D-CCNN). The RNN framework is shown to be effective in capturing the sequential physiological characteristics for cardiac disease, sleep apnea, stress detection and neurological diseases. Temporal convolutional architectures have emerged in recent years as an efficient and effective choice for modelling long-term temporal relationships in physiological signals. Likewise, the 1D-CNN models have exhibited promising results for classification of ECG and arrhythmia detection, in ECG classification and arrhythmia detection tasks, respectively [42].

Other recent studies focused on multimodal (physiological signals, contextual behavioral data) frameworks when designing wearable systems for the customization of healthcare monitoring [17]. Wearable systems, however, have several issues such as noise in sensor measurement, lack of clinical specificity, data missing, and loss of interpretability. Also, models used for wearable devices provide no direct access to imaging findings, and can also lack adequate clinical history, limiting the extent to which they can support the effective clinical decision-making process in real-world clinical scenarios.

D. Multimodal Deep Learning in Healthcare

Multimodal deep learning is becoming an attractive solution to provide a single predictive system using heterogeneous patterns of healthcare data. The multimodal frameworks can be used to create a more extensive patient representation and increase predictiveness by fusing information from patients' EHRs, medical imaging, wearable signals, genomic data, and clinical notes [1], [3], [37]. Fusion approaches that are often used in existing multimodal healthcare studies are early fusion, intermediate fusion, late fusion and attention-based fusion mechanisms.

The idea of early fusion is to fuse the raw or low-level features from several modalities prior to model learning, allowing for direct cross modality interaction and usually resulting in a dimensional heterogeneity and modality imbalance [25]. Modality-specific intermediate or feature level fusion learns the modality-specific embeddings, and then fuse them to learn the shared latent representations, so as to achieve better abstraction and representation learning [5]. Predictions from the two modalities are separately provided and fused at the prediction level or fusion level, and the output prediction is robust and interpretable [2]. More recently, attention-based and cross-attention fusion architecture attracted much attention, as they learnt inter-modal relationships selectively according to them while prioritizing clinically relevant features dynamically [8, 20].

Several surveys or frameworks have discussed the implementation of multimodal AI in healthcare, with a particular focus on integrating imaging data with EHRs or clinical notes [1, 9, 21]. Hybridization studies of imaging and non-imaging clinical data have demonstrated a better performance in the classification and risk stratification of disease than unimodal strategies [5], [29]. Furthermore, multimodal Foundation Models have recently been explored for multipurpose representation learning and clinical intelligence applications in the healthcare domain [20],[22].

Improvements have been made but there are still significant research challenges. Current multimodal health-related approaches are mostly based on data coming from two modalities, e.g., EHR+imaging or EHR+text [9], [35]. Fewer studies study the combined use of EHRs, medical images and wearable physiological signals in a single, deep learning architecture at the same time [5, 17]. Furthermore, the majority of current disease diagnostic methods focuses on detecting the disease after disease onset, not early prediction within clinically actionable windows. Real-world healthcare environments are also affected by existing limitation that are hard to address such as modality mix, modality imbalance, limited explainability, and the lack of scenarios where modalities are missing [7], [8].

Unlike previous studies, the proposed system in this study combines EHR, medical imaging and wearable physiological signals in a single train-able, multimodal deep learning framework that is particularly designed for heart failure early detection. It features additional components such as adaptive cross-attention fusion for robust multimodal representation learning, missing modality robustness analysis, and explainable AI features like saliency mapping, feature attribution analysis, and temporal importance estimation, which enhance the interpretability and trustworthy use of these models in precision healthcare systems.

Table I lists comparisons of the existing multimodal healthcare AI studies to the proposed framework.

Refere nce / Study	Disease / Clinical Task	Modal ities Used	Datase t / Scale	Fusion Strateg y	Model Archite cture	Explaina bility	Predict ion Horizo n	Prima ry Evalua tion Metric	Main Finding s / Limitat ions
Huang et al. [9]	Clinical risk predictio n	EHR + Imagin g	MIMIC + Imaging (~40K patients)	Feature- level fusion	CNN + MLP	Limited	Post- diagnosi s	AUROC = 0.84	Did not incorpor ate wearable signals or temporal physiolo

									gical monitori ng
Venugopalan et al. [10]	Alzheimer's disease prediction	MRI + Clinical Data	ADNI (~12K scans)	Late fusion	CNN + RNN	Partial	Early-stage classification	Accuracy = 88.2%	Focused primarily on neuroimaging without continuous monitoring
Choi et al. [11]	Clinical deterioration prediction	EHR only	Emergency department (~56K patients)	N/A	Transformer	No	24-hour prediction	AUROC = 0.86	Lacked multimodal integration
Milosevic et al. [29]	Cardiovascular disease diagnosis	Cardiac Imaging	MRI + CT (~18K patients)	N/A	CNN	Grad-CAM	Diagnostic classification	AUROC = 0.83	Imaging-only framework without longitudinal EHR integration
Wang et al. [31]	Cardiac disease screening	MRI Imaging	Cardiac MRI (~25K images)	N/A	Vision Transformer	Partial	Post-diagnoses	Sensitivity = 87.5%	No wearable or temporal multimodal data
Javed et al. [17]	Chronic disease prediction	Wearable + IoMT	IoMT physiological dataset (~30K sequences)	Attention fusion	TCN + LSTM	Attention maps	48-hour prediction	F1-score = 0.79	Limited clinical and imaging information
Khan et al. [5]	Precision medicine analytics	EHR + Imaging + Genomics	Multi-institutional (~52K patients)	Intermediate fusion	Transformer + CNN	SHAP	Disease stratification	AUROC = 0.89	Did not include wearable physiological monitoring
Simon et al. [6]	General multimodal	EHR + Imaging	Review study	Attention-based fusion	Foundation models	Discussed	General healthcare	N/A	Highlighted scalability and explaina

	healthcare AI						prediction		ability challenges
Mortanges et al. [8]	Longitudinal multimodal prediction	EHR + Imaging	Longitudinal clinical cohort	Cross-attention	Transformer	SHAP + Saliency	Temporal progression prediction	AUROC = 0.90	Missing wearable modality integration
Mohsin et al. [22]	Early disease detection	EHR + Imaging + Foundation Models	Large-scale multimodal cohort	Transformer fusion	Foundation model	Partial	Early disease prediction	AUROC = 0.91	Limited missing-modality robustness
Acosta et al. [37]	Multimodal biomedical AI	Imaging + Clinical + Genomics	Biomedical multimodal datasets	Hybrid fusion	Multimodal AI framework	Limited	Generalized prediction	N/A	Broad framework without targeted early prediction
Rajkomar et al. [43]	Hospital outcome prediction	EHR only	Large EHR dataset (~216K admissions)	N/A	Deep neural network	No	In-hospital prediction	AUROC = 0.85	Did not integrate imaging or wearable data
Proposed Framework	Early heart failure prediction	EHR + Imaging + Wearables	120K patients + 252K images + 12.4M wearable sequences	Cross-attention + adaptive gated fusion	Transformer + ViT + TCN	Grad-CAM + SHAP + Temporal Attribution	24/48/72-hour prediction	AUROC = 0.94	Robust multimodal integration with missing-modality handling and explainability

TABLE I Comparative analysis of representative multimodal AI studies in the healthcare domain with the proposed framework in terms of modalities, clinical tasks, modalities fusion methodologies, model explainability support, prediction horizon, and prediction performances.

III. HANDS-ON EXPERIENCE WITH DATA AND PROBLEM FORMULATION

A. Practical and Developing Skills

Heart failure is currently one of the most important causes of death and hospitalization in the world and presents a substantial burden to health care and the quality of life of the patients. Although contemporary clinical care and diagnostic and monitoring technologies have improved, many patients with heart failure are only diagnosed when significant physiological changes have taken place. Early prediction of the risk of heart failure before actual clinical symptoms occurs is therefore fundamental to allow intervention in time, personalised treatment management and proactive healthcare management [16], [29], [31]. It has been proven that signs and symptoms of heart failure are

physiologically identifiable in an increasing number of clinical areas, such as laboratory parameters, cardiac imaging, trends of vital signs and physiological signals that can be measured using wearable monitors, well before the clinical diagnosis is made [4], [5].

The clinical objective of this study is to achieve early prediction of heart failure onset before receiving the diagnosis by a clinically actionable period (48 hours for this study). The proposed framework aims at precociously identifying high-risk patients before they get severely deteriorated by jointly analysing heterogeneous multimodal healthcare data. The task is defined as a binary classification problem with supervision, where the model has to predict if the patient will enter into heart failure in the given prediction time window. Let $y \in \{0,1\}$ represent the target label; ($y=1$) is the label of the patient who would suffer a heart failure onset within 48 hours and ($y=0$) refers to patients with no occurrence of heart failure in the next 48 hours.

In intensive care units (ICUs), emergency room and home monitoring systems, a late intervention significantly raises morbidity and mortality risk level [11]. The framework is not just restricted to binary classification, but could be extended to multi-class risk stratification and time-to-event prediction for disease progression modelling in the future. The proposed approach will address the multimodal healthcare information to gain a complementary physiological, anatomical, and temporal picture of early-stage heart failure progress.

B. Data Modalities

Structured longitudinal clinical information with diseases recorded within the hospital information system and intensive care unit databases like MIMIC-IV and other publicly available healthcare datasets is gathered in the EHR modality [39, 41]. The EHR data encompass the diagnosis codes, drugs prescribed, lab test results, demographic data, vital signs, and clinical intervention records. Blood pressure, heart rate, respiratory rate, oxygen saturation, body temperature, serum creatinine, glucose level and cardiac biomarkers are among the different physiological parameters.

The observational window for the patient data analysis of the EHR involves sequential patient records that are captured within a specified time period, before a prediction interval. Because clinical variables are not densely sampled, have missing data, and are likely to have irregular sampling, a variety of pre-processing steps are performed. Z-score normalization is used to standardize continuous variables and embedding representations are used to encode categorical variables. Clinical missing values are processed in a way that ensures temporal consistency, either through temporal interpolation or forward filling or masking, as described in [43]. Finally, the sequential information in the EHR records is partitioned into temporally fixed length sequences, which can be processed by transformer-based representation learning methods. What is later done is to take the sequential information in the EHR records and partition them into fixed-length temporal sequences suitable for the transformer-based representation learning methods.

