International Journal of Basic and Applied Sciences, 14 (1) (2025) 453-463

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/x1f6be61
Research paper

Multimodal contrastive learning for optimized integration of ECG signals and clinical reports for arrhythmia classification

Ragavan Veerarajan [1] *, Vanitha V. [2]

[1] *Research Scholar Sri Ramachandra Faculty of Engineering and Technology, Sri Ramachandra Institute of Higher Education Education and Research, Chennai-India*
[2] *Associate Professor, Sri Ramachandra Faculty of Engineering and Technology, Sri Ramachandra Institute of Higher Education Education and Research, Chennai-India*
**Corresponding author E-mail: ragavanveerarajan24@gmail.com*

Received: May 1, 2025, Accepted: May 21, 2025, Published: May 28, 2025

Abstract

Early and accurate detection of cardiac arrhythmias is essential for timely diagnosis and effective treatment. Traditional ECG-based classification models primarily rely on signal processing techniques, often neglecting the valuable contextual information contained in clinical notes. This study introduces a novel multimodal contrastive learning framework that integrates ECG signals and clinical notes for arrhythmia clas-sification. This approach is not previously explored in ECG-based diagnostics. Unlike existing methods that process ECG and clinical text separately, our framework employs contrastive learning for early feature alignment, ensuring meaningful representation of both modalities. Additionally, we propose an alternative alignment strategy that enables ECG-text integration without requiring direct patient-level mapping, addressing a major challenge in real-world healthcare applications. A neural network extracts temporal features from ECG signals, while a BERT-based model processes textual clinical data. Contrastive learning is employed to align multimodal representations, improving the model's ability to differentiate between normal and abnormal arrhythmias. Experimental results demonstrate 97.3% accuracy and an AUC-ROC of 0.98, significantly outperforming unimodal approaches. The proposed method enhances model generalization, robustness, and interpretability, making it suitable for real-world clinical applications. This study contributes to the development of AI-driven diagnostic tools that can integrate multimodal patient data, improving the reliability of automated arrhythmia detection in healthcare settings.

Keywords: *Contrastive Learning; Electrocardiogram; Multimodal Learning; Arrhythmia.*

1. Introduction

In the field of machine learning (ML) and artificial intelligence (AI), healthcare has been one area that has witnessed tremendous developments, especially in the use of diagnostic equipment like the electrocardiogram (ECG) to identify cardiovascular conditions [1]. Conventional ECG-based arrhythmia classification models, while successful in most instances, are usually limited by their use of a single modality of data like ECG signals [2]. These models tend to concentrate on raw signal analysis, e.g., identifying irregularities or patterns in heart rhythms, but do not incorporate the richer context that clinical reports and other textual information provide. Clinical notes, while important to understanding the medical history and subtleties of patient conditions, are still underutilized in AI-based diagnostics. New breakthroughs in multimodal learning [3] have showcased the capability of combining multiple sources of data, like time-series ECG signals and textual clinical documents, to enhance the accuracy and reliability of AI models. One such method that is attracting attention in machine learning is contrastive learning [4]. Contrastive learning, through training to compare and align representations from multiple modalities, allows the models to discover useful relationships between clinical texts and ECG signals, resulting in more context-based predictions [5]. Apart from improving diagnosis accuracy, it also assists with generalizing the models across varying datasets, eliminating the possibility of overfitting, and making it easier for models to adapt to new and unseen sources of data. Our research introduces a new multimodal contrastive learning method that combines both ECG signals and clinical text reports for arrhythmia classification. Through the process of training with contrastive learning, the model learns to map features between two modalities, creating a better understanding of the condition of the patient. This model hopes to create more precise, context-rich predictions that consider both the electrophysiological signals, and the wealth of descriptive text within clinical reports. This is a method that is expected to enhance model generalization across different datasets, improve diagnostic accuracy, and ultimately aid healthcare professionals in making more informed and accurate decisions.
In contrast to past work using either only ECG signals or only textual descriptions alone, we present a new multimodal contrastive learning paradigm that learns representations for both inputs jointly. Our method is different from previous multimodal ECG research in that it (1)

uses contrastive learning to enhance signal and text data alignment, (2) utilizes a common embedding space to enhance feature interaction, and (3) converts a multi-class dataset into a binary classification environment for enhanced generalization. These advances allow for improved arrhythmia classification performance while resolving multimodal fusion challenges in healthcare AI.

2. Related work

The application of machine learning (ML) in the field of medicine has greatly helped in the analysis of bio signals [6] and clinical text [7]. Specifically, arrhythmia classification from ECG signals and clinical report analysis have been areas of interest for research. However, these two modalities have often been studied in isolation, with little exploration of their potential to complement each other. Our research aims to address this gap by combining ECG signals and clinical reports through a multimodal contrastive learning approach [8], which could improve prediction accuracy and generalization. ECG-based arrhythmia classification has seen great progress in recent years, with the introduction of deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) that excel at capturing spatial and temporal patterns in ECG data [9]. These methods have demonstrated success in arrhythmia detection [10], focusing primarily on the ECG signals themselves. However, these models often overlook the additional context provided by clinical reports, which may include relevant patient history and other diagnostic insights that could help in more accurate decision-making. At the same time, the field of clinical text mining has witnessed substantial advancements, particularly with the advent of transformer-based models like BERT and Bio BERT [11]. These models have greatly improved the understanding of clinical narratives, extracting meaningful information from unstructured text data. Despite these improvements, most clinical text mining approaches focus solely on textual data, without considering the powerful diagnostic signals that ECG data can provide. Recognizing the value of combining multiple modalities, multimodal learning has emerged as an effective method to integrate diverse data types, such as images, signals, and text [12]. Previous work has shown the potential of multimodal models in healthcare tasks like disease diagnosis, where integrating medical images and clinical records improved prediction accuracy [13]. However, most existing approaches to multimodal learning tend to combine modalities late in the process, after each modality has been processed separately. This late-stage fusion can limit the model's ability to exploit the inherent relationships between modalities early on, which might reduce the model's effectiveness. One promising technique in this regard is contrastive learning, which has shown success in aligning similar samples while separating dissimilar ones, allowing for better representation learning. In healthcare, contrastive learning has been applied to integrate medical images with corresponding clinical reports, improving diagnostic accuracy [14]. While contrastive learning has demonstrated potential in multimodal settings, its application to combine ECG signals with clinical reports remains underexplored. In our research, we explore how contrastive learning can align these two distinct modalities such as ECG signals and clinical text, by learning joint representations that can be used for more robust arrhythmia detection. Despite the success of individual ECG classification and clinical text mining models, several gaps remain in the current state of the art. First, the limited integration of ECG and clinical text data in predictive models reduces the potential for context-aware, accurate predictions. Contrastive learning provides an opportunity to bridge this gap, allowing both modalities to be represented in a shared space that enhances predictions. Additionally, there is a challenge in improving generalization across datasets. Many existing models are prone to overfitting to specific data sources, which impacts their performance when applied to new data [15]. the use of early-stage fusion in multimodal learning has been relatively unexplored. By utilizing contrastive learning for early-stage fusion, it is possible to improve the integration of ECG and clinical text, allowing the model to learn richer, more accurate representations that combine both types of data seamlessly. By addressing these gaps, multimodal contrastive learning could significantly enhance the accuracy and generalization of arrhythmia classification models, providing more reliable and contextually informed predictions. This approach opens new avenues for improving healthcare decision-making, where both clinical context and diagnostic signals are considered holistically.

3. Methodology

The goal of our approach is to learn joint embeddings for ECG signals and clinical notes. The task involves learning a model that can map ECG features and corresponding clinical notes into a shared embedding space such that:

- Positive pairs (ECG + corresponding clinical notes) are close together in the embedding space.
- Negative pairs (ECG + mismatched clinical notes) are pushed further apart.

This representation learning problem is cast as a contrastive learning task, where the network is trained to minimize the contrastive loss between positive and negative pairs of ECG and clinical notes.

3.1. Data collection

For this study, we utilized two distinct datasets: the MIT-BIH Arrhythmia ECG Signals Dataset [16] and the MIMIC-III Clinical Notes Dataset [17], to develop a multimodal contrastive learning approach for arrhythmia classification. We selected three arrhythmia classes—Super ventricular Premature Beats (S), Normal Beats (N), and Unclassifiable Beats (Q), and excluded Fusion Beats and Ventricular Ectopic Beats because a binary classification framework simplifies the task, leading to improved performance, better feature learning, and more reliable decision-making. We have also used down down-sampling technique [18] to balance the MIT-BIH dataset; Table 1 describes the MIT-BIH Arrhythmia dataset population, and Figure 1 depicts actual data distribution and sampled data distribution taken for this study. Clinical notes related to arrhythmias were extracted from the MIMIC-III dataset using ECG-like text filters to provide contextual information for the ECG signals. Due to the absence of shared patient identifiers between MIT-BIH and MIMIC-III, we aligned ECG signals with clinical notes using arrhythmia status rather than direct patient correspondence. While this does not guarantee perfect patient-level alignment, previous research in multimodal learning has demonstrated that label-based alignment can still yield meaningful feature associations when direct mapping is unavailable [19]. This approach enables the model to learn shared representations of ECG and clinical text, enhancing its ability to generalize across real-world clinical scenarios. Future work will explore incorporating additional metadata, e.g., demographic patterns, signal characteristics, to refine the alignment strategy. The clinical text was pre-processed and paired with the corresponding ECG signals for contrastive learning. Binary labels were assigned to the ECG signals: 0 (Normal cases) for Normal Beats (N), and 1 (Abnormal cases) for Supraventricular Premature Beats (S) and Unclassifiable Beats (Q). Unclassifiable Beats (Q) were categorized as abnormal based on clinical significance, as they often indicate irregular cardiac activity requiring further evaluation. These signals were included in the negative pair for contrastive learning to enhance model robustness. This approach integrates both ECG signal data and clinical notes to improve arrhythmia classification by leveraging complementary information from both modalities

a) Original ECG Signal Population Distribution.
b) Down-Sampled ECG Signal Population Distribution

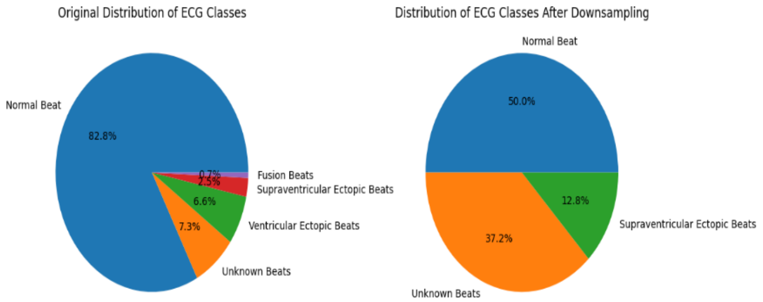

Fig. 1: ECG Data Population Analysis Before and After Down-Sampling.

Table 1: Summary of Arrhythmia Types, Sampling Strategy, and Binary Labels of MIT BIH Dataset

Arrhythmia type	Actual Samples	Consider-ation	Down-sam-pled Samples	Binary label	Remarks
Normal Beat(N)	72470	Included	8654	0	Normal beats are included to serve as a reference for distinguishing abnormal arrhythmias.
Supraventricular Ec-topic Beats (S)	2223	Included	2223	1	These beats represent abnormal heart rhythms from above the ven-tricles and are included to detect supraventricular arrhythmias.
Ventricular Ectopic Beats (V)	5788	Excluded	-	-	Excluded likely to focus on supraventricular arrhythmias instead of ventricular ones.
Fusion Beats (F)	641	Excluded	-	-	Fusion beats are rare and difficult to classify, often leading to poor model generalization.
Unknown Beats (Q)	6431	Included	6431	1	Includes unclassified beats, which are clinically relevant for detect-ing artificial pacemaker signals or unknown arrhythmias.

We utilized the Google BigQuery platform to extract clinical notes from the MIMIC-III dataset. applying text filters for ECG, Arrhythmia, and relevant ICD-9 codes [19] as outlined in Table 2. These codes were employed to enhance the filtering process. The extracted clinical notes were then mapped to the MIT-BIH Arrhythmia dataset, focusing on Normal Beats (N), Supraventricular Ectopic Beats (S), and Unknown Beats (Q). The multi-class dataset was subsequently transformed into a binary classification dataset, distinguishing between normal and abnormal cases. This refined dataset was used to train a contrastive learning-based multimodal ECG classifier and a clinical note generator. Table 3 presents sample clinical notes corresponding to both normal and abnormal arrhythmia cases. Before accessing the MIMIC III dataset, we completed the training provided by PhysioNet and obtained the necessary permissions to work with their data. This ensured compliance with ethical and regulatory guidelines for handling sensitive medical information.

Table 2: Mapping of ICD-9 Codes to MIMIC-III Codes for Arrhythmia

ICD-9 Code	MIMIC-III Code	Description
427.1	4271	Paroxysmal ventricular tachycardia (VT)
427.2	4272	Ventricular fibrillation and flutter
427.31	42731	Atrial fibrillation (AFib)
427.32	42732	Atrial flutter
427.41	42741	Ventricular fibrillation
427.42	42742	Ventricular flutter
427.5	4275	Cardiac arrest
427.60	42760	Premature beats, unspecified
427.61	42761	Supraventricular premature beats
427.69	42769	Other premature beats

Table 3: Sample Clinical Notes for Normal and Abnormal Arrhythmia Cases

Arrhythmia Case	Clinical Note Summary
Normal	• Sinus rhythm • Left anterior fascicular block and intraventricular conduction delay • Inferior/lateral ST-T changes • Since last ECG, the ST-T wave changes may be less
Abnormal	• Atrial fibrillation with rapid ventricular response • Left axis deviation - left anterior fascicular block • Consider incomplete right bundle branch block • Ant/septal+lateral ST-T changes may be due to myocardial ischemia • Since last ECG, rate slightly slower

3.2. Data preprocessing

To ensure the data is in a suitable format for deep learning models, several preprocessing steps were performed on both the ECG signals and clinical reports

3.2.1. ECG signal preprocessing

The ECG encoder is a two-hidden-layer neural network designed to extract meaningful features from the ECG signal. It processes the sequential data and captures temporal patterns, enabling effective feature representation for classification and contrastive learning. Figure 2 depicts the ECG Feature extraction process.

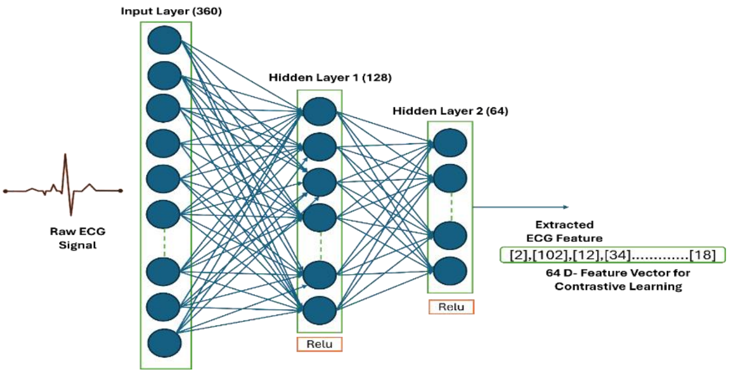

Fig. 2: ECG Encoder Architecture.

Depicts how the encoded ECG features are transformed into a 64-dimensional vector for downstream tasks.

3.2.2. Clinical notes processing

The clinical notes are processed using the BERT tokenizer [20], which breaks the raw text into subword tokens. These tokenized sequences are then fed into the BERT model to generate contextualized embeddings. The tokenizer produces two key outputs: "input_ids", which are numerical representations of the tokens, and "attention_mask", which differentiates real tokens from padding using 0s and 1s. This process is illustrated in Figure 3.

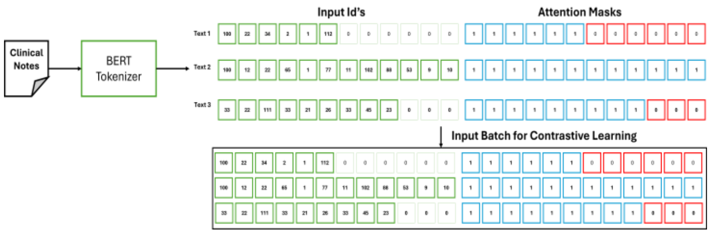

Fig. 3: BERT Tokenization.

The embeddings from the clinical notes and the processed ECG features are concatenated together. This combined feature set represents both the clinical context from the clinical notes and the raw physiological data from the ECG signals. Along with the tokenized text, an attention mask is created. This mask informs the BERT model about which tokens should be ignored and which are actual content. The attention mask helps BERT focus on relevant tokens during processing. These concatenated features are fed into a classification to predict the class label whether the ECG indicates arrhythmia or not. The model essentially learns to contrast the information from both modalities, ECG and clinical text, for classification.

3.2.3. Generating positive and negative pairs for contrastive learning

(a) Positive Pairs: A positive pair is formed by combining the ECG signal of a patient with its corresponding clinical note. This pair should be semantically similar, representing the same patient.
(b) Negative Pairs: A negative pair is formed by pairing an ECG signal with a mismatched clinical note from a different patient. The goal is to create a representation space where similar pairs are closer, and dissimilar pairs are farther apart.
The final pre-processed data consists of ECG features as tensors, tokenized clinical notes, and corresponding attention masks for each note. Positive and negative pairs are ready for contrastive loss computation.

3.3. Research design

We designed the study based on a multimodal contrastive learning classification framework, where the ECG signals and clinical reports are used as input features to two separate encoders. The encoders learn feature representations for each modality, which are then aligned in a shared embedding space using a contrastive loss function. The overall model design is shown in Figure 4.

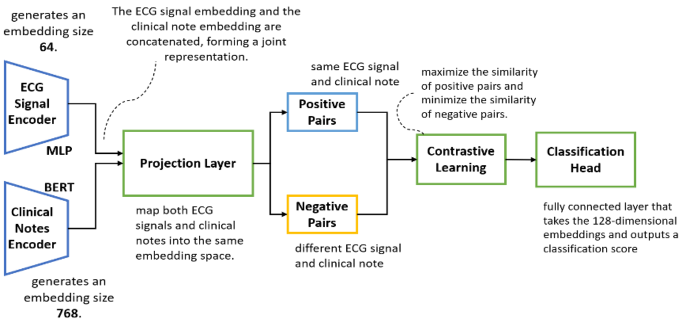

Fig. 4: Multimodal Contrastive Learning Model for Arrhythmia Classification.

This architecture combines ECG signal processing with BERT-based clinical note understanding to perform both contrastive learning and binary classification (Normal vs. Abnormal).

3.3.1. ECG encoder

Neural Network processes the ECG signals and extracts meaningful features. The encoder is designed to handle sequential data and capture temporal patterns in the ECG signal. The ECG encoder is a processes ECG signals and maps them to a lower-dimensional latent space. X be the input ECG feature vector which consist of 360 samples (raw ecg data), where each dimension corresponds to a different extracted ECG feature. The first transformation is a linear mapping from 360 dimensions to 128 dimensions

$$h_1 = W_1 X + b_1$$

Where,
$W_1 \in R^{128 X 360}$ is the weight matrix for the first layer
$b_1 \in \mathbb{R}^{128}$ is the bias vector.
$h_1 \in \mathbb{R}^{128}$ is the intermediate hidden representation before activation.
Applying the ReLU activation function

$$h_1' = \mathrm{ReLU}\,(h_1)$$

Where,

$$\mathrm{ReLU}(z) = \max(0, z)$$

This ensures that all negative values are set to zero, introducing non-linearity to the network.
The second transformation further reduces the dimensionality from 128 to 64

$$z = W_2 h_1' + b_2$$

Where:
$W_2 \in \mathbb{R}^{64 X 128}$ is the weight matrix for the second layer.
$b_2 \in \mathbb{R}^{64}$ is the bias vector.
$z \in \mathbb{R}^{64}$ is the final ECG embedding.
The output of the ECG encoder, z is a 64-dimensional vector that serves as a compressed representation of the ECG signal. This representation is then used in further tasks such as paired with clinical note embeddings, contrastive learning, and classification.

3.3.2. Clinical notes encoder

A pre-trained transformer-based model, BERT, processes the clinical reports. The model is fine-tuned on the clinical text data to generate contextual embeddings that capture the semantics and relationships in the reports. A clinical note is a sequence of words (tokens) represented as

$$C = \{w_1, w_2, \dots \dots w_T\}$$

T is the number of tokens in the input sequence.
Each word w_i is first converted into an integer token ID using a vocabulary.
The input representation consists of three embeddings

1) Token Embeddings: Each word w_i is mapped to an embedding vector $e_i \in \mathbb{R}^d$ using a pre-trained word embedding matrix.
2) Position Embeddings: Since transformers don't have recurrence, positional embeddings $p_i \in \mathbb{R}^d$are added to retain word order information.
3) Segment Embeddings: If multiple sentences exist, segment $s_i \in \mathbb{R}^d$differentiate them.

The final input embedding for each token is

$x_i = e_i + p_i + s_i$ is the input vector for each token.

BERT processes the input sequence using multiple self-attention layers and feed-forward networks.
Self-Attention Mechanism: Each token attends to all other tokens using the scaled dot-product attention

$$\mathrm{Attention}\,(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where,
$Q = X\,W_Q, K = X\,W_K, V = X\,W_V$ are query, key, and value matrices,
$X \in \mathbb{R}^{T X d}$ is the matrix of token embeddings,
$W_Q, W_K, W_V \in \mathbb{R}^{d X d_k}$ relearnable weight matrices,
d_k is the dimension of keys
The multi-head attention mechanism is

$$\mathrm{MultiHead}\,(X) = \mathrm{Concat}\,(\mathrm{head}_1, \mathrm{head}_2, \dots \dots \mathrm{head}_h)W_O$$

Where each head is computed using independent Q, K, and V projections.
The final layer of BERT produces hidden representations for each token

$H = \{h_1, h_2, \dots\dots, h_T\}$

Where
$H \in \mathbb{R}^{T \times d}$ is the set of contextualized embeddings for all tokens.
Each $h_i \in \mathbb{R}^d$ captures information from surrounding words.
To obtain a fixed-length representation from BERT, we extract the CLS token embedding

$$Z_{BERT} = h_{CLS}$$

Where:
$h_{CLS} \in \mathbb{R}^d$ is the hidden state of the special [CLS] token.
This vector represents the entire clinical note.
Alternatively, we can average all token embeddings

$$Z_{BERT} = \frac{1}{T}\sum_{i=1}^{T} h_i$$

This sentence embedding is then used for downstream tasks like contrastive learning and classification. The clinical note embedding, $Z_{BERT} \in \mathbb{R}^d$ is a dense vector representation of the input note, capturing its meaning and context. BERT tokenization involves two key outputs

1) Input id's X_{Token} : Numerical Representation of Tokens
2) Attention id's M_{attn} : Binary mask to distinguish actual tokens from padding

Tokenization function T as a mapping from text to tokenized outputs

$$T : \text{Sentence} \rightarrow (X_{Token}, M_{attn})$$

For a given clinical note $S = (w_1, w_2, \dots., w_n)$

a) Tokenize the text into subwords using Word Piece tokenization

$$T(S) = (t_1, t_{2,} \dots., t_L)$$

Where t_i are subword tokens, and L is the number of tokens after tokenization.

b) Convert tokens to numerical IDs using the BERT vocabulary V

$$X_{Token} = (v_1, v_2, \dots\dots, v_L)$$

Where v_iIs the index of the token t_i.

c) Generate attention mask M_{attn}

- If $t_i It$ is a real token, $M_{attn,i} = 1$
- If t_i is padding, $M_{attn,i} = 0$

$$M_{attn} = (m_1, m_2, \dots., m_L), m_i \in \{0,1\}$$

3.3.3. Projection layer

The concatenated embeddings from the ECG and clinical note encoders are projected into a common 128-dimensional space to align the two modalities. ECG embeddings $z \in \mathbb{R}^{64}$ and Clinical note embeddings $Z_{BERT} \in \mathbb{R}^d$ has concatenated and represents as $Z_{concat} = [z;\ Z_{BERT}] \in \mathbb{R}^{(64+d)}$ The concatenated embedding is passed through a projection layer that maps it into a lower-dimensional contrastive space $\mathbb{R}^{128}$

$$Z_{proj} = W_{proj} Z_{concat} + b_{proj}$$

Where

- $W_{proj} \in \mathbb{R}^{128 \times (64+d)}$ is the learned projection matrix.
- $b_{proj} \in \mathbb{R}^{128}$ is the bias vector.
- $Z_{proj} \in \mathbb{R}^{128}$ is the final projected representation.

3.3.4. Contrastive learning

The embeddings generated from the ECG and clinical report encoders are passed into a contrastive loss function. This loss function maximizes the similarity between embeddings of matching pairs (i.e., ECG and its corresponding clinical report) while minimizing the similarity between embeddings of mismatched pairs (i.e., ECG and a non-corresponding clinical report).
Let, $z_i \in \mathbb{R}^{128}$ be the projected ECG embedding for the sample i
$z_i' \in \mathbb{R}^{128}$ be the projected BERT embedding for the same sample i
The goal is to maximize the similarity between (z_i, z_i') while minimizing similarity with embeddings of other samples.
We use cosine similarity between the projected embeddings

$$\text{sim}\ (z_i, z_i') = \frac{z_i . z_i'}{\|z_i\| . \|z_i'\|}$$

This measures how aligned the two vectors are in the shared space.

3.3.5. Contrastive loss function

The contrastive loss function used in our study ensures that the model learns a shared embedding space where related ECG signals and clinical reports are close to each other, while dissimilar signals are pushed apart. This helps improve the overall learning process and enables the model to make more accurate predictions when new, unseen data is introduced. we use the Normalized Temperature-Scaled Cross-Entropy Loss (NT-Xent) [21]. The NT-Xent loss is used to optimize the embedding space. It encourages the model to bring similar (positive) pairs closer and push dissimilar (negative) pairs apart. The loss is computed using the cosine similarity between the embeddings of positive and negative pairs, with a temperature scaling factor to control the sharpness of the similarity measure.

The NT-Xent is defined as

$$l_i = -\log \frac{\exp(\mathrm{sim}\left(\frac{z_i \cdot z_i'}{\tau}\right))}{\sum_{j \neq i} \exp(\mathrm{sim}\left(\frac{z_i \cdot z_j'}{\tau}\right))}$$

Where τ is a temperature scaling parameter (typically 0.05)
The denominator includes embeddings from another sample $j \neq i$, enforcing separation.
The final contrastive loss across all samples in a batch of size N

$$L = \frac{1}{N} \sum_{i=1}^{N} l_i$$

3.3.6. Classification layer

The contrastive loss helps in learning discriminative representations [22], but for actual diagnosis or downstream tasks, we need classification. the concatenated embedding $[z \parallel z_{BERT}]$ are passed to a fully connected classification head. The model is trained using cross-entropy loss and Adam optimizers [23] are used to update model parameters. The gradients from the classification loss are propagated back to update the ECG and BERT feature extractors.

4. Experimental setup

The proposed framework was implemented using Python with PyTorch. The ECG encoder and BERT model were trained using an AMD Radeon 530 with 16GB VRAM. The dataset was preprocessed using standard normalization techniques, and all experiments were conducted using PyTorch 2.0 with the Hugging Face Transformers library [24] for BERT embeddings.

4.1. Training and optimization

The model was trained using the Adam optimizer with an adaptive learning rate. The training process involved minimizing the contrastive loss with multiple epochs, with periodic evaluation on the validation set to avoid overfitting. Early stopping was applied to prevent the model from training for too many epochs and to ensure that the best model was selected based on validation performance.

4.2. Evaluation metrics

To evaluate the model's performance, we used several metrics, including accuracy, precision, recall, F1 score, and AUC-ROC [25]. These metrics were used to assess both the classification performance of the arrhythmia detector and the overall effectiveness of the multimodal contrastive learning approach.

4.3. Model evaluation and validation

To evaluate the effectiveness of the multimodal contrastive learning approach, we conducted several experiments. We compared the performance of our proposed multimodal model with traditional models that use only ECG signals or clinical reports individually. This comparison helped demonstrate the added value of combining both modalities using contrastive learning. To assess the model's generalizability, we validated the performance on multiple datasets, ensuring that the model can effectively handle new and unseen data sources.

5. Results

The proposed multimodal framework integrating ECG signals and clinical text using contrastive learning demonstrated superior performance compared to unimodal models. The results of our experiments indicate that the combined feature representations significantly improved classification accuracy, demonstrating the effectiveness of contrastive learning in aligning information from heterogeneous medical data sources.

5.1. Classification performance

Figure 5 explains the confusion matrix of our classifier, and Table 4 presents the classification performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, also compared with existing methods. The results show that our proposed method consistently outperforms baseline models that utilize only ECG signals or clinical text separately.

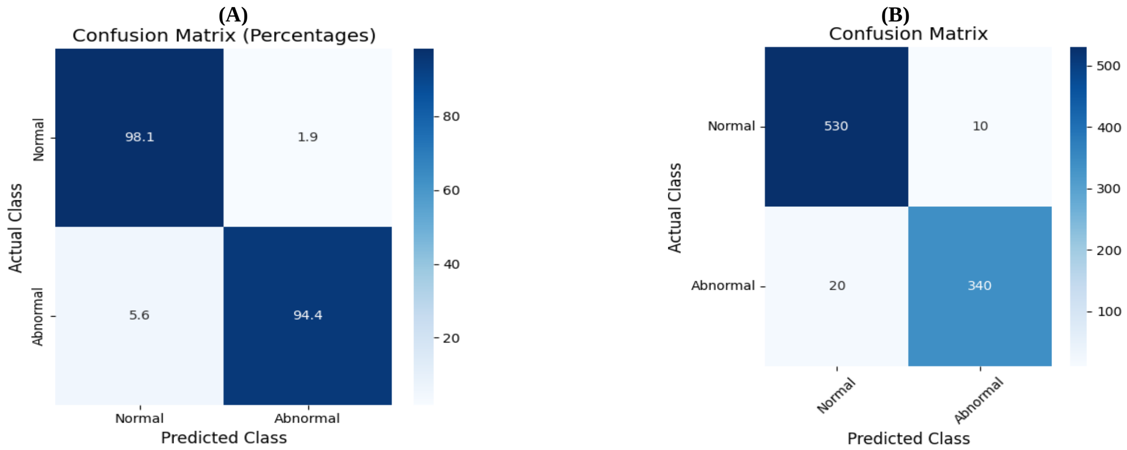

Fig. 5: Multimodal Contrastive Learning Model Performance Results

5(a) Illustrate TP, FP, TN, FN percentage values, 5(b) illustrate TP, FP, TN, FN values

Table 4: Multimodal Contrastive Learning Model Performance Comparison with Different Model Variants

Model Variant	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
ECG-only Model (CNN Model)	91.3	91.3	91.3	91.3	91.3
ECG-only Model (LSTM- Model)	97.5	87.9	86.4	87.1	89.2
Clinical Text-Only Model (BERT-Based Text Model)	84.2	84.2	84.2	84.2	84.2
Without Contrastive Learning (CNN with BERT)	89.7	89.7	89.7	89.7	89.7
Without Shared Projection Space (CNN with BERT)	87.8	87.8	87.8	87.8	87.8
Proposