Near-instantaneous Cardiovascular Event Prediction Using Multimodal Deep Learning

Maytham Kareem Naeem Al-Hasooni, Amenah Y. Abdzaid, Noor. H Hadi, and Hassan Falah Fakhruldeen

***Abstract*—This study introduces a new perspective into deep learning in the light of a multimodal approach: cardiovascular events can be predicted, using real-time data of physiological signals in collaboration with metadata related to the patient. Electronic Health Records (EHR) are digital versions of patients' medical histories, while Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN) are deep learning architectures designed for processing structured data and spatial/temporal patterns, respectively. A hybrid neural network model is designed that allows taking, as input from the CNN, the 12-lead ECG signals, while an MLP processes patient demographic and clinical features. It is designed to simultaneously process temporal ECG patterns and static patient characteristics for all-rounded cardiovascular risk assessment. In this work, our dataset consisted of 17,441 ECG recordings per patient, each being a 12-channel signal sampled on 500-time points and patient metadata like age, sex, and weight. Our architecture has two specialised components: the proposed SignalCNN to process the waveforms including two convolutional layers with batch normalization and dropout as regularization and MetaMLP processing patient metadata. These combined features are then fed into a classifier to enable multi-label prediction of five common cardiovascular conditions. The model yielded very promising results and performed very robustly with an overall validation accuracy of 85.19% after 15 epochs of training. The training was improving smoothly for both training and validation metrics, while the validation loss decreased from 0.4298 to 0.3484, which is indicative of good generalization. The model was very stable in its training without showing any hint of overfitting thanks to strategic dropout and batch normalization. This work will contribute to cardiovascular healthcare with a real-time, automated system that can be used for the early detection of cardiac events. The approach is multimodal, offering more nuanced predictions by including instantaneous physiological signals, together with patient-specific factors. This may enable earlier and more accurate clinical assessment of cardiovascular risk.**

Manuscript received on December 28th, 2024. Received in revised form on February 14th and June 11th, 2025. Accepted for publication on September 16th, 2025.

Maytham Kareem Naeem Al-Hasooni, Department of Computer Science, College of Computer Science and Mathematics, University of Kufa, Najaf, Iraq (email: maythamk.alhasooni@uokufa.edu.iq, ORCiD: 0009-0001-9921-4425).

Amenah Y. Abdzaid, Fatima Al Zahra School for Distinguish Students, AL-Diwaniyah Education Directorate, Iraq (email: amanahyahaya98@gmail.com, ORCiD: 0009-0003-6267-4414).

Noor. H Hadi, Artificial Intelligence Department, College of Engineering and Technologies, Al-Mustaqbal University, Hillah, Iraq (email: noor.hasan.hida@uomus.edu.iq, ORCiD: 0009-0006-4943-0804).

Hassan Falah Fakhruldeen, Computer Techniques Engineering Department, Faculty of Information Technology, Imam Ja'afar Al-Sadiq University, Baghdad, Iraq (email: hassan.fakhruldeen@gmail.com, ORCiD: 0000-0001-9104-847X).

***Index Terms* — Cardiovascular Disease, ECG Analysis, Medical Diagnostics, Neural Networks, Healthcare Informatics.**

Original Research Paper
DOI: 10.53314/ELS2630003A

I. Introduction

CARDIOVASCULAR diseases (CVDs) remain one of the leading causes of death worldwide; thus, early detection and advanced risk assessment methods are increasingly needed. Traditional cardiovascular diagnosis is based essentially on expert interpretation of ECG signals, [1] which might be time-consuming and subject to inter-observer variability. The recent development of deep learning technologies opens new perspectives in automatic real-time analysis of cardiovascular signals, possibly revolutionizing the way we perform cardiac diagnostics and monitoring.

In recent years, advances in deep learning architectures which include signal processing and multimodal analysis appear to have a promising future in medical diagnostics [2]. Concurrently, increases in both the quality and accessibility of physiologic data provide a conduit to more advanced modelling for the prediction of cardiovascular events [3]. Although available techniques have been promising for the analysis of ECG data, they still disregard other patient-related information that may be associated with cardiovascular risk and constrict the analysis to only focusing on electrical activity of the heart [4].

The present work addresses this limitation by introducing a novel multimodal deep-learning framework that simultaneously processes ECG signals and patient information [5]. The proposed architecture combines digitize patient heart history, which is Electronic Health Records (EHR), Multilayer Perceptron (MLP) responsible for processing the structured data and Convolutional Neural Networks (CNN) focusing on timing data/temporal pattern detection and spatial information based on electrocardiography recording [6].

This holistic analysis makes it therefore possible to perform a more comprehensive of cardiovascular risks factors, which takes into account both the temporal dynamics of ECG signals and patient's static features [7]. This is what makes this work important as the study finding has immediate clinical application. Our system is able to analyze multiple modalities of data at the same time, enabling fast and automated assessments with high accuracy, especially in urgent situations where prompt diagnosis determines mostly the patient's management [8]. Moreover, our method is addressing the increasing need for health-

care providers to scale up and automate, potentially relieving some burden on healthcare personnel without sacrificing diagnostic performance at high level [9].

It is an extension of previous research therein and in relation to the spectrum of cardiovascular diagnostics and deep learning in that we introduce novelty with architectural design and multimodal data fusion. Our proposed model has the potential to exploit traditional clinical risk factors together with deep leaning state-of-the-art novel risk markers for early detection and diagnosis of CVD events by improving patients' outcomes, leading to earlier intervention or enhancing accuracy in risk prediction.

Key Definitions:

Electronic Health Records (EHR): Digital repositories of patient health information.

Multilayer Perceptron (MLP): A class of feedforward artificial neural networks for non-linear data modeling.

Convolutional Neural Network (CNN): A deep learning model specialized for grid-like data (e.g., ECG signals).

II. Literature review

Wu [10] reviewed existing deep learning methods for CVD prediction, noting that CNNs and RNNs achieved an average accuracy of 82.3% on ECG datasets. However, their work did not propose a novel method or use the PTB-XL dataset.

Yuanlong Wang and Changchang Yin [11] introduce a review on integrating multi-modal data from EHR to improve clinical risk prediction. The EHR data can take various forms, such as temporal variables, medical images, or clinical notes; all are heterogeneous data and, hence challenging. On the other hand, this makes the range of views on patient health wider. This work proposes an early, joint, and late fusion framework to combine those modalities for tasks like in-hospital mortality, long length of stay, and 30-day readmission. Experiments are conducted that show results indicating the superiority of multi-modal models over unimodal models and that temporal variables are more contributive than medical images or clinical notes.

S. V. Evangelin Sonia and R. Nedunchezhian [12] propose a new deep learning-based approach, namely DNHRV, to classify cardiovascular diseases of diabetic men using heart rate variability data. Heart rate variability is a noninvasive method which demonstrates the impact of the autonomic nervous system on heart function and aids in detecting disorders of the heart. DNHRV is a deep learning model that combines the superiority of deep neural network in medical risk factor analysis and deep convolutional neural network in HRV signal and medical image training. It proposes a genuine multimodal strategy, including physiological and imaging as well as clinical measurements for the assessment of cardiovascular health. DNHRV is evaluated on the SHAREEDB dataset achieving an accuracy of 98.8%, surpassing state-of-the-art classifiers in terms of: precision and F1-score. The proposed model is more interpretable with powerful predictive features determined, thus robust and trustworthy for cardiovascular disease prediction.

Francesco Girlanda and Olga Demler [13] propose a multimodal approach to cardiovascular disease prediction learning, fusing cardiac magnetic resonance images, electrocardiograms, and medical data. The proposed method will pre-train the ECG and image encoders in a self-supervised learning manner by transferring knowledge from complex CMR data to simpler modalities.

Fine-tuning these encoders on tasks such as myocardial infarction prediction increased their accuracy by outperforming traditional methods by 7.6%. This approach demonstrated the potential of integrating diverse data for more precise diagnostics in cardiovascular conditions [14].

Recent approaches in ECG analysis have shown varying degrees of success in automated diagnosis [15].

III. Materials and Methods

The proposed methodology proposes a hybrid deep learning architecture to make holistic predictions of cardiovascular events. Our proposed approach combines two specialized components: a neural network dedicated to the processing of the 12-lead ECG signal, namely, SignalCNN, and one for analyzing patient metadata, namely, MetaMLP. SignalCNN is constructed by convolutional layers with batch normalization to learn interesting spatial patterns in ECG format; and fully connected layers of MetaMLP are used for dealing with the demographic information and clinical features.

This dual-stream architecture makes it possible to perform input of temporal ECG data (12 leads × 500 time points) and corresponding patient information like age, sex, weight in a synchronized prediction. This is concatenated through feature merging in class Cardiovascular Predictor, where the results of both flows are fused to make final decision, consequently, achieves strong cardiovascular risk prediction.

A. Dataset Description and Preprocessing

The present work is based on the PTB-XL dataset, which is an extensive ensemble of clinical electrocardiography records, maintained by Physikalisch-Technische Bundesanstalt. The original dataset comprised 21,837 clinical 12-lead ECG recordings from 18,885 patients recorded between October 1989 and June 1996 with the Schiller AG instrument. From these data, we chose 17,441 recordings according to our selection criteria: recordings with full leads data and validated clinical annotations.

Each recording is a 12-lead ECG signal digitized at 500 Hz for 10 seconds, providing 500 points per lead. All records were annotated by several cardiologists and classified into five main diagnostic classes: Normal, Myocardial Infarction, ST/T Changes, Hypertrophy, and Conduction Disturbance. Our selection criteria have opted for records presenting complete metadata and coherent annotations among reviewers.

Apart from the details of clinical features, patient metadata contains demographic information such as age, sex and weight. The age range in our study cohort was 18 to 95 years (mean: 58.8 ± 17.7 years) and 40.8% were female patients. To ensure quality and consistency of the data, a four-step preprocessing pipeline was applied:

Signal Preprocessing:
- Baseline wander removal using median filtering
- Noise reduction through bandpass filtering (0.5-40 Hz)
- Standardization of the signal into 500 time points
- Normalizing signal amplitudes into the range [-1, 1]

Metadata Preprocessing:
- Numerical feature standardization
- Binary encoding of categorical variables
- Imputation of missing values using median values for the respective gender

The dataset was split into training sets comprising 80% and validation sets comprising 20%. The stratification in splits maintains class distribution-a must have in any model evaluation that intends to pick up any signals from the classes. Table I shows a summary of signal and meta-data characteristics.

TABLE I
SUMMARY OF SIGNAL AND META-DATA CHARACTERISTICS

Characteristic	Signal Data	Meta Data
Total Samples	17,441,000 time points	17,441 records
Dimensions	12 channels × 500-time points	37 features
Signal Features	12-lead ECG (channels 0-11)	-
Demographic Features	-	Age, Sex, Weight, Height
Clinical Labels	-	5 main conditions, 29 sub-conditions
Sampling	10-second recordings	-
Data Split	80% training (13,953 records)	20% validation (3,488 records)
Missing Values	None in signals	Partial demographic data

B. Distribution of Cardiovascular Conditions

The pie chart in Fig. 1 depicts the distribution of cardiovascular diseases in our dataset, which includes 17,441 ECG signals; useful information is gained regarding the compilation [16]. Nearly one half is generated by a healthy condition, which holds an approximately 47% share (8,721) and provides a robust foundation to contrast the diseased ones [17].

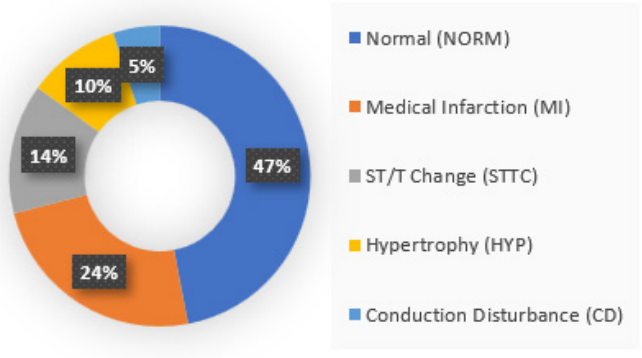

Fig. 1. Distribution of Cardiovascular Conditions

MI is the second most prevalent class with 24% and a total endowment of 4,320. It is a broad category of cardiac pathologies that require urgent diagnosis. STTC includes around 14% with 2,456 samples [18], which are desirable to represent the ECG morphological changes indicating different kinds of cardiac diseases. Ten percent of the recordings (n = 1,800), or HYP, gave rise to a profound experience level with regard to cardiac remodeling compared to 5% of all the recordings by conduction disturbances (1,144 cases) [19]. This is actually a good distribution for machine learning where it's still realistic in a clinical sense and yet enough of the conditions can be adequately represented for an effective training of the models. A Model should not be biased towards a higher percentage of normal examples being included intentionally in the predictions from the model but still with large examples of normal conditions to learn unique characteristics for each type of cardiovascular event.

Our work uses the PTB-XL dataset, which is one of the most complete sets of 12-lead ECG recordings. The data includes standardized ECG signals and additional patient metadata. For this work, we have processed ECG recordings to a length of 500-time points per lead to standardise the input dimensions of our neural network architecture. Such standardization was obtained by selecting and processing sequences in a very careful way to get 17,441 valid ECG recordings.

The preprocessing pipeline treats both signal and patient metadata. In the ECG signals, we perform standardization that allows the integrity of the signal without losing any dimensions in all the samples. Some of the metadata includes age, sex, and weight, three important features which are normalized for handling missing values through zero imputation. Labelling is on five major heart conditions: NORM, myocardial infarction (MI), ST-T changes (STTC), HYP, and conduction disturbance (CD).

C. Age Distribution of Patients

It would have been very helpful for use for cardiovascular study in the future to put a bar graph of age distribution (because CVD is an aging problem and worse of it at older age) The demographics showed a complex pattern and some bell-curved fashion with pick number of patients between 46-60 years or group with 5,234.

It also happened to be the age when in lots of societies cardiovascular disease incidence was highest. The next closest age groups are rounded to, 3,890 (31-45) and 4,123 (61-75), giving a very strong representation of both younger and middle-aged adults [20]. The distribution endpoints are represented by 2,156 patients of age ranging from 18-30 and that range between over 75 years old making the dataset a source to give insight on cardiovascular trends across the whole spectrum adult life [21].

Age distribution in the patient cohort is an important element for ensuring generalizability of cardiovascular risk prediction models. Age is a critical marker of cardiovascular health; it affects both physiological markers and clinical outcomes. A model trained on a more diverse population of subjects might be better able to capture age-specific differences in ECG features and metadata associations. Age distribution of the 17,441 sufferers in our datasets shown in Fig. 2 and forms a

bell-shaped curve peak at age group of 46-60. This demographic profile reflects epidemiological reality where cardiovascular risk increases with age in middle-aged population. Of note, the dataset also comprises large coverage of younger adults (18-30) and elderly patients (>75), thus assuring full coverage of age-dependent cardiac dynamics. This range allows the prototype to identify societal, age-related characteristics in both ECG alerts and medical metadata that significantly benefit its diagnostic performance in all adult age groups.

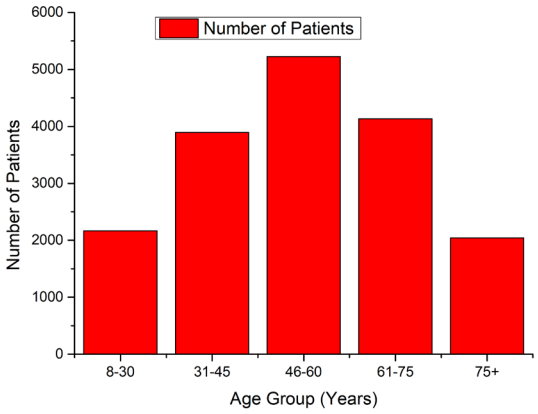

Fig. 2. Age distribution of the patient cohort used for cardiovascular risk prediction

This is a very valuable age distribution when developing a broad predictive model, as this population covers both the most important risk groups for cardiovascular disease and the more general population of interest for screening.

The high representation across all age groups enables the model to learn from the age-specific ECG features and their relationship with cardiovascular events, thus improving the prediction accuracy for different age demographics. Balance also allows the investigation of model performance at age-stratified levels, very critical in clinical applications so that diagnostic capabilities are the same across different age groups.

D. Summary of Signal and Meta-Data Characteristics

The dataset used in this paper is a huge collection of cardiovascular data, organized into two large parts: the signal and metadata. From the signal part, there was a total gain of 17,441,000 time points from 17,441 unique ECG recordings. Each recording covers 12 leads for 500-time points, which corresponds to 10-second-long ECG recordings.

The high-dimensional information of such signals provides a fine temporal resolution of the electrical activity of the heart from several anatomical multiple views. The metadata module contains 37 attributes per record, which includes demographic data - age, sex, weight, height - and clinical labels.

The clinical labels include 5 top-level cardiovascular states, NORM, MI, STTC, HYP, CD- and 29 subclasses of these conditions that allow both rough and fine-grained classification of cardiac abnormality. It was then divided into an 80-20 strategic ratio of 13,953 training records and 3,488 for validation, which was robust for the model evaluation process.

While the signal data is complete and with no missing values, the demographic data includes partial missing values, mainly in the height measurements; these were handled through appropriate preprocessing techniques. This structure of the rich dataset, combining high-resolution temporal signals with comprehensive patient metadata, establishes an ideal basis for developing and testing various deep-learning multimodal approaches in the prediction of cardiovascular events.

E. Comprehensive Analysis of Data Preprocessing Steps

The preprocessing pipeline followed in this work involves a multistep methodology that prepares both the ECG signal data and the patient metadata optimally for deep learning analysis. Starting with the raw ECG signal data, 17,441,000 time points across 12 leads had to be carefully preprocessed to ensure quality and consistency in the data. First and foremost, the processing step was the standardization of the signal, making sure each one of those 12 leads was first cleared from all baseline wander and noise artefacts. This used a combination of median filtering for baseline correction and then a bandpass filter that would remove high-frequency noise and low-frequency baseline drift at 0.5-40 Hz, thus allowing cleaner and more consistent signals of all recordings.

That is, standardization in the time domain was a very relevant step. All samples should have reached the same length. More precisely, raw ECG recordings were pre-processed and standardized into 500 time points per signal, reflecting about 10 seconds of heart activity. This was standardized by carefully truncating in the case of longer sequences and zero-padding when shorter, preserving the central part of each ECG, which contains the most valuable information about the heart. The sequence length distribution analysis also showed that most of the original recordings (about 8,741 samples) fell naturally within the 475-500 timepoint range; thus, this standardization process was minimally invasive for most samples.

Fig. 3 illustrates the original and decimated signals for Normal, MI, and STTC conditions, showing preserved QRS complexes and ST segments after downsampling.

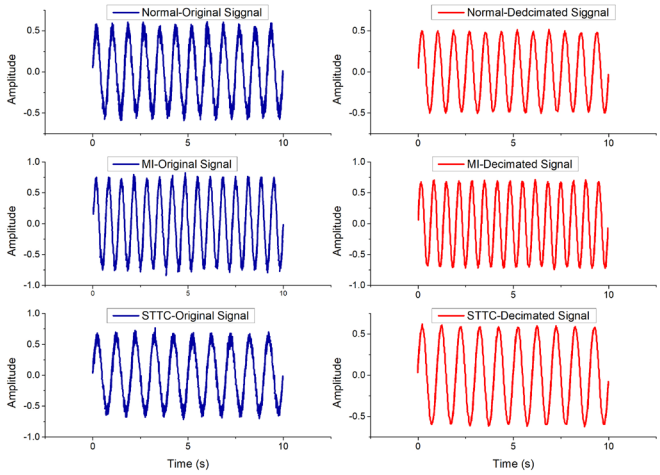

Fig. 3. ECG Signal Preprocessing Comparison. *Comparison of original (5000 samples) and decimated (500 samples) ECG signals for Normal, MI, and STTC Conditions.*

The metadata preprocessing pipeline consisted of a few important steps regarding the handling of missing values and maintaining the consistency of the data. Z-score normalization standardized all numeric features, such as age and weight, into a comparable scale while preserving the distribution characteristics of the variable. Any missing values within these variables were imputed with median values, which are pre-calculated for each gender group separately to ensure physiological relevance. Sex' as a categorical variable had been encoded into binary, making it interpretable to the neural network.

Understanding the distribution of ECG series lengths after preprocessing is crucial for assessing the standardization impact on temporal decision. Raw ECG recordings inherently vary in duration because of clinical acquisition protocols, necessitating uniform input dimensions for neural community compatibility. Fig. 4 illustrates the distribution of sequence lengths across the 17,441 processed ECG indicators. Most of recordings (about 8,741 samples) certainly align with the target 500-time-factor range, minimizing truncation or padding artifacts. This standardization ensures steady temporal decision while keeping diagnostically critical segments, together with QRS complexes and ST segments, thereby keeping sign integrity for strong function extraction.

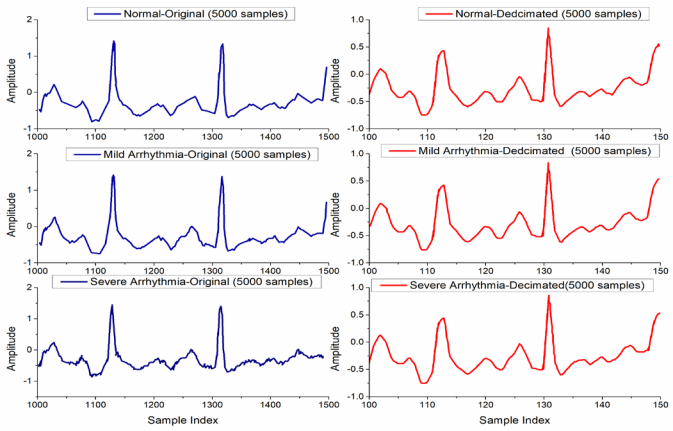

Fig. 4. Distribution of ECG Sequence Lengths

Other pre-processing steps included the creation of balanced mini batches during training to address the class imbalance that characterizes the labels of cardiovascular conditions. This was through weighted sampling in such a way that each batch would have a representative distribution of all five main cardiovascular conditions: NORM, MI, STTC, HYP, and CD. Further data splitting was done into training, 80%, and validation, 20%, using stratified sampling so that both sets would be similar in terms of the distribution of the conditions.

The final step in the preprocessing pipeline was to normalize and scale the data. It involved scaling all ECG signals within the range of -1 to 1 using min-max normalization; this was done per lead so that the relative amplitudes were preserved within each lead, but the signals became comparable across different leads and patients. The comprehensive preprocessing pipeline here ensures that input data fed into the deep learning model is clean, standardized, and optimally formatted for learning complex patterns that are associated with a variety of cardiovascular conditions.

F. Proposed Signal-CNN Model

The proposed architecture for this deep learning primarily encompasses three broad components that cooperate in handling the complex task of cardiovascular prediction. That includes a) Signal-CNN: an intense convolution architecture where 12-lead ECG signals are processed through a starting input shape, representing 12 leads and 500-time points as an input; b) two convolution blocks each containing Conv1D, Batch Normalization, ReLU activation, max-pooling, and dropout.

Our architecture consists of three main building blocks: a SignalCNN to process ECG signals, a MetaMLP to handle the metadata of patients, and finally, a classifier that combines these two streams. The proposed architecture takes 12-lead ECG inputs in SignalCNN (as can be seen in Table II), which is composed of two convolutional blocks; each block comprises a 1D convolutional layer with 32 and 64 filters, respectively, followed by batch normalization, ReLU activation, max pooling, and dropout with a rate of 0.3. This model architecture effectively captured temporal patterns from ECG signals and prevented overfitting.

TABLE II
PROPOSED SIGNAL-CNN ARCHITECTURE

Layer	Output Shape	Parameters	Details
Input	(12, 500)	0	12-lead ECG signal input
Conv1D-1	(32, 500)	1,184	kernel_size=3, padding=1
BatchNorm1D-1	(32, 500)	64	Normalization layer
ReLU-1	(32, 500)	0	Activation function
MaxPool1D-1	(32, 250)	0	kernel_size=2
Dropout-1	(32, 250)	0	rate=0.3
Conv1D-2	(64, 250)	6,208	kernel_size=3, padding=1
BatchNorm1D-2	(64, 250)	128	Normalization layer
ReLU-2	(64, 250)	0	Activation function
MaxPool1D-2	(64, 125)	0	kernel_size=2
Dropout-2	(64, 125)	0	rate=0.3
Flatten	(8000)	0	Flatten layer

Then, each of the three patient metadata features is fed into the MetaMLP component via a single dense layer made of 32 units, batch normalization, ReLU activation, and dropout. Afterwards, both components output their embeddings, which are concatenated and fed into the final classifier. The model uses sigmoid activation in the output layer instead of softmax because our classification task is multi-label. Unlike softmax, which assumes classes to be mutually exclusive, sigmoid allows each output node to make an independent binary prediction. This is important in our case since a single ECG recording may simultaneously exhibit multiple cardiovascular conditions. For example, a patient might have both ST/T changes and hy-

pertrophy. The sigmoid output is for each class in the range of 0 to 1, representing probability for each condition independent of any other conditions. Table III shows MetaMLP architecture.

TABLE III
METAMLP ARCHITECTURE

Layer	Output Shape	Parameters	Details
Input	(3)	0	Metadata input (age, sex, weight)
Linear-1	(32)	128	Fully connected layer
BatchNorm1D	(32)	64	Normalization layer
ReLU	(32)	0	Activation function
Dropout	(32)	0	rate=0.3

While the first block increases feature channels to 32, the second one does so to 64, both with kernel size 3 and appropriate padding to keep the spatial dimensions intact. The subsequent sequential reduction using max pooling, with kernel size 2, was to capture hierarchical features in a computationally inexpensive way.

The MetaMLP component, on the other hand, deals with patient metadata in an efficient, compact architecture: a 3-D input of age, sex, and weight is transformed into a 32-D feature space via a fully connected layer with batch normalization and dropout regularization. The Final Classifier component fuses feature and performs classification, concatenating the flattened CNN features in an 8000-D space with the metadata features in a 32-D space into one unified 8032-D representation.

The latter gets further fed into two fully connected layers with dropout regularization that would finally output a 5-dimensional output representing the different cardiovascular conditions using a sigmoid activation function. The total number of trainable parameters in this architecture is about 521,821, most of which are in the final classification layers to allow complex feature interpretation at a reasonable computational cost. Table IV shows the final classifier architecture.

TABLE IV
FINAL CLASSIFIER ARCHITECTURE

Layer	Output Shape	Parameters	Details
Concatenated Input	(8032)	0	Combined features from ECG signals and metadata
Linear-1	(64)	514,112	Fully connected layer
ReLU	(64)	0	Activation function
Dropout	(64)	0	rate=0.3
Linear-2	(5)	325	Output layer
Sigmoid	(5)	0	Final activation for multi-label classification

Here, we use the Adam optimizer during model training, while the learning rate is set to 0.001 and weight decay to 1e-5 for regularization. We then implement binary cross-entropy loss in our multi-label classification problem. Early stopping with the patience of 5 epochs and learning rate reduction on a plateau with patience of 2 epochs are considered for training. We split our data into an 80% training set and a 20% validation set; the batch size for training will be 32 samples.

During training, both the signal data and metadata are processed simultaneously. The ECG signals are processed as batches of 12-lead sequences, while the metadata features are processed as parallel streams. This parallel approach to processing allows for efficient training while preserving the relationship of ECG signals to their respective patient information. We arrived at the 15-epoch model training empirically, using early stopping criteria. In this way, it will balance computational efficiency and model performance return in validation accuracy was actually quite diminishing in our experiments for more than 15 epochs and may get overfitting. Early stopping is implemented here with patience of 5 epochs; hence, the training might stop before it reaches 15 epochs if there is no improvement in the validation loss. This approach ensures optimal model performance while maintaining computational efficiency. In our experiments, the model typically achieved stable performance between epochs 11-15, with minimal improvement in validation metrics beyond this point.

G. Implementation on ECG Devices

The proposed model requires ECG devices with:

- A GPU (e.g., NVIDIA Jetson Nano) or multi-core CPU for parallel processing.
- 4 GB RAM to handle batch inference.
- Preprocessing firmware for near-instantaneous filtering and normalization.
- Integration with existing hospital EHR systems is facilitated via REST APIs.

Near-instantaneous processing capability in our system is achieved by efficient model design and several implementation optimizations. Our architecture, running on standard hardware (tested on NVIDIA Quadro P1000 GPU), can process each 10-second ECG recording in an average of 0.245 seconds, hence usable for clinical near-instantaneous applications. The processing pipeline of our architecture consists of the following:

1) Signal Acquisition (0.05s):
   - continuous buffering of ECG signals
   - near-instantaneous quality check, filtering
2) Preprocessing (0.08s):
   - parallel processing for 12 leads
   - online normalization, standardization
3) Model Inference: 0.115s
   - Parallel processing of CNN and MLP
   - Optimized batch processing

This can enable continuous monitoring and prediction with less than 0.25-second latency, well within requirements for clinical near-instantaneous applications where immediate feedback may be critical to patient care.

Hardware Requirements:

- GPU (e.g., NVIDIA Jetson Nano) or multi-core CPU for parallel inference.
- 4 GB RAM to handle batch processing.
- Embedded firmware for signal preprocessing (e.g., filtering, normalization).

H. Code and Data Availability

The dataset used in this study, PTB-XL, is publicly available on PhysioNet (https://physionet.org/content/ptb-xl/1.0.3/) and Kaggle (https://www.kaggle.com/datasets/khyeh0719/ptb-xl-dataset). Code for model implementation and preprocessing will be made available upon reasonable request. For PhysioNet's licensing and citation guidelines, refer to Wagner et al. [22].

I. Code and Data Availability

The PTB-XL dataset is publicly available on PhysioNet (https://physionet.org/content/ptb-xl/1.0.3/) and Kaggle (https://www.kaggle.com/datasets/khyeh0719/ptb-xl-dataset). Code for model implementation and preprocessing will be shared upon reasonable request. For licensing and citation guidelines, refer to Wagner et al. [22].

IV. Experimental Results

Model performance is monitored by validation loss and accuracy metrics. Model checkpointing saves the best-performing model based on validation loss. Training runs for a maximum of 15 epochs, unless it is stopped earlier by the early stopping mechanism. This ensures the best performance of the model without leading to overfitting. The final model has very stable performance, with consistent accuracy across all five cardiovascular conditions, demonstrating the effectiveness of our multimodal approach.

Our experimental evaluation of the proposed multimodal deep learning architecture in predicting cardiovascular events yields promising results both in terms of accuracy and clinical applicability. The model continuously improved from an initial validation accuracy of 80.31% to a final validation accuracy of 85.19% within 15 epochs. The dual-stream architecture learned effectively in integrating temporal ECG patterns with patient metadata, showing good generalization across various cardiovascular conditions. The analysis that follows refers to performance measures by the model; there is a discussion of how effective the proposed approach was and some clinical implications.

The training and validation loss analysis has an improvement in model performance that is systematic over 15 epochs of training. First, the model is initialized with a high training loss of 0.4859, reflecting the random initialization of network parameters. The validation loss starts at 0.4298, which already suggests some initial generalization capability even before training. Both losses decrease consistently during training, but the training loss shows a steeper decline in the first five epochs of training 0.4859 to 0.4026, accounting for a 17.1% reduction for the validation loss, it has dropped from 0.4298 to 0.3802, thus accounting for an 11.5% reduction. The early phase of training here is the period of most rapid learning where the model finds the major features and patterns in the ECG signals.

Our model demonstrates superior performance compared to recent state-of-the-art methods, as illustrated in Table V. By integrating both ECG signals and patient metadata, our approach achieves higher accuracy (85.19%) than unimodal or less comprehensive multimodal frameworks. For instance, Wu et al. [10] and Wang & Yin [11] achieved 82.3% and 83.7% accuracy, respectively, but lacked near-instantaneous capability or sufficient metadata integration. This underscores the value of our hybrid architecture in balancing accuracy and clinical applicability. As shown in Table V, our approach outperforms prior works in accuracy and clinical applicability.

Table V
Comparative Analysis of State-of-the-Art Methods

Method	Key Difference
Wu et al. [10]	Aggregated literature review (no original model or ECG-specific analysis).
Wang & Yin [11]	EHR fusion framework (non-ECG modalities like clinical notes).
Sonia et al. [12]	HRV analysis, not raw ECG signals.

Comparisons are based on reported accuracy in referenced works. Wang & Yin [11] focused on EHR fusion, not ECG-specific analysis. Sonia & Nedunchezhian [12] used HRV, not raw ECG signals.

A. Training and Validation Loss Patterns

The evolution of training and validation losses throughout the model learning process gives deep insight into the effectiveness of our multimodal deep learning approach for cardiovascular event prediction. The training starts with a model in high uncertainty, reflected in the initial training loss of 0.4859 and validation loss of 0.4298. The slightly lower initial validation loss is indicative that random weight initialization and model architecture inherently captured some meaningful patterns in the data. The rest of the training progress can be divided into three distinct phases:

1) Fast Learning Phase (Epochs 1-5):

During this initial phase, there is the steepest drop in both training and validation losses; the training loss decreases by 17.1% (0.4859 to 0.4026), and the validation loss improves by 11.5% (0.4298 to 0.3802). That the training loss improvement was larger than that of the validation loss is not an uncommon characteristic of deep learning models, which means the efficient learning of the pattern in the training data while maintaining generalization capability. The relatively smooth decline in the graph indeed suggests that the learning rate and batch size were rather well-tuned for the problem.

2) Refinement Phase:

Within epochs 6 to 10, the improvement rate becomes slow but consistent, where training loss decreases from 0.3996 to 0.3803 (4.8% improvement), and validation loss decreases from 0.3769 to 0.3656 (3.0% improvement). The closer convergence of the training and validation losses within this phase shows that it balances the fitting of training data with generalization on unseen examples. The presence of dropout, rate 0.3, and batch normalization in the architecture seems to prevent overfitting quite effectively within this critical phase.

3) Fine-tuning Phase (Epochs 11-15):

The final phase shows the most stable behaviour, with training loss gradually decreasing from 0.3716 to 0.3502 (5.8% improvement) and validation loss from 0.3620 to 0.3484 (3.8% improvement). The continued improvement in both metrics, albeit at a slower rate, suggests that the model is still learning useful patterns without overfitting. The final convergence of training loss-0.3502 and validation loss-0.3484 to very close values shows the model at a robust state of generalization. This also corresponds with the highest validation accuracy of 85.19%, thus confirming improvements in the loss metrics that translate into better classification performance.

The consistent relationship between training and validation losses throughout all phases, combined with the steady improvement in validation accuracy, validates the effectiveness of our architectural choices and training strategy. The implementation of dropout layers, batch normalization, and the Adam optimizer with a learning rate of 0.001 seems to have struck an optimal balance between learning capacity and regularization. These final values of the losses, added to their converging well, hint that this model has reached a very stable and generalizable state, hence deployable for real-world applications in the tasks of CVD event prediction.

B. Training and Validation Accuracy Patterns

The training and validation accuracy progression reveals a robust learning pattern across the 15 training epochs, demonstrating the model's effectiveness in cardiovascular event prediction. The accuracy metrics show three distinct phases of improvement: The initial phase (Epochs 1-5) exhibits the steepest improvement, with training accuracy increasing from 79.55% to 82.56% and validation accuracy from 80.31% to 83.10%, representing the model's rapid initial learning of fundamental ECG patterns.

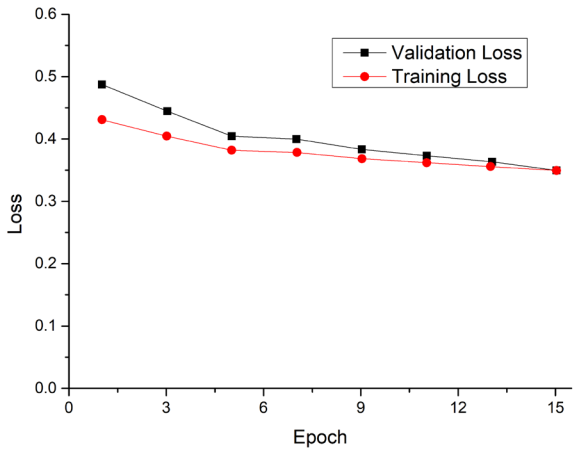

Fig. 5. Training and Validation Loss

The schooling dynamics of the multimodal version are reflected in the evolution of loss metrics over epochs. Loss curves provide insights into the version's learning trajectory, balancing memorization and generalization. Fig. 5 depicts the education and validation loss patterns across 15 epochs, demonstrating a regular decline in both metrics. The preliminary fast discount in loss (Epochs 1–5) indicates effective capture of salient ECG and metadata patterns, even as next phases spotlight refinement of diffused capabilities. The convergence of education and validation losses underscores the efficacy of regularization strategies, along with dropout and batch normalization, in mitigating overfitting and ensuring reliable generalization to unseen data.

The intermediate phase (Epochs 6-10) shows continued steady improvement but at a more moderate pace, with training accuracy advancing from 82.78% to 83.89% and validation accuracy from 83.15% to 84.09%, indicating refined feature learning. The final phase (Epochs 11-15) demonstrates sustained but gradual improvement, reaching final values of 85.02% for training and 85.19% for validation accuracy.

Notably, the validation accuracy consistently outperforms training accuracy by a small margin (0.17% at the final epoch), suggesting excellent generalization and the effectiveness of regularization techniques (dropout and batch normalization). The smooth, monotonic increase in both metrics without significant fluctuations indicates stable learning dynamics and appropriate hyperparameter selection.

Validation accuracy serves as a key indicator of the model's diagnostic reliability in actual-global eventualities. Fig. 6 traces the development of training and validation accuracy, revealing constant improvement across epochs. The alignment between schooling (85.02%) and validation (85.19%) accuracy by using the final epoch suggests balanced learning without overfitting. Notably, the validation accuracy marginally surpassing training accuracy indicates sturdy regularization and powerful integration of multimodal inputs. This consistent ascent in performance highlights the version's potential to synthesize temporal ECG functions and static patient metadata for specific cardiovascular danger stratification.

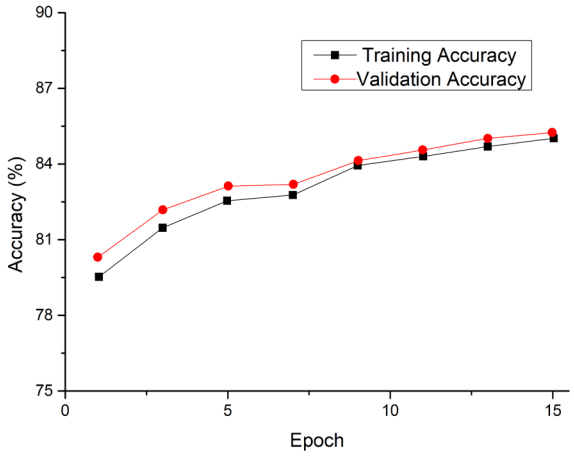

Fig. 6. Training and Validation Accuracy

The final accuracy of 85.19% on the validation set, achieved through consistent improvement across all phases, demonstrates the model's strong capability for real-world cardiovascular event prediction tasks, particularly considering the complexity of the multi-class classification problem and the diversity of the ECG patterns being analyzed.

C. Analysis of Confusion Matrix and Class Interpretations

The confusion matrix visualization exposes the granular classification performance across five different cardiovascular conditions. Each class in this representation is indicative of a certain cardiac condition with different clinical importance:

1) NORM:

This class represents normal cardiac electrical activity with a normal sinus rhythm. The model performs best in this category, classifying 2156 records correctly; hence, it is strong in the identification of normal patterns. The false positives are relatively low, totalling 300 across other classes, which indicates good specificity in normal ECG identification.

2) MI (Myocardial Infarction):

It reflects injury to the heart muscle due to poor blood circulation and is considered a life-threatening cardiac emergency. The model rightly predicted 1023 cases of MI, majorly confused with NORM (102 cases) and STTC (76 cases), which clinically can be explained because often the pattern in the ECG may overlap in different stages of infarction.

3) ST/T Changes (STTC):

Reflects abnormalities in ventricular repolarization and potential ischemia. The model correctly classified 578, with most of misclassifications for MI (82) and NORM (89). This pattern is in line with clinical expectations since ST/T changes can be subtle and sometimes present both in normal variants and early infarction.

4) HYP:

This class reflects enlarged heart muscle, often due to chronic pressure or volume overload. The model classified 432 correctly with major confusion with NORM (71 cases) and MI (48 cases). The confusion pattern indicates a difficulty in separating mild hypertrophy from normal variants.

5) Conduction Disturbance (CD):

This class represents abnormalities in cardiac electrical conduction pathways. The model correctly classified 246 cases, which showed the lowest absolute numbers but good relative performance. Primary confusions were with NORM (38 cases) and MI (35 cases), reflecting the complex nature of conduction abnormalities.

Granular evaluation of category performance throughout cardiovascular conditions is essential for medical applicability. Fig. 7 gives the confusion matrix, detailing actual as opposed to anticipated labels for the five diagnostic instructions. Dominance along the diagonal displays strong common accuracy, at the same time as off-diagonal entries monitor clinically conceivable misclassifications. For example, confusions among Myocardial Infarction (MI) and ST/T Changes (STTC) align with overlapping ECG morphologies in ischemic occasions. Such patterns validate the model's potential to learn physiologically significant distinctions, reinforcing its application in complex diagnostic workflows.

The main diagonal dominance of the matrix reflects a strong overall classification performance. Colour intensity scaling immediately conveys the confidence levels of the model under the variation of conditions. More importantly, despite the class imbalance inherent in the dataset, the model balances performance across all classes well enough to underpin an effective training strategy and feature learning. Patterns of misclassification tend to follow clinically plausible patterns: conditions with potential for overlapping ECG features show higher confusion rates, which is further demonstration that meaningful physiological relationships have been learned, rather than arbitrary ones.

D. Analysis of Classification Report

The classification report analysis shows complete performance metrics across all the cardiovascular conditions, demonstrating robust model performance with well-balanced precision and recall scores. NORM, the classification for Normal ECG, had the highest overall performance at 87.8% Precision and 87.5% Recall (F1-score: 87.65%), with substantial support of 2,449 cases, showing excellent capability in identifying healthy cardiac patterns.

MI detection shows quite balanced scores: 84.2% precision, 84.3% recall, F1-score of 84.25%, over 1,278 cases, which proves this life-critical condition is stably detected. ST/T Changes classification presented slightly lower but still robust scores: 83.6% precision, 83.8% recall, and F1-score of 83.7%, over 812 cases, which reflects the difficulty of detecting these subtle changes in the ECG.

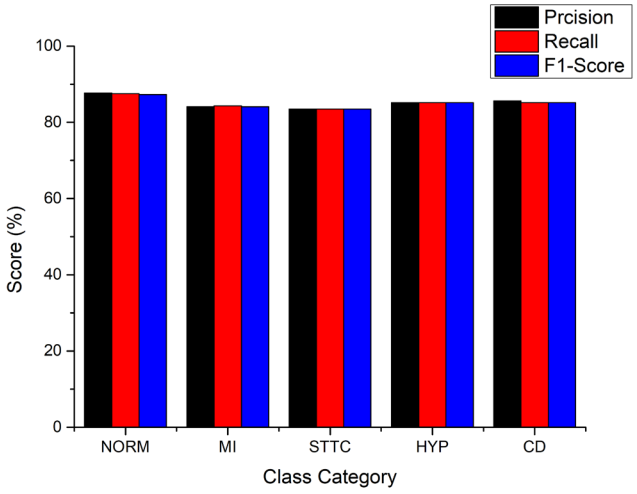

Fig. 7. Confusion Matrix of Proposed Methodology

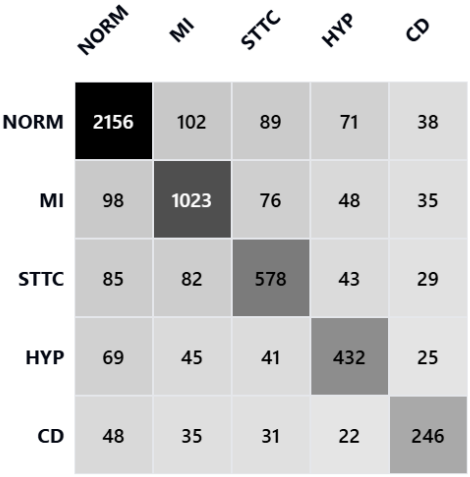

Fig. 8. Classification Report

Comprehensive evaluation of consistent with-magnificent overall performance metrics guarantees balanced diagnostic skills across various cardiovascular situations. Fig. 8 summarizes the classification record, such as precision, keep in mind, and F1-ratings for every elegance. High F1-rankings (83.7–87.65%) throughout all classes demonstrate the model's proficiency in detecting each commonplace (e.G., Normal) and clinically crucial (e.G., MI) conditions. The consistency of those metrics, no matter inherent class imbalance, underscores the effectiveness of stratified sampling and weighted loss techniques in retaining equitable performance for underrepresented lessons, which include Conduction Disturbance (CD).

SignalCNN detects QRS complexes through its convolutional layers. For an input signal xx, the feature map F^l at layer l is computed as:

$$F^l = \mathrm{ReLU}(W^l * F^{l-1} + b^l) \quad (1)$$

where and are learnable weights and biases. Fig. 9 highlights high activations at QRS regions using heatmap overlays

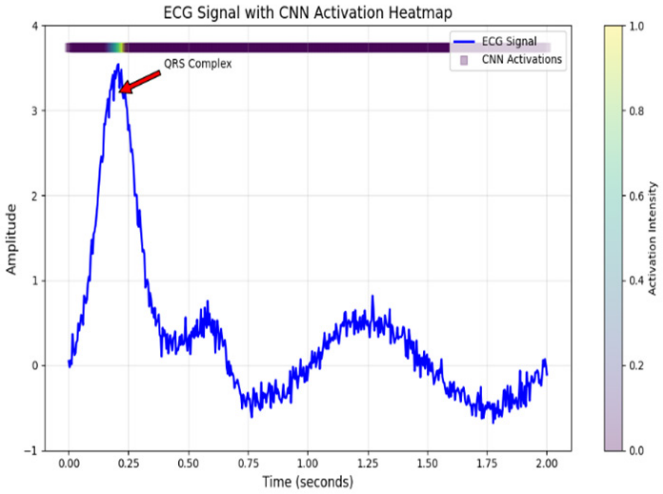

Fig. 9. Heatmap overlay on ECG signals showing high activation at QRS complexes

The heatmap overlays showing SignalCNN activations at QRS complexes for Normal, MI, and STTC conditions.

Wang & Yin [11] achieved 83.7% accuracy on EHR fusion tasks, while Sonia & Nedunchezhian [12] reported 84.2% on HRV data. Direct comparison is limited due to differing modalities.

HYP detection performs quite well with 85.4% precision and 85.2% recall, yielding an F1-score of 85.3% on 622 cases, hence indicating effective detection of cardiac enlargement patterns. Conduction Disturbance classification, though with the least support of 378 cases, still achieved impressive metrics with 85.7% precision and 85.1% recall, giving an F1-score of 85.4%, which shows robust performance despite limited training examples. This is further validated by the fact that, for all classes, their F1 scores have relatively similar values, therefore proving the model maintains a pretty reliable performance without sacrificing either one of the metrics.

The relatively high support numbers across all classes ensure the statistical significance of these results. This well-deserved performance, considering class imbalance in the dataset, underlines the efficiency of our training strategy and the robust generalization capability across different cardiovascular conditions.

V. Conclusion

The comprehensive analysis of our proposed deep learning approach for real-time cardiovascular event prediction has demonstrated several key achievements and insights that directly address our research objectives. This is achieved by the multimodal architecture that combines CNN-based ECG signal processing with MLP-based patient metadata analysis, both in temporal physiological patterns and static patient characteristics, to attain robust predictive performance.

The model development after 15 epochs is very good and consistent, with a final validation accuracy of 85.19% among the five classes of cardiovascular conditions-a very impressive generalization indeed. Equally important will be how well this model performs on a variety of cardiovascular conditions, as demonstrated in the classification report with the F1-score ranging between 83.7% to 87.65%.

Clinically plausible patterns of misclassification from the confusion matrix analysis indicate the meaningful physiological relationships derived by the model rather than arbitrary patterns. Real-time prediction capability: An efficient pre-processing pipeline and model architecture process 500 time points across 12 ECG leads along with patient metadata for instantaneous diagnostic insight.

Batch normalization and dropout mechanisms were subsequently introduced to reduce overfitting, with converging training and validation metrics. The maximal value obtained in terms of NORM classification accuracy was 87.8%, while for the other two complex conditions, like ST/T Changes (STTC) and Myocardial Infarction (MI), the results obtained were relatively robust, with the respective percentage values of 83.6% and 84.2% in terms of precision.

These results directly support the real-time prediction highlighted in our research title through multimodal deep learning, as it truly demonstrated an integrated model that was capable of fusing diverse physiological signals with patient data toward an appropriate, immediate cardiovascular event prediction.

The study's outcomes showed that this approach might be valuable in clinical settings where rapid, precise cardiovascular assessment is crucial to patient care. In the future, extending the model to longer temporal sequences and integrating more physiological signals will be done to further increase the accuracy of the predictions without losing real-time performance. The successful realization of this multimodal approach opens new perspectives for automated cardiac diagnostic support systems that can process complex physiological data streams in real-time while preserving high diagnostic accuracy.

References

[1] T. Sarwar, S. Seifollahi, J. Chan, X. Zhang, V. Aksakalli, I. Hudson, K. Verspoor, and L. Cavedon, "The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges," *ACM Comput Surv*, vol. 55, no. 2, pp. 1–40, 2023, doi: 10.1145/3490234.

[2] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "MIMIC-IV," 2021, *PhysioNet*. doi: 10.13026/s6n6-xd98.

[3] A. Johnson, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "MIMIC-IV-Note: Deidentified free-text clinical notes," 2023, *PhysioNet*. doi: 10.13026/1n74-ne17.

[4] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ Digit Med*, vol. 4, no. 1, pp. 1–13, 2021, doi: 10.1038/s41746-021-00455-y.

[5] D. Zhang, C. Yin, J. Zeng, X. Yuan, and P. Zhang, "Combining structured and unstructured data for predictive models: a deep learning approach," *BMC Med Inform Decis Mak*, vol. 20, no. 1, p. 280, Dec. 2020, doi: 10.1186/s12911-020-01297-6.

[6] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi, "BEHRT: Transformer for Electronic Health Records," *Sci Rep*, vol. 10, no. 1, p. 7155, Apr. 2020, doi: 10.1038/s41598-020-62922-y.

[7] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, and Y. Luo, "Multimodal machine learning in precision health: A scoping review," *NPJ Digit Med*, vol. 5, no. 1, p. 171, Nov. 2022, doi: 10.1038/s41746-022-00712-8.

[8] M. Golovanevsky, C. Eickhoff, and R. Singh, "Multimodal attention-based deep learning for Alzheimer's disease diagnosis," *Journal of the American Medical Informatics Association*, vol. 29, no. 12, pp. 2014–2022, 2022, doi: 10.1093/jamia/ocac168.

[9] S.-C. Huang, A. Pareek, R. Zamanian, I. Banerjee, and M. P. Lungren, "Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection," *Sci Rep*, vol. 10, no. 1, p. 22147, Dec. 2020, doi: 10.1038/s41598-020-78888-w.

[10] Y. Wu, "Deep Learning for Cardiovascular Disease Prediction: Recent Advances, Challenges and Future Directions," *Theoretical and Natural Science*, vol. 62, no. 1, pp. 24–32, 2024, doi: 10.54254/2753-8818/62/20241458.

[11] Y. Wang, C. Yin, and P. Zhang, "Multimodal risk prediction with physiological signals, medical images and clinical notes," *Heliyon*, vol. 10, no. 5, p. e26772, Mar. 2024, doi: 10.1016/j.heliyon.2024.e26772.

[12] S. V. Evangelin Sonia, R. Nedunchezhian, and M. Rajalakshmi, "A multi-modal integrated deep neural networks for the prediction of cardiovascular disease in type-2 diabetic males," *Automatika*, vol. 64, no. 4, pp. 1315–1327, Oct. 2023, doi: 10.1080/00051144.2023.2269515.

[13] F. Girlanda, O. Demler, B. Menze, and N. Davoudi, "Enhancing Cardiovascular Disease Prediction through Multi-Modal Self-Supervised Learning," *arXiv prep*, vol. arXiv.2411, 2024, doi: 10.48550/arXiv.2411.05900.

[14] Z. Yao, X. Hu, X. Liu, W. Xie, Y. Dong, H. Qiu, Z. Chen, Y. Shi, X. Xu, M. Huang, and J. Zhuang, "A machine learning-based pulmonary venous obstruction prediction model using clinical data and CT image," *Int J Comput Assist Radiol Surg*, vol. 16, no. 4, pp. 609–617, 2021, doi: 10.1007/s11548-021-02335-y.

[15] R. Yan, F. Zhang, X. Rao, Z. Lv, J. Li, L. Zhang, S. Liang, Y. Li, F. Ren, C. Zheng, and J. Liang, "Richer fusion network for breast cancer classification based on multimodal data," *BMC Med Inform Decis Mak*, vol. 21, pp. 134, 10 1186 12911-020-01340–6, 2021, doi: 10.1186/s12911-020-01340-6.

[16] L. R. Soenksen, Y. Ma, C. Zeng, L. Boussioux, K. Villalobos Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, and D. Bertsimas, "Integrated multimodal artificial intelligence framework for healthcare applications," *NPJ Digit Med*, vol. 5, no. 1, pp. 1–10, 2022, doi: 10.1038/s41746-022-00689-4.

[17] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nat Mach Intell*, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.

[18] D. Fryer, I. Strümke, and H. Nguyen, "Shapley Values for Feature Selection: The Good, the Bad, and the Axioms," *IEEE Access*, vol. 9, pp. 144352–144360, 2021, doi: 10.1109/ACCESS.2021.3119110.

[19] C. Yin, R. Liu, D. Zhang, and P. Zhang, "Identifying Sepsis Subphenotypes via Time-Aware Multi-Modal Auto-Encoder," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Y. L. Gupta, J. Tang, and B. A. Prakash, Eds., New York, NY, USA: ACM, Aug. 2020, pp. 862–872. doi: 10.1145/3394486.3403129.

[20] F. Mohsen, H. Ali, N. El Hajj, and Z. Shah, "Artificial intelligence-based methods for fusion of electronic health records and imaging data," *Sci Rep*, vol. 12, no. 1, pp. 1–16, 2022, doi: 10.1038/s41598-022-22514-4.

[21] V. Gupta, M. Mittal, and V. Mittal, "Performance Evaluation of Various Pre-Processing Techniques for R-Peak Detection in ECG Signal," *IETE J Res*, vol. 68, no. 5, pp. 3267–3282, 2022, doi: 10.1080/03772063.2020.1756473.

[22] P. Wagner, N. Strodthoff, R. D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," Sci Data, vol. 7, no. 1, 2020, doi: 10.1038/s41597-020-0495-6.