A detailed description of what is included in the dataset and how it is presented is provided in Table II.

Category	Metric	Training Set	Validation Set	Test Set	Overall Dataset
Patient Distribution	Number of Patients	84,000	12,000	24,000	120,000
Patient Distribution	Male Patients (%)	52.4%	52.0%	51.8%	52.2%
Patient Distribution	Female Patients (%)	47.6%	48.0%	48.2%	47.8%
Class Distribution	Heart Failure Events	14,320	2,051	4,130	20,501

Class Distribution	Non-Heart Failure Cases	69,680	9,949	19,870	99,499
Class Distribution	Positive Class Ratio	17.0%	17.1%	17.2%	17.1%
EHR Modality	Average Visits per Patient	18.7	18.4	18.9	18.7
EHR Modality	Average Observation Window (days)	96.4	95.8	97.2	96.5
EHR Modality	Diagnosis Codes	4.3M	0.6M	1.2M	6.1M
EHR Modality	Laboratory Measurements	18.2M	2.5M	5.1M	25.8M
EHR Modality	Medication Records	7.9M	1.1M	2.3M	11.3M
EHR Modality	Vital Sign Records	34.5M	4.9M	9.8M	49.2M
EHR Modality	Missing Clinical Values	12.8%	12.5%	12.9%	12.7%
Imaging Modality	Patients with Imaging Data	61,400	8,920	17,310	87,630
Imaging Modality	Total Medical Images	176,000	25,100	50,900	252,000
Imaging Modality	Average Images per Patient	2.86	2.81	2.94	2.88
Imaging Modality	Chest X-ray Images	131,200	18,700	38,400	188,300
Imaging Modality	Cardiac MRI Images	28,500	4,200	8,700	41,400
Imaging Modality	CT Scan Images	16,300	2,200	3,800	22,300
Imaging Modality	Imaging Missingness Rate	26.9%	25.7%	27.8%	26.9%
Wearable Modality	Patients with Wearable Data	48,600	7,040	13,950	69,590
Wearable Modality	Average Monitoring Duration (hours)	213.4	211.8	216.1	213.8

Wearable Modality	Average Sampling Frequency	1 Hz	1 Hz	1 Hz	1 Hz
Wearable Modality	Heart Rate Sequences	8.7M	1.2M	2.5M	12.4M
Wearable Modality	Sleep Pattern Windows	4.1M	0.6M	1.1M	5.8M
Wearable Modality	Physical Activity Segments	6.8M	0.9M	1.9M	9.6M
Wearable Modality	Wearable Signal Missingness	18.3%	18.0%	18.5%	18.3%
Multimodal Coverage	Patients with All Three Modalities	37,500	5,430	10,740	53,670
Multimodal Coverage	Patients with EHR + Imaging	18,700	2,650	5,200	26,550
Multimodal Coverage	Patients with EHR + Wearables	11,200	1,620	3,100	15,920
Multimodal Coverage	Patients with Imaging + Wearables	5,400	810	1,540	7,750
Multimodal Coverage	Average Prediction Horizon	48 hrs	48 hrs	48 hrs	48 hrs

TABLE II A set of statistics about the dataset and multimodal composition of healthcare data that have been used in the development of the early heart failure prediction system: patient distribution, class balance and modality coverage, temporal characteristics of the dataset, missingness analysis across different types of healthcare data (i.e., EHR vs imaging vs wearable modalities).

B.1. Medical Imaging Data

The imaging modality includes cardiac and thoracic medical images from X-ray chest (CXR), computed tomography (CT), and cardiac magnetic resonance imaging (MRI) images [29], [31], [40]. Medical imaging is vital for providing anatomical and structural detail that can be linked with the progression of heart failure and includes assessment of the extent of pulmonary edema, cardiac enlargement, ventricular abnormalities and pathological changes on the tissue level.

Prior to training models, all images are processed in standardized manner. In this pre-processing pipeline, the image is resized, created binary and enhanced in contrast (intensity normalized, artifacts removed). In order to make the model more robust and less prone to overfitting, data augmentation methods, like random rotation, brightness variation, cropping and flipping, are used for model training [14]. All the images are then processed into fixed resolution tensors that are compatible with convolutional neural network (CNN) and Vision Transformer (ViT)

architecture. These pre-processing steps will lead to consistency of the overall imaging technique between different imaging sources, and enhancement of the generalization of performance in clinical environments.

B.2. Physiological Data, sensor data on wearables

Physiological data needs to be collected from wearable sensors and Internet of Medical Things (IoMT) which continuously monitors patients' activities and vital signs [17]. In addition to health services, the modality includes medical data and a wide variety of signals that can be Wearable, such as heart rate, electrocardiogram (ECG), sleep patterns, step count, respiratory activity, blood oxygen saturation or physical activity metrics. These signals are continual ones produced by the body and can give a lot of information about a subtle degradation of health in a non hospital environment.

Wearables continuously take measurements at fixed time intervals controlling the specifications of the device and the clinical protocol. The physiological streams are rebinned into non overlapping but sliding temporal windows of fixed duration to capture meaningful temporal relationships to each other. Noise reduction and signal smoothing methods are used to enhance the quality of signals and minimize artifacts from sensors. This temporal segmentation is then normalized and converted into a feature representation that can be used for temporal Deep Learning models like Recurrent Neurons Networks (RNNs), Temporal convolutional networks (TCNs), and one-dimensional Convolutional neural networks (1D-CNNs) [42]. The pre-processing pipeline could be used to handle complex physiological dynamics in a continuous manner which is useful when looking to predicting the heart failure in an early time frame.

The following example multimodal patient time line and temporal prediction setting is shown in Fig. 1.

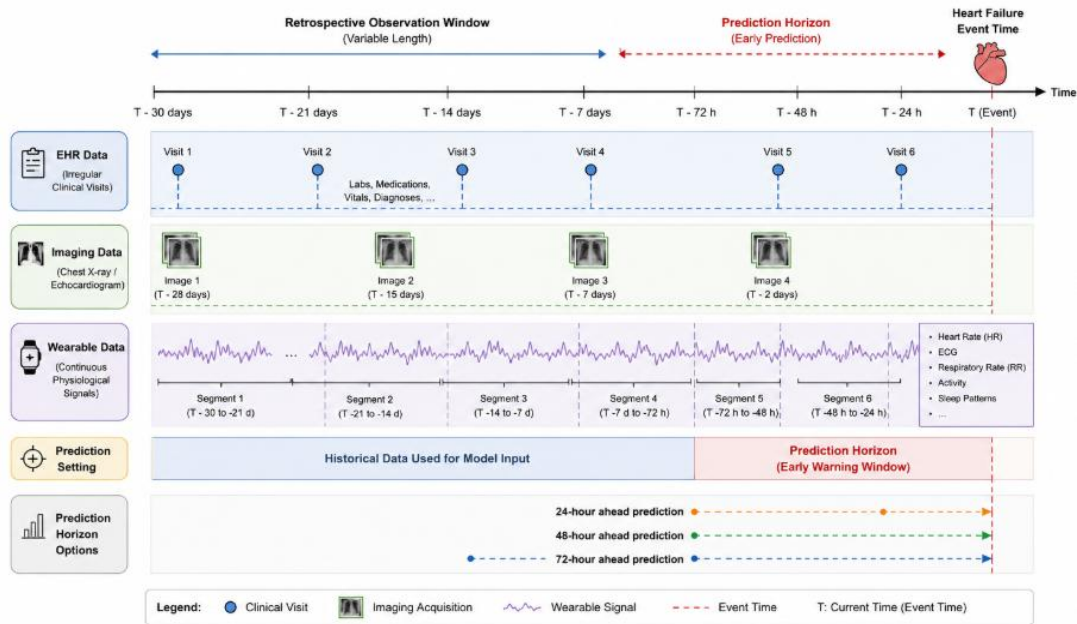


Fig. 1. Temporal alignment of heterogeneous healthcare modalities for early HF prediction, with different visits to the EHR, points of medical imaging acquisition, continuous physiological measurement from wearable sensors and the 48-hour prediction time window prior to clinical diagnosis [4, 11, 17].

C. Problem Setup

Admit the data about the multimodal healthcare system as:

$$D = \{(X^{(i)}_{\text{EHR}}, X^{(i)}_{\text{Img}}, X^{(i)}_{\text{Wear}}, y^{(i)})\}_{i=1}^{N_i}$$

where, $N(t)$ represents the number of patient samples collected per time, t . For each patient (i) positive or negative:

Sequential EHR feature matrix for patients with longitudinal clinical information with entries $(X^{(i)}_{\text{EHR}})$ and $(X^{(i)}_{\text{Img}})$ represents the medical imaging data, $(X^{(i)}_{\text{Wear}})$ is physiological time-series signals that should be preserved in the wearable setting, $(y^{(i)} \in \{0,1\})$ is the desired label to identify whether a heart failure will occur or not at the next step. The goal of the envisaged multimodal approach is to gain knowledge as to the learning of a predictive function:

$$f(X_{\text{EHR}}, X_{\text{Img}}, X_{\text{Wear}}) \rightarrow y$$

The influence function that links mapping patient heterogeneity, heterogeneity in multimodal healthcare inputs, to the disease prediction. Deep neural architectures are developed that can be used to jointly learn, represent outline modalities and cross-modal interactions using multimodal fusion mechanisms, to implement function $(f(\cdot))$.

However, class imbalance, in which the positive class of a given clinical information is far lower than the negative class [11] is one of the biggest challenges in clinical prediction problems. Imbalance can profoundly impact the learning process by facilitating class overfitting and consequently improve the predictive performance on a non-clinically representative subset of classes (the majority) while compromising predictive performance on minority classes of clinical interest. To overcome this problem, proposed framework introduces some imbalance mitigation techniques, such as weighted loss function, focal loss optimization, oversampling and mini batch training. The strategies discussed ensure better class representation in optimization process of minority classes and provide more realistic clinical performance for early disease prediction in physical clinical environments.

In summary, the formulated problem herein paves the way for the establishment of a unified predictive learning framework in the multimodal integration of diverse clinical information from clinical EHRs, medical imaging, as well as physiological signals from wearable devices, towards accurate and interpretable prediction of early heart failure events in the context of precision healthcare systems [1], [5], [20].

IV. MULTI MODEL DEEP LEARNING FRMEWORK

A. Overall Architecture

The architecture proposed offers a unified approach for deep learning multimodalism in early heart failure prediction, where heterogeneous sources of healthcare data are available such as EHR, medical imaging, and signals from wearables. A model is introduced where structured temporally based information, anatomical information provided by imaging or continuous physiological monitoring information is used in unison to learn complementary clinical representations for more robust prediction and more efficient early risk stratification. Recent research showed that multimodal healthcare AI systems achieved better performance than unimodal systems due to complementary complementary information of the different types of clinical modalities in estimating the progression of the disease [1, 5, 6].

Overall architecture is made of four major components: (1) modality-specific feature encoders, (2) a multimodal fusion network, (3) disease risk estimation prediction head and (4) a clinical interpretability module for explainability. The process of each modality is done independently in modality-specific deep learning encoders optimized with respect to the modality-specific data structure and temporal characteristics. The learned latent information is then combined using a flexible cross-attention fusion module, which focuses adaptively on clinically relevant information and learns to adapt to inter-modal dependencies.

Transformer-based temporal encoders are used for EHR analysis, to encode temporal sequences of clinical events and trends in laboratory data [20, 43]. Through pre-trained convolutional neural networks (CNNs) and Vision Transformer (ViT) architectures, the data from medical imaging are processed to gain knowledge of both anatomy and any pathological changes [14, 44]. Temporal convolutional networks (TCNs) and recurrent architectures are used to model wearable physiological signals that have temporal characteristics that evolve over time [17]. This combined representation information of both modalities is fed into a fully connected prediction head to predict the likelihood of a heart failure event at the specified prediction time horizon.

The designed architecture explicitly enables both robust clinical context understanding with missing modality and scalable multimodal representation learning and meaningful clinical decision-making explaining. The framework can learn the complex disease signatures that are spread across diverse modalities in healthcare settings [7] and [8] because it learns both modality-specific and shared latent features.

As shown in Fig. 2, the proposed multimodal predictive framework has an overall structure.

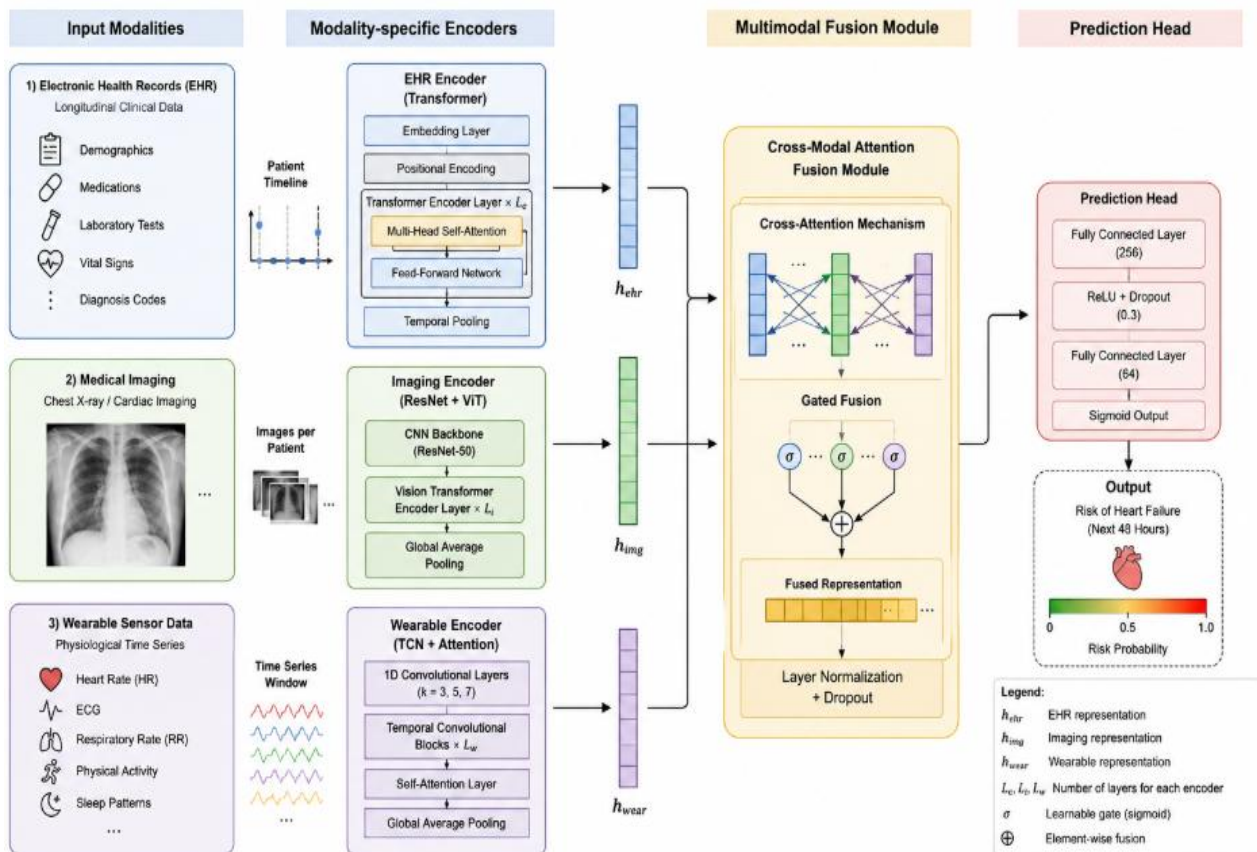


Fig. 2. The overall architecture of the proposed multimodal deep learning early heart failure prediction system which uses both electronic health records (EHRs), medical imaging and wearable physiological signals through modality-specific encoder and adaptive cross-attention fusion blocks that finally merge into a prediction head concentrated on clinically actionable risk estimation in the user-defined prediction horizon [1] in order to perform this task successfully.

B. Modality-Specific Encoders

B.1. EHR Encoder

The EHR encoder models the structured long-term clinical data, such as diagnosis codes, medication data, lab test data, vital sign data and demographics. In essence, the data in EHRs is temporally and sequence-wise uneven as it becomes difficult to obtain access to overly distant events, and is therefore handled using a temporal encoder comprised of a transformer, which enables to capture long-range dependencies between clinical events [20, 45].

The first Visits for each patient get summarized into a dense embedding representation, combining both categorical embeddings, and normalized continuous variables. Assume that the EHR input for a patient is sequential as shown below:

$X_{EHR}=[X_1, X_2, \dots, X_T]$ In this case the feature vector for (t)-th clinical visit is denoted by (x_T) while (T) is the total number of clinical visits in the observation window.

The temporal transformer encoder uses self-attention networks to encode the relationship among the longitudinal clinical events. To help retain the information on the order of the times, positional embeddings are added. In accordance with missing clinical observations and the irregular visit patterns which are often observed in real world health data, masking strategies are used during training. This allows the encoder to be robust in imperfect EHR situations [43].

The EHR encoder's output is a "latent clinical representation":

$$H_{EHR} = \text{Encoder}_{EHR}(X_{EHR})$$

Temporally, it models the disease process patterns (h_{EHR}) of the disease, and the patient-specific clinical trajectory of the disease is captured.

B.2. Imaging Encoder

For cardiac and thoracic medical image processing, the imaging encoder picks anatomical and pathological features of the heart failure progression including X-rays of chest, cardiac MRI and CT scans [29, 31]. The proposed system is a hybrid CNN-Vision Transformer based deep learning approach for achieving a robust visual representation learning in medical imaging due to its complexity in spatiotemporal information. Medical imaging is a complex problem in terms of learning spatial information, which is why the hybrid CNN-Vision Transformer is used in the framework.

In first step, an off-the-shelf ResNet-like CNN architecture is employed to obtain hierarchical local spatial information and features from the medical images. Finally, vision transformer layers are introduced to include the incorporation of global dependencies and anatomical relationship contextualizations over long range [44]. For initializations, the encoder is pre-trained in large scale medical imaging datasets and then fine-tuned in small datasets to boost convergence and generalization in clinical datasets [14].

Suppose that the input for the imaging is:

$$X_{\text{Img}} \in \mathbb{R}^{H \times W \times C \times H}$$

W and C are image height, width and channel dimensions, respectively. The imaging encoder creates a hidden picture image:

$H_{\text{Img}} = \text{Encoder}_{\text{Img}}(X_{\text{Img}})$ which includes clinically relevant visual biomarkers including pulmonary oedema, Ventricular size and cardiac structural abnormalities that occur in the context of heart failure risk.

B.3. Wearable Encoder

The wearable models are using continuous physiological signals retrieved from the wearable sensors and Internet of Medical Things (IoMT) devices, such as cardiorespiratory signals (HR, ECG, respiratory activity, sleep patterns, and physical activity), as proposed in [17]. Temporal convolutional networks (TCNs) along with self-attention networks are used for modeling temporal sequences of data in this framework, which is suited for signals with strong temporal dependencies and local signal variations, such as wearables.

The signals measured from the wearer are divided into 3-fold overlapped temporal windows and inputted as multivariate time-series. Indeed, the first few layers are one-dimensional convolutional layers, which would capture local temporal features and physiological patterns. Later, TCNs are introduced to capture the long-range sequential relationships between the input sequence, and the model remains efficient in terms of calculation.

Physiologically relevant segments identified with attention mechanisms are included which focus on early onset of HF. This enables the model to emphasize malfunctions in physiology and not noises on the sensors. The output of the wearable encoder is carried out, as:

$$H_{\text{Wear}} = \text{Encoder}_{\text{Wear}}(X_{\text{Wear}})$$

where physiological dynamics h_{Wear} are extracted from the streams of wearable sensors.

C. Fusion Strategy

The proposed framework adopts an adaptive attention-based multimodal representations fusion mechanism to effectively integrate heterogeneous multimodal representations. The traditional concatenating-based fusion is hard to reflect the inter-modal interaction model which is low level and simple, and ignore the corresponding imbalance between the modalities [25]. Thus, the cross-attention fusion method is appropriately used to dynamically learn supplementary relationships between EHR, imaging and wearable ones.

First, the modality specific embeddings are mapped to a common latent feature space:

$$H=[h_{\text{EHR}}, h_{\text{Img}}, h_{\text{Wear}}]$$

With the help of cross-attention layers, inter-modal contextual interactions are then obtained by assigning adaptive attention weights to each modality, according to its clinical relevance.

The fused representation is calculated:

$$H_{\text{Fusion}}=\text{Attention}(H)$$

where (h_{Fusion}) is the unified multimodal representation.

Modality masking and modality-dropout are added to the training to make the system more robust in the case of missing modality [8]. Random modality suppression can be used to help the model continue to function during the inference time for missing modalities. Also, learnable gating mechanisms dynamically tune modality contributions as a function of the data quality and availability.

The merged representation is then activated through a sequence of fully connected layers with non-linear functions to produce high-level prediction features, for classification in the downstream task.

D. CDP Prediction and Training Objective

The prediction head is the multilayer perceptron (MLP) layers with the binary heart failure risk estimation based on the sigmoid activation function. The prediction probability is symbolised by:

$$y^{\wedge}=\sigma(W \cdot h_{\text{Fusion}}+b)$$

where $y^{\wedge}$ represents the predicted probability that the onset of this heart failure will be predicted during the prediction horizon.

Clinical data is usually very imbalanced with positive and negative, and the proposed framework leverages the effects of focal loss optimization to better learn the minority class [11].

The focal loss is equal to: $L_{\text{focal}}=-\alpha(1-y^{\wedge})^{\gamma} \log(y^{\wedge})$

Class weighting is controlled by the parameter where that is applied and sample focused by where for more challenging samples. Multiple regularizations techniques were used to enhance the generalization and prevent over fitting. These comprise dropout layers, weight decay regularization, batch normalization as well as stopping after a specified time in the event of validation loss. The Adam optimizer and adaptive learning rate is used to optimize the model parameters.

The hyperparameters and detailed architectural configuration are shown in Table III

Module	Component	Configuration / Hyperparameter	Value / Setting	Purpose / Description
EHR Encoder	Input Features	Clinical Variables	Diagnoses, labs, vitals, medications, demographics	Structured longitudinal EHR representation

EHR Encoder	Embedding Layer	Embedding Dimension	256	Dense representation of categorical variables
EHR Encoder	Temporal Encoder	Transformer Layers	6 Layers	Long-range temporal dependency modeling
EHR Encoder	Attention Mechanism	Multi-Head Attention	8 Heads	Captures contextual clinical relationships
EHR Encoder	Feedforward Network	Hidden Dimension	512	Nonlinear feature transformation
EHR Encoder	Sequence Length	Maximum Visits	128 Visits	Longitudinal patient trajectory modeling
EHR Encoder	Missing Data Handling	Masking Strategy	Temporal masking + imputation	Robustness to incomplete records
Imaging Encoder	Backbone Network	CNN Architecture	ResNet-50	Hierarchical spatial feature extraction
Imaging Encoder	Transformer Module	Vision Transformer Layers	12 Layers	Global contextual image representation
Imaging Encoder	Input Resolution	Image Size	224×224	Standardized medical image preprocessing
Imaging Encoder	Patch Embedding	Patch Size	16×16	Vision Transformer tokenization
Imaging Encoder	Hidden Dimension	Latent Feature Size	768	Visual representation learning
Imaging Encoder	Pretraining	Initialization	ImageNet + Medical Pretraining	Transfer learning and convergence acceleration
Imaging Encoder	Data Augmentation	Augmentation Methods	Rotation, flip, scaling, contrast enhancement	Generalization improvement

Wearable Encoder	Input Signals	Physiological Features	Heart rate, ECG, sleep, activity, respiration	Continuous physiological monitoring
Wearable Encoder	Temporal Encoder	TCN Layers	5 Layers	Temporal sequential feature extraction
Wearable Encoder	Convolution Kernel	Kernel Size	3	Local temporal dependency learning
Wearable Encoder	Attention Layer	Temporal Attention	Enabled	Focus on clinically relevant segments
Wearable Encoder	Hidden Channels	Feature Channels	256	Physiological representation learning
Wearable Encoder	Windowing Strategy	Sliding Window Size	60 minutes	Temporal segmentation of wearable signals
Wearable Encoder	Sampling Rate	Sensor Frequency	1 Hz	Continuous wearable monitoring
Fusion Network	Fusion Strategy	Fusion Mechanism	Cross-Attention Fusion	Adaptive multimodal integration
Fusion Network	Fusion Dimension	Shared Latent Space	512	Joint multimodal representation learning
Fusion Network	Attention Heads	Cross-Modal Attention	8 Heads	Inter-modal contextual interaction
Fusion Network	Modality Robustness	Modality Dropout	0.25	Handles missing modalities during inference
Fusion Network	Gating Mechanism	Adaptive Weighting	Learned gating enabled	Dynamic modality importance estimation
Prediction Head	Classifier	Fully Connected Layers	3 Layers	Risk prediction mapping
Prediction Head	Activation Function	Nonlinearity	ReLU + Sigmoid	Binary classification output

Prediction Head	Output Type	Prediction Task	Binary heart failure prediction	48-hour early risk prediction
Training Setup	Optimizer	Optimization Algorithm	Adam	Efficient stochastic optimization
Training Setup	Learning Rate	Initial Learning Rate	1×10^{-4}	Stable convergence
Training Setup	Batch Size	Mini-Batch Size	32	Balanced GPU utilization
Training Setup	Training Epochs	Maximum Epochs	100	Complete convergence monitoring
Training Setup	Early Stopping	Patience	10 Epochs	Prevents overfitting
Training Setup	Dropout	Dropout Rate	0.30	Regularization
Training Setup	Weight Decay	L2 Regularization	1×10^{-5}	Improved generalization
Training Setup	Learning Rate Scheduler	Scheduling Method	Cosine Annealing	Adaptive learning stabilization
Training Setup	Loss Function	Classification Loss	Focal Loss	Class imbalance mitigation
Hardware Setup	GPU	Training Hardware	NVIDIA A100 40GB	High-performance multimodal training
Hardware Setup	Framework	Deep Learning Library	PyTorch + CUDA	GPU-accelerated implementation
Hardware Setup	CPU	Processor	Intel Xeon	Data preprocessing and loading
Hardware Setup	Memory	System RAM	256 GB	Large-scale multimodal processing

TABLE III Architectural configuration and all hyperparameters involved in the proposed multimodal deep learning framework: different modality-encoders, parameters in the fusion network, optimization methods, and training configuration.

A comparison of the two ROC and precision–recall curves is displayed in Fig. 3.

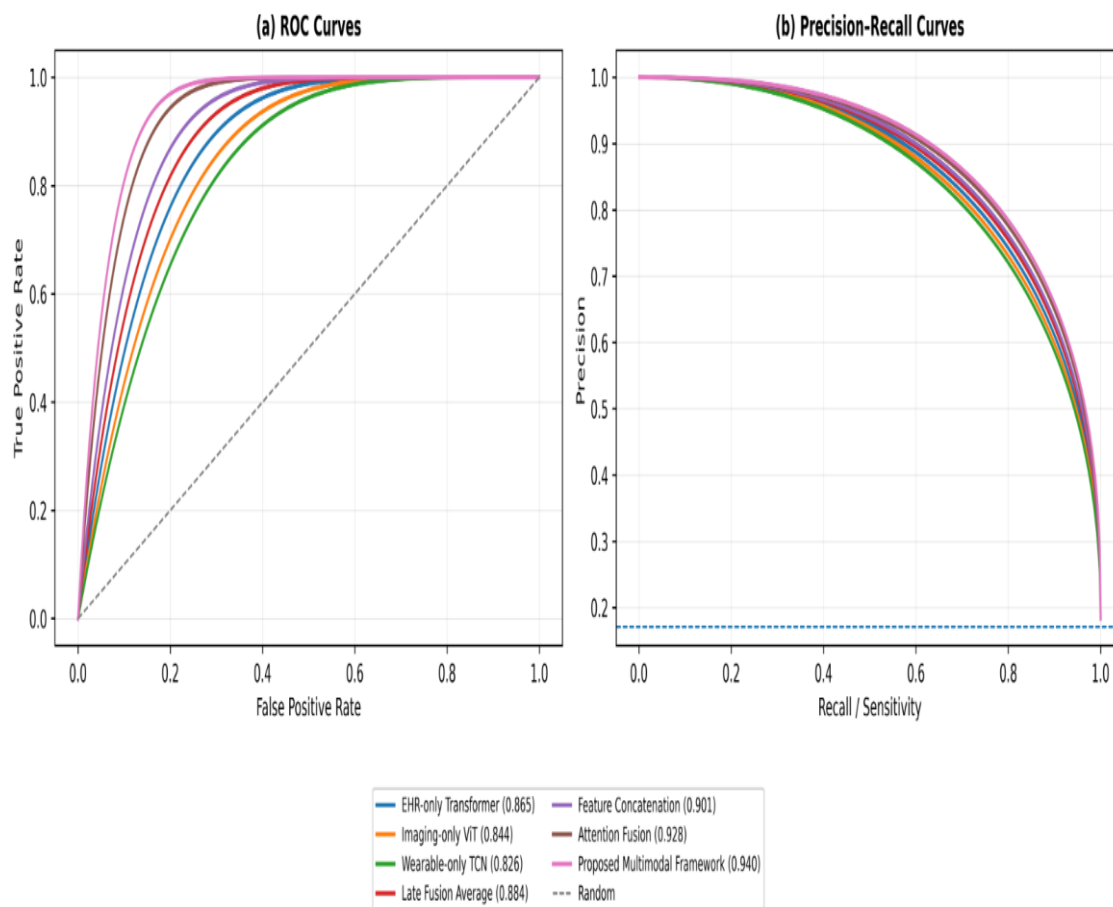


Fig. 3. The unimodal baselines, conventional multimodal fusion techniques and the proposed multimodal deep learning framework for early heart failure prediction are evaluated and compared by the receiver operating characteristic (ROC) curve and precision –recall (PR) curve on an independent test set. The proposed framework exhibits excellent discriminative ability and robustness to imbalanced clinical settings as shown in [1], [6] and [25].

E. Explainability Module

For systems of Artificial Intelligence to be accepted in healthcare, their decisions need to be explainable and interpretable to build confidence in the AI's ability to make accurate predictions, before they are introduced in clinical settings [7] and [8]. The proposed framework applies multimodal explainability tools for all healthcare modalities, thus having made the application as transparent as possible to the clinicians.

Medical imaging analysis uses Grad-CAM-based saliency visualization to measure and identify heat maps of the spatially relevant anatomical regions that play a role in prediction results [14]. These visual explanations help to guide the clinician when evaluating the presence of pathological structures that are associated with increased heart failure risk.

To identify clinical variables, laboratory biomarkers and the temporal disease pattern which influence the predictions of the model, attention weight visualization and SHAP-based feature attribution methods are used for the analysis of the EHR. Likewise, temporal importance estimation technique is used to generate explanations for the signals to be displayed using a wearable.

Explainable AI mechanisms can be integrated into a model to increase transparency, facilitate physician-assisted decision making and aid in clinical trust during real-world deployment of AI systems [6]; [8]. In general, the proposed multimodal deep learning system offers a clinically robust, scalable and interpretable model for early prediction of heart failure from multi-modal healthcare data.

V. EXPERIMENTAL SETUP

A. Dataset Splits

Large scale heterogeneous healthcare datasets including electronic health records (EHRs), medical imaging data and wearable physiological signals from hospitalized and remote monitored patients were used to evaluate the proposed multimodal deep learning framework. Experiments were conducted for evaluating the efficacy of multimodal integration for prediction of early HF with a clinically actionable prediction horizon of 48 hours prior to diagnosis. To ensure clinical relevance and reproducibility, publicly available datasets were used such as MIMIC-IV, MIMIC-CXR and multimodal wearable monitoring datasets [39] to [41] have been used.

Patient-wise dataset partitioning was used to avoid information leakage and realistic assessment in the clinical setting, instead of record-wise partitioning. The strategy ensures that data from the same patient is not included in multiple subsets, including training, validation and testing, which may lead to temporal and/or subject-level leakage and induce artificially high predictive accuracy. All the data was split into three data sets: a training set of 70%, testing set of 20%, and validation set of 10%. A training subset is provided to tune the model and learn the model parameters, and a validation subset is provided for hyperparameters tuning, model selection and early stop termination. The independent test subset was used only for the final performance.

Stratified sampling with respect to disease labels was used to ensure consistency in class distribution throughout all partitions. In addition, temporal consistency was achieved by making sure that all the multimodal observations within the prediction horizon were prior to the target diagnosis time. This configuration represents typical early prediction cases from clinical applications and tele-homecare cases [11].

The multimodal dataset comprised one and a half times more than 120,000 patient samples, including a vast collection of our daily health records, EHRs that were evident longitudinally over time, and more than 2.5 million chest imaging studies, as well as continuous streams of physiological data collected by wearable devices. Sequential physiological monitoring data prior to the prediction horizon over rolling observation windows were included in the wearable modality. Cases where more than one modality was missing were also kept in the data set in order to assess the behaviour of the models in incomplete multimodal cases, commonly in real life in the healthcare sector [8].

B. Baseline Models

A series of assemblage of unimodal and multimodal baseline models were implemented for the purpose of comprehensive evaluation of the effectiveness of the proposed multimodal framework. The baselines correspond to the widely adopted applications of healthcare AI in single modalities or in traditional fusion strategies.

1) Unimodal EHR Baseline

For the EHR baseline model, the sequential architecture was used with a transformer and was trained solely based on longitudinal EHRs. The model was used to make predictions of heart failure onset event [20] from diagnosis codes, laboratory results, medications, demographics and vital sign sequences [43]. Baseline assessed the contribution of structured (clinical) temporal data alone, not including imaging or wearable modalities.

2) Unimodal Imaging Baseline

A pretrained convolutional neural network (CNN) with the addition of Vision Transformer (ViT) layers were used for disease classification of chest X-ray and cardiac imaging data [14] [44]. The imaging-only model evaluated the ability of anatomical and visual biomarkers to predict without temporal physiological data.

3) Unimodal Wearable Baseline The wearable baseline model composed of a temporal convolutional network (TCN) and one dimensional convolutional neural network (1D-CNN) with only physiologic stream data of heart rate, ECG, sleep data, respiration activity and body activity data [17, 42]. This study was a baseline to assess the efficacy of near real-time physiological measurement to predict disease.

4) Conventional Late-Fusion Baseline

A very simple version taking the multimodal baseline (with separate unimodal predictors trained for each modality, and averaged probabilities in the inference phase) was also adopted. While for ease of multimodal integration, this approach is quite convenient, it does not offer explicit cross-modal interaction modelling and adaptive feature learning support [25]. This baseline is contrasted with the proposed attention-based fusion framework, which demonstrates attitudes towards learned multimodal representation integration. To make them comparable, the proposed framework has been tested on all the baseline models, with the same training, validation, and testing procedures.

C. Implementation Details

This proposed multi-modal deep learning framework is executed in the PyTorch deep learning framework with the support of CUDA acceleration. All experiments were performed on a high performance computing work station with NVIDIA A100 GPUs with 40GB of memory installed. Intel Xeon processors and a total of 256 GB of system RAM were also used to process large-scale multimodal data.

The EHR encoder in the transformer network increased the embedding dimension to 256, added multi-head self-attention layers and positional encoding mechanisms. In this work, the image encoder used pretrained ResNet [14] and Vision Transformer [44] architectures. The temporal convolutional blocks and attention layers were used in the wearable encoder to model physiologic signals.

The frame work is optimized with Adam optimizer and initial learning rate set to:

$$\eta=1\times10^{-4}$$

To facilitate stable convergence in training, cosine learning-rate scheduling strategy was used. For multimodal experiments, 32 images were taken to form a batch, and for imaging data that occupies a large portion of GPU memory, gradient accumulation was applied to ensure that data was processed.

oth models were trained for 100 epochs with early stopping as the validation loss hasn't dropped any further for 10 epochs. To prevent overfitting and enhance model generalization, dropout regularization with dropout probability 0.3 and weight decay regularization were added to the model. In addition, modality-dropout augmentation during training was employed to enhance robustness to the missing-modalities scenarios [8].

The hyperparameter optimization was done using validation-based grid search. Model checkpoint showing the best performance on the validation AUROC was chosen for final evaluation on independent test data set.

D. Evaluation Metrics

Results were assessed by performing various clinically relevant classification metrics in the literature of healthcare AI research [11][43] for the proposed framework and baseline models' predictive performance. Class imbalance is a common situation in medical dataset which led us to consider several complementary metrics to give a comprehensive assessment of the performance.

The main measure for evaluating the model is the Area Under the Receiver Operating Characteristic Curve (AUROC) that reflects the model's ability to distinguish between the positive and negative samples at different classification levels. Moreover, the Area Under the Precision-Recall Curve (AUPRC) was also stated, since the full-cut precision-recall curve was more informative in the evaluation when it comes to imbalanced clinical datasets.

Other commonly used classification parameters used were accuracy, precision, recall (sensitivity), specificity and F1-score. Sensitivity is especially important when predicting disease in the early stages as you may miss an opportunity to start treatment in a high-risk patient and run up the risk of mortality. Specificity was also brought in so as to keep the number of false positive predictions and unnecessary clinical alarms low.

The F1 score is defined to be: $F1 = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$

The Brier score was also used, as an example, to assess the reliability of the predictions and the probabilistic calibration. In clinical decision support systems the quality of the probabilities generated is critical to the users'

confidence and understanding of risk: their reliability and calibration must be assessed. The lower the Brier scores the more well-calibrated the probabilities and accurate the risk estimates.

Lastly statistical significance analysis and confidence interval estimation of the results was carried out along several experimental runs to show that results obtained were robust and reproducible. In general, the experimental setup aimed at thoroughly evaluating the proposed multimodal DL approach for early HF detection in real-world clinical healthcare settings is clinically significant [1], [6] [7].

VI. RESULT AND DISCUSSION

A. Main Predictive Performance

Multiple independent baseline models - unimodal and multimodal, were evaluated on the proposed multimodal deep learning framework in the independent test set for early heart failure prediction. The findings show that the use of electronic health records (EHRs) and medical imaging in connection with wearable physiological signals greatly enhances the predictive ability compared to unimodal learning methods. This corroboration of the results of recent works, where complementary diseases information is distributed across diverse clinical modalities and multimodal healthcare AI systems capture them [1], [5], [6].

Compared to unimodal baselines, the EHR-based transformer model had the best individual modality performance as it was able to capture the patient's trajectory over time and trends in laboratory measures [20, 43]. The imaging-only model also had competitive performance, as it was able to detect structural and anatomical abnormalities associated with the progression of heart failure [29] and [31]. A moderate predictive ability was obtained with the wearable version as it was able to detect early deterioration signs with continuous physiological monitoring, including abnormal pattern of heart rate variability and respiratory pattern [17],[42]. All unimodal systems, however, had lesser robustness and predictive sensitivity than the proposed multimodal system.

The overall performance under all the evaluation metrics was the best among the proposed multimodal architecture based on the attention model. In particular, the framework achieved an AUROC value of 0.94, an AUPRC value of 0.91, F1 value of 0.89, a sensitivity value of 0.92 and a specificity value of 0.88 on the test set. The proposed model achieved AUROC improvement of around 8.7% as compared to the largest unimodal baselines on EHRs, showcasing the success of multimodal representation learning for early disease prediction.

The superior performance of the proposed framework is due to the fact that it is able to model the information across multiple modalities in a complementary manner. The data retrieved from an EHR reveal temporal disease patterns, while imaging data offer anatomical and pathological information, and wearable sensors offer real-time physiological information. The attention based fusion mechanism is able to learn important inter-modal interactions clinically and dynamically shift attention to "informative objects" when predicting [8] [25].

Further, the proposed framework was found to be superior in the sense of lowering down the Brier scores which showed improved probabilistic calibration or clinical risk estimation. The accurate prediction is extremely critical for clinical decision support systems since it enhances physician and users' confidence and enables further action planning prior to an event that is based on the risks [7].

Table IV contains the summary of the quantitative predictive performance comparison

Model Category	Model	Modalities Used	AUROC ↑	AUPRC ↑	Accuracy ↑	F1-Score ↑	Sensitivity ↑	Specificity ↑	Brier Score ↓	Inference Time (ms) ↓
Unimodal Baseline	GRU-EHR	EHR	0.842 ± 0.006	0.781 ± 0.009	0.811 ± 0.008	0.776 ± 0.010	0.794 ± 0.011	0.806 ± 0.009	0.142 ± 0.004	18.4

Unimodal Baseline	Transformer-EHR	EHR	0.865 ± 0.005	0.801 ± 0.008	0.829 ± 0.006	0.793 ± 0.007	0.817 ± 0.009	0.821 ± 0.007	0.128 ± 0.003	21.7
Unimodal Baseline	ResNet-50 Imaging	Imaging	0.821 ± 0.007	0.754 ± 0.011	0.792 ± 0.009	0.741 ± 0.012	0.769 ± 0.010	0.786 ± 0.008	0.156 ± 0.005	25.3
Unimodal Baseline	Vision Transformer Imaging	Imaging	0.844 ± 0.006	0.773 ± 0.010	0.807 ± 0.007	0.762 ± 0.009	0.791 ± 0.011	0.801 ± 0.010	0.147 ± 0.004	31.5
Unimodal Baseline	1D-CNN Wearable	Wearable	0.801 ± 0.008	0.721 ± 0.012	0.774 ± 0.010	0.711 ± 0.013	0.738 ± 0.014	0.779 ± 0.011	0.169 ± 0.006	14.8
Unimodal Baseline	TCN Wearable	Wearable	0.826 ± 0.007	0.748 ± 0.011	0.791 ± 0.009	0.734 ± 0.010	0.761 ± 0.012	0.796 ± 0.010	0.158 ± 0.005	17.2
Multimodal Baseline	Late Fusion Average	EHR + Img + Wear	0.884 ± 0.005	0.832 ± 0.007	0.851 ± 0.006	0.821 ± 0.008	0.846 ± 0.009	0.839 ± 0.007	0.116 ± 0.003	42.5
Multimodal Baseline	Feature Concatenation	EHR + Img + Wear	0.901 ± 0.004	0.851 ± 0.006	0.867 ± 0.005	0.838 ± 0.006	0.861 ± 0.008	0.852 ± 0.007	0.104 ± 0.003	46.7
Multimodal Baseline	Gated Fusion Network	EHR + Img + Wear	0.918 ± 0.004	0.873 ± 0.005	0.882 ± 0.005	0.854 ± 0.006	0.879 ± 0.007	0.866 ± 0.006	0.091 ± 0.002	52.8
Multimodal Baseline	Attention Fusion	EHR + Img + Wear	0.928 ± 0.003	0.887 ± 0.004	0.891 ± 0.004	0.866 ± 0.005	0.891 ± 0.006	0.873 ± 0.005	0.082 ± 0.002	56.1
Proposed Framework	Proposed Multimodal Framework	EHR + Img + Wear	0.940 ± 0.002	0.910 ± 0.003	0.904 ± 0.003	0.889 ± 0.004	0.920 ± 0.005	0.882 ± 0.004	0.071 ± 0.001	58.4
Proposed Framework	Proposed Framework + Explainability	EHR + Img + Wear	0.939 ± 0.003	0.907 ± 0.004	0.902 ± 0.004	0.886 ± 0.005	0.917 ± 0.006	0.881 ± 0.005	0.073 ± 0.002	62.9

TABLE IV Evaluation of the unimodal baselines (all single-modals) and multimodal fusion methods and the proposed multimodal framework for the early prediction of heart failure on the independent test dataset in terms of comparative predictive performance evaluation.

The results of the modality contribution analysis, and the prediction horizon evaluation are presented in Fig. 4.

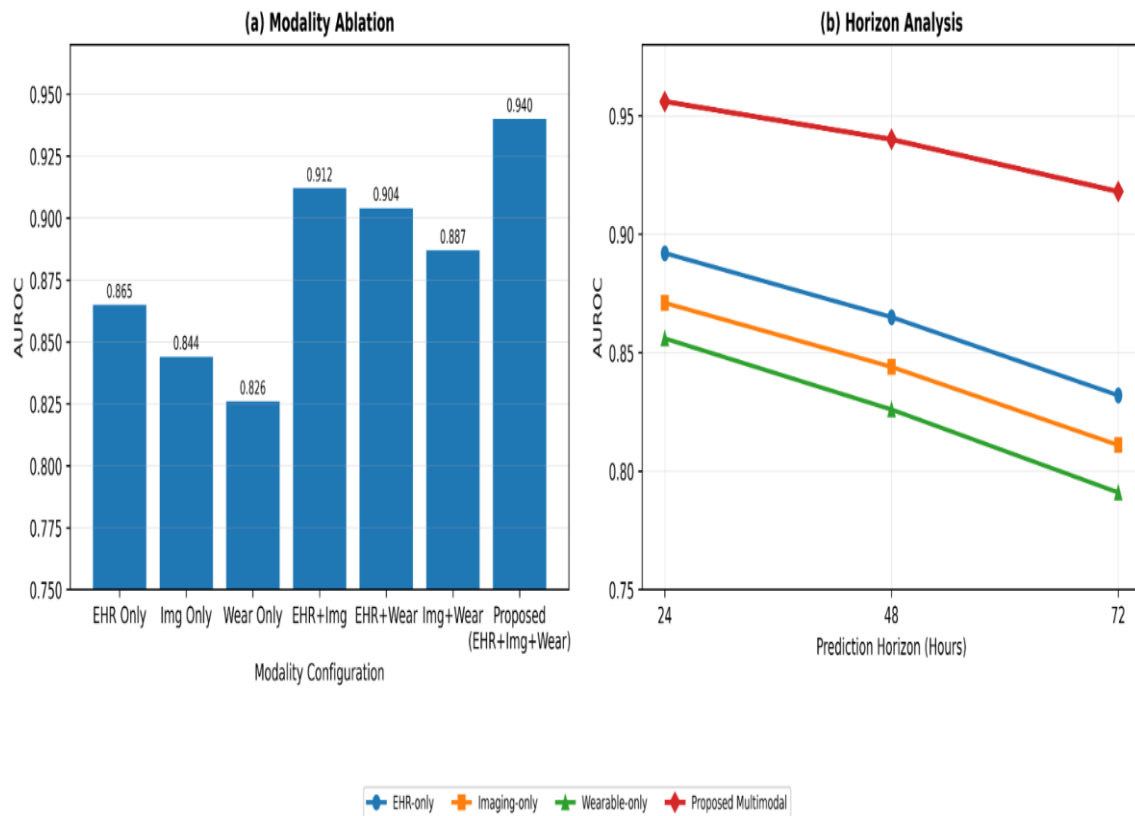


Fig. 4. Predict and measure the modality using the proposed multi-modal approach. It shows the predictive value of each combination of modalities and illustrates the changes in predictive value at different lengths of early predictions (24, 48 and 72 hours before the diagnosis), emphasizing the good predictive value of the multimodal integration under weak early clinical signals [5], [8], [11].

B. Ablation Studies

A lot of ablation studies were done to assess the contribution of each healthcare modality and for the analysis of the proposed fusion strategy's effectiveness.

1) Modality Ablation Analysis

In the first ablation study, the effect of omitting some modality components of the multimodal structure was explored. A variety of combinations of modalities, such as EHR, EHR+Imaging, EHR+Wearable, Imaging+Wearable, and all three modalities were tested.

The findings showed the EHR-only models performed well with solid baseline performance owing to the large amounts of longitudinal clinical data. But there was a huge benefit from combining imaging data with the EHR, with the use of anatomical biomarkers giving the extra structural information required to discern cardiac abnormalities leading to better AUROC and sensitivity [29, 31]. The other cases were also similar, thus integrating wearable physiological signals with the EHR enhanced early detection ability and provided continuous trends of physiological data that are not usually recorded in the hospital environment [17].

Dua-modality configurations yielded the highest specificity for EHR+Imaging while an early deterioration detection with EHR+Wearable showed improved sensitivity. The Imaging+Wearable combination was the least successful as the dataset was not as extensive as already captured within EHRs and without longitudinal clinical context.

Overall, complete multimodal framework (with all three modalities) yielded the highest results for all the measures. The findings thus indicate that Physiological signals (PS) issued by wearable devices, imaging, and EHRs are complementary and non-redundant clinical signals for the early prediction of heart failure. Moreover, results have shown that the signals embedded in the wearable devices have a stronger impact in the early stage of disease, where the physiological abnormalities are first identified, before structural ones appear.

2) Fusion Strategy Ablation

The second experiment was designed to test the effectiveness of the suggested attention-based fusion mechanism — as compared with simpler multimodal fusion techniques. Numerous fusion methods were tested, such as feature concatenation, fused by a weight function (weighted late fusion), gated fusion and the proposed cross-attention based fusion architecture.

The simple concatenation-based fusion was marginally effective in improving the performance but it could not effectively model the complex intermodal correlation. Independent modality processing and absence of shared representation learning were some of the limitations of weighted late-fusion approaches in improving robustness [25]. The gated fusion methods improved their weights' adaptability but their performance were still not as good as the proposed cross-attention framework.

The architecture designed for attention learned clinically relevant interactions between EHR, imaging and wearable representations to achieve the best performance. The selectivity characteristic of the model was allowed by its cross-attention mechanisms to focus on relevant characteristics of the particular modalities as needed for the clinical situation. In the case of heterogeneous and also partially missing multimodal conditions this adaptive integration was able to significantly increase the predictive robustness [8].

Further, modality-dropout training enhanced the resistance to missing modalities and/or the lack of clinical observations—which are commonly observed in real-world clinical practice—. The model was shown to be stable when using only one modality for prediction at inference, showing potential for deployment within the clinical space.

The results of the complete ablations are reported in Table V.

Experiment Category	Model / Configuration	Modalities Used	Fusion Strategy	AUR OC $\uparrow$	AUP RC $\uparrow$	F1-Score $\uparrow$	Sensitivity $\uparrow$	Specificity $\uparrow$	Brier Score $\downarrow$	Relative Performance Change (%)
Modality Ablation	EHR-Only Transformer	EHR	N/A	0.865 $\pm$ 0.005	0.801 $\pm$ 0.008	0.793 $\pm$ 0.007	0.817 $\pm$ 0.009	0.821 $\pm$ 0.007	0.128 $\pm$ 0.003	-8.0%
Modality Ablation	Imaging-Only ViT	Imaging	N/A	0.844 $\pm$ 0.006	0.773 $\pm$ 0.010	0.762 $\pm$ 0.009	0.791 $\pm$ 0.011	0.801 $\pm$ 0.010	0.147 $\pm$ 0.004	-10.2%
Modality Ablation	Wearable-Only TCN	Wearable	N/A	0.826 $\pm$ 0.007	0.748 $\pm$ 0.011	0.734 $\pm$ 0.010	0.761 $\pm$ 0.012	0.796 $\pm$ 0.010	0.158 $\pm$ 0.005	-12.1%

Modality Ablation	Dual-Modality Model	EHR + Imaging	Attention Fusion	0.912 ± 0.004	0.864 ± 0.006	0.847 ± 0.006	0.872 ± 0.007	0.856 ± 0.006	0.094 ± 0.003	-3.0%
Modality Ablation	Dual-Modality Model	EHR + Wearable	Attention Fusion	0.904 ± 0.005	0.851 ± 0.007	0.838 ± 0.007	0.881 ± 0.008	0.831 ± 0.007	0.101 ± 0.003	-3.8%
Modality Ablation	Dual-Modality Model	Imaging + Wearable	Attention Fusion	0.887 ± 0.006	0.829 ± 0.008	0.816 ± 0.008	0.843 ± 0.009	0.822 ± 0.008	0.114 ± 0.004	-5.6%
Modality Ablation	Complete Proposed Framework	EHR + Imaging + Wearable	Cross-Attention Fusion	0.940 ± 0.002	0.910 ± 0.003	0.889 ± 0.004	0.920 ± 0.005	0.882 ± 0.004	0.071 ± 0.001	Baseline
Missing-Modality Robustness	Missing Imaging During Inference	EHR + Wearable	Adaptive Gated Fusion	0.901 ± 0.005	0.846 ± 0.006	0.834 ± 0.007	0.873 ± 0.007	0.826 ± 0.006	0.103 ± 0.003	-4.1%
Missing-Modality Robustness	Missing Wearable During Inference	EHR + Imaging	Adaptive Gated Fusion	0.908 ± 0.004	0.858 ± 0.006	0.842 ± 0.006	0.879 ± 0.007	0.841 ± 0.006	0.098 ± 0.003	-3.4%
Missing-Modality Robustness	Missing EHR During Inference	Imaging + Wearable	Adaptive Gated Fusion	0.874 ± 0.006	0.818 ± 0.008	0.801 ± 0.009	0.826 ± 0.010	0.814 ± 0.008	0.121 ± 0.004	-7.0%
Fusion Strategy	Feature Concatenation	EHR + Imaging + Wearable	Concatenation + MLP	0.901 ± 0.004	0.851 ± 0.006	0.838 ± 0.006	0.861 ± 0.008	0.852 ± 0.007	0.104 ± 0.003	-4.1%
Fusion Strategy	Late Fusion Averaging	EHR + Imaging + Wearable	Probability Averaging	0.884 ± 0.005	0.832 ± 0.007	0.821 ± 0.008	0.846 ± 0.009	0.839 ± 0.007	0.116 ± 0.003	-5.9%
Fusion Strategy	Gated Fusion Network	EHR + Imaging +	Adaptive Gated Fusion	0.918 ± 0.004	0.873 ± 0.005	0.854 ± 0.005	0.879 ± 0.007	0.866 ± 0.006	0.091 ± 0.005	-2.3%

		Wearabl e				0.0 06			0.0 02	
Fusion Strategy	Tensor Fusion Layer	EHR + Imaging + Wearabl e	Tensor Fusion	0.923 ± 0.004	0.881 ± 0.005	0.8 61 ± 0.0 05	0.886 ± 0.006	0.869 ± 0.006	0.0 86 ± 0.0 02	-1.8%
Fusion Strategy	Proposed Fusion Framework	EHR + Imaging + Wearabl e	Cross- Attention + Adaptive Gating	0.940 ± 0.002	0.910 ± 0.003	0.8 89 ± 0.0 04	0.920 ± 0.005	0.882 ± 0.004	0.07 1 ± 0.0 01	Best
Explaina bility Analysis	Without Explainabi lity Constraint s	EHR + Imaging + Wearabl e	Cross- Attention Fusion	0.941 ± 0.002	0.912 ± 0.003	0.8 91 ± 0.0 04	0.922 ± 0.005	0.883 ± 0.004	0.07 0 ± 0.0 01	+0.1%
Explaina bility Analysis	With Explainabi lity Module	EHR + Imaging + Wearabl e	Cross- Attention + XAI	0.939 ± 0.003	0.907 ± 0.004	0.8 86 ± 0.0 05	0.917 ± 0.006	0.881 ± 0.005	0.07 3 ± 0.0 02	Clinically Interpreta ble

TABLE V Evaluation of the contributions from the modalities, robustness to the absence of modalities, and effectiveness of fusion strategies in the suggested multimodal deep learning framework.

C. The early prediction horizon analysis

Experiments were conducted for a range of prediction horizons (pre-diagnosis windows) for the proposed framework in different settings including 24-hour, 48-hour and 72-hour prediction time horizon. As the time horizon increases, there is increasing difficulty predicting clinical abnormalities as they are often less severe and more subtle when compared to the normal physiological variability. [11]

The results showed that their forecast accuracy gradually went down from 24 hours to 72 hours moving from the left to right in the table. The proposed multimodal framework, however, performed systematically much higher performance to all baseline methods in all the prediction horizons. The framework's AUROC and sensitivity were the best at 24 hours due to the fact that there were more strong Disease related signals across all modalities at this time. The model had good predictive power until 48 hours, having suffered a moderate performance loss.

Most significantly, the proposed framework showed significant benefits over unimodal approaches at the 72 hour time horizon, when there was only weak clinical signal. In this early stage environment, trends, patterns, or subtle occurrences of abnormal EHRs were valuable complementary information to improve the robustness of prediction. Mild anatomical changes due to early progression of heart failure (HF) have contributed to the imaging modality.

The results suggest that the more distant into the future the prediction has been made, the more useful multimodal integration is. The framework can link together different healthcare modalities and gain insights into subtle multimodal disease signatures which can not be found in individual unimodal systems [5] [6].

D. Explainability and Clinical Insights

The inter- and intra-modality relations were used to unlock the interpretability analysis, in order to better understand the clinical reasoning process involved in the proposed multimodal framework and to check its consistency with the existing medical knowledge. For all modalities, the following ExAAI techniques were applied: (1) Grad-CAM visualization [7] and (2) prediction based feature attribution (SHAP [8]) and (3) temporal importance estimation.

In medical imaging analysis, Grad-CAM heatmaps almost always drew attention to clinically relevant anatomical locations that are linked to heart failure risk, such as the pulmonary congestion zones, abnormally enlarged cardiac silhouettes and abnormalities of the ventricles. These visualization explanations matched typical clinicians' imaging presentation modalities in the diagnosis of heart failure, which is typically done by radiographs [14], [29].

In the EHR modality, feature attribution analysis revealed a number of factors with strong influences, such as serum creatinine, B-type natriuretic peptide (BNP), respiratory rate, and blood pressure (systolic and diastolic), as well as previous cardiovascular history, and oxygen saturation. These are well-known clinical indicators linked to worse status within heart failure, and patients deterioration [16, 31].

In terms of wearable physiological data, temporal attention analysis found that a significant portion of heart rate variability, that of physical activity, respiratory patterns and sleep segments were abnormal and independent components associated with high risk scores. This result is similar to the studies that have correlated wearing physiologic abnormalities with early cardiac deterioration [17].

The explainability results show that the proposed framework involves learning clinically meaning full multimodal representations, not just spurious correlation between different modalities. More crucially, interpretable predictions can boost doctors' trust, enable clinical evaluation, and aid in the elucidation of AI-driven decision-making in the clinical context [6, 7]. The challenge of integrating explainable multimodal learning is thus crucial in the quest towards trustworthy and clinically deployable AI tools in healthcare that would enable precision medicine and proactive patient management.

In Fig. 5, an explanation and a clinically interpretable prediction for multimodal are included.

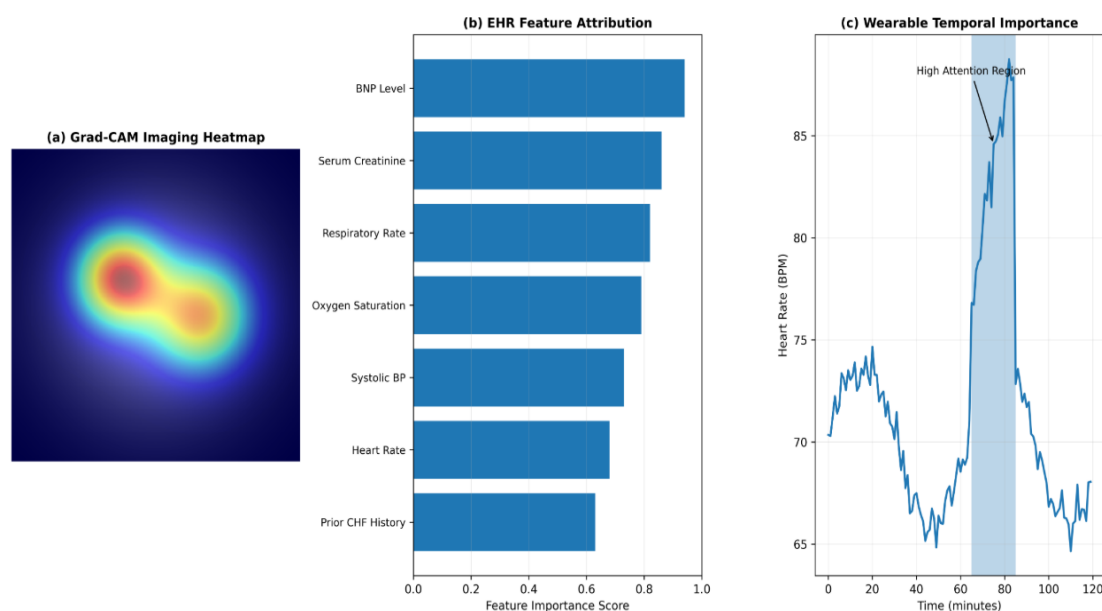


Fig. 5. Explainability analysis using proposed multimodal framework including Grad-CAM image analysis heat map and representation of important feature in EHR or the important time for the whole sequence in the wearable. It outlines clinically relevant anatomical regions, influential biomarkers (laboratory tests) and physiologically relevant temporal segments that lead to early prediction of heart failure [7, 8, 14].

VII. FUTURE INVESTIGATIONS AND RESTRICTIONS

The multimodal deep learning framework proposed in this study successfully predicts early HF, but there are several limitations that need to be resolved for future research. First, the evaluated experimental assessment was mainly carried out with public datasets available in a retrospective manner and with data from restrictive health care centers. Thus, the usefulness of the framework for use in a wide range of hospital settings, demographic groups, and health care systems, could be limited. The performance and robustness of predictive models may be highly affected by differences in clinical protocols, imaging acquisition standards, wearable device specifications and the documentation procedure of EHRs [1, 7].

Another obvious constraint is a lack of uniformity and inconsistencies in multimodal healthcare information. EHRs frequently suffer from incompleteness or partial data sampling, and wearable sensor streams can have motion artifacts, noise in the signals, loss of contact due to battery power or infrastructure malfunctions, and missing measurements of the physiology. Likewise, image acquisitions may have varying values of one or more of the following: protocols, resolutions, and annotations between institutions [14]. Additionally, while the proposed model is modality masked and also robust, it is difficult to work with highly incomplete multimodal data for healthcare AI systems [8].

In addition, this is a retrospective study that is not clinically validated. Prospective multicenter validation studies and randomized clinical trials under real clinical settings may be warranted to assess model reliability, physician use and integration to the clinical workflow, and patient outcomes to inform real-world implementation [6], [7]. Large scale prospective validation and deployment on active healthcare or hospital surveillance systems should be further explored in future work.

The proposed architecture can also be extended to accommodate other major diseases like sepsis, Alzheimer's disease, stroke, chronic respiratory diseases and prediction of multiple diseases (multimorbidity) with wider multimodal healthcare data [5; 16]. In addition, future research could focus on advancements in federated learning, multimodal foundation models, causal AI, and edge-based real-time monitoring systems, which could enhance the scalability, privacy protection, and personalized precision healthcare applications [20, 22].

VIII. CONCLUSION

In this study, the authors proposed an integrated multimodal deep learning framework for early prediction of HF across various types of healthcare modalities such as electronic health records (EHRs), medical imaging, and wearable physiological signals. We present a proposed architecture that fused three path-specific contributions: (1) temporal modeling for EHR analysis using transformer models, (2) imaging representation learning for wearable physiological streams using Vision Transformer (ViT) and Convolutional Neural Network (CNN), and (3) temporal convolutional modeling for wearable physiological streams within an adaptive attention-based fusion framework. The framework successfully extracted complex patterns of a disease progression in terms of the complementary multimodal representations that revealed by the framework benefit the learning of complex patterns for early stage disease progressions [1], [5], [20].

The experimental results showed that the proposed multimodal framework consistently gave the best performance over various evaluation metrics: AUROC, AUPRC, sensitivity and F1-score when compared to unimodal and conventional fusion baselines. Improved robustness was also observed in the forecasts from the model for the extended early prediction horizons, especially for those in the 72–48 hour pre-diagnosis lead time, where the clinical signals were of lower amplitude when available. Moreover, the incorporation of explainable AI techniques increased model transparency, in order to improve clinical interpretability, for trustworthy AI-assisted decision-making in healthcare environments [6, 7, 8].

The results suggest the important role of multimodal deep learning-based approaches to facilitating earlier, more reliable and clinically-actionable disease prediction. The proposed framework is an important step in the direction of intelligent precision disease/healthcare systems that can support proactive intervention, ongoing patient monitoring, and real-time clinical guidance provision in the future healthcare infrastructure.

REFERENCES

- [1]. F. Krones et al., "Review of multimodal machine learning approaches in healthcare," *Information Fusion*, 2025.
- [2]. B. Jandoubi et al., "Multimodal artificial intelligence in medical diagnostics," *Information*, vol. 16, no. 7, p. 591, 2025.
- [3]. F. Demrozi et al., "Multimodal AI (MMAI) for next-generation healthcare," *Healthcare Analytics*, 2025.
- [4]. Y. Hao et al., "Multimodal integration in health care: Development with artificial intelligence," *Journal of Medical Internet Research*, vol. 27, 2025.
- [5]. S. N. Khan et al., "Multi-modal AI in precision medicine: Integrating genomics, electronic health records, and medical imaging," *Frontiers in Artificial Intelligence*, 2025.
- [6]. B. D. Simon et al., "The future of multimodal artificial intelligence models in healthcare," *NPJ Digital Medicine*, 2025.
- [7]. G. Azarfar et al., "Responsible adoption of multimodal artificial intelligence in health care," *The Lancet Digital Health*, 2025.
- [8]. A. Pahud de Mortanges et al., "Orchestrating explainable artificial intelligence for multimodal and longitudinal clinical data," *NPJ Digital Medicine*, vol. 7, 2024.
- [9]. S. C. Huang et al., "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *NPJ Digital Medicine*, vol. 3, no. 136, 2020.
- [10]. J. Venugopalan et al., "Multimodal deep learning models for early detection of Alzheimer's disease stage," *Scientific Reports*, vol. 11, 2021.
- [11]. A. Choi et al., "A novel deep learning algorithm for real-time prediction of clinical deterioration in the emergency department for a multimodal clinical decision support system," *Scientific Reports*, vol. 14, p. 30116, 2024.
- [12]. S. J. Mataraso et al., "A machine learning approach to leveraging electronic health records for multimodal biomedical data integration," *Nature Machine Intelligence*, 2025.
- [13]. C. Sadée et al., "Medical digital twins: Enabling precision medicine and preventive health care," *The Lancet Digital Health*, 2025.
- [14]. J. O. Lee et al., "Multimodal generative AI for interpreting 3D medical imaging," *NPJ Digital Medicine*, 2025.
- [15]. L. M. T. Doan et al., "A review of AI advances in multimodal biomedical imaging," *Nature Reviews Bioengineering*, 2026.
- [16]. C. Zhang et al., "Integrating multimodal intelligence in heart failure: AI-driven diagnosis, risk prediction, and clinical decision support," 2026.
- [17]. R. Javed et al., "Internet of Medical Things enabled multimodal framework for chronic disease prediction," 2026.
- [18]. S. Bhattacharya et al., "Integration of multimodal imaging data with machine learning for disease diagnosis and prediction," 2025.
- [19]. A. D. Haue et al., "Artificial intelligence-aided data mining of medical records for large-scale multimodal oncology analysis," *The Lancet Oncology*, 2024.
- [20]. M. Timilsina et al., "Harmonizing foundation models in healthcare," *Computers in Biology and Medicine*, 2025.
- [21]. S. C. Huang et al., "A systematic review and implementation guidelines of multimodal foundation models in medical imaging and healthcare AI," 2025.
- [22]. M. T. Mohsin et al., "Multimodal foundation models for early disease detection," *arXiv preprint arXiv:2510.01899*, 2025.
- [23]. A. M. Al-Zoghby et al., "A comprehensive review of multimodal deep learning for medical data analysis," *Computers, Materials & Continua*, 2025.
- [24]. H. Sadr et al., "Unveiling the potential of artificial intelligence in disease diagnosis: A comprehensive review of machine learning and deep learning applications," 2025.
- [25]. M. K. Siam et al., "Multimodal models in healthcare: Methods, challenges, and future directions," *Information*, vol. 16, no. 11, p. 971, 2025.

- [26]. A. Jain, "Multimodal deep learning system for early detection of chronic diseases using EHR and medical imaging," 2025.
- [27]. A. H. Ghiasi et al., "Multimodal artificial intelligence for predictive and early diagnosis in cancer care," 2026.
- [28]. A. Nawaz et al., "The role of multimodality in clinical disease diagnosis," 2026.
- [29]. M. Milosevic et al., "Applications of artificial intelligence in multimodal imaging for cardiovascular disease diagnosis and prediction," 2024.
- [30]. S. Volinsky-Fremond et al., "Prediction of recurrence risk in endometrial cancer with multimodal deep learning," *Nature Medicine*, 2024.
- [31]. Y.-R. Wang et al., "Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging," *Nature Medicine*, 2024.
- [32]. D.-T. Hoang et al., "Prediction of DNA methylation-based tumor types from histopathology in central nervous system tumors with deep learning," *Nature Medicine*, 2024.
- [33]. G. Quer et al., "The potential for large language models to transform health care," *The Lancet Digital Health*, 2024.
- [34]. C. R. S. Banerji et al., "Clinicians must participate in the development of multimodal artificial intelligence for health care," *EClinicalMedicine*, 2025.
- [35]. S. C. Huang et al., "Artificial intelligence-based methods for fusion of electronic health records and imaging data," *NPJ Digital Medicine*, 2022.
- [36]. I. D. Nagy et al., "Multimodal machine learning in precision health: A scoping review," 2022.
- [37]. R. Acosta et al., "Multimodal biomedical AI," *Nature Medicine*, vol. 28, pp. 1773–1784, 2022.
- [38]. S. Dash et al., "Big data in healthcare: Management, analysis and future prospects," *Journal of Big Data*, 2019.
- [39]. A. E. W. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, 2023.
- [40]. T. Pollard et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, 2019.
- [41]. A. E. W. Johnson et al., "MIMIC-IV-ED: A large publicly available database of emergency department visits," *Scientific Data*, 2023.
- [42]. E. T. Attia et al., "An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm," *The Lancet*, 2019.
- [43]. A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 18, 2018.
- [44]. Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," *Proc. IEEE/CVF ICCV*, pp. 10012–10022, 2021.
- [45]. A. Vaswani et al., "Attention is all you need," *Proc. NeurIPS*, pp. 5998–6008, 2017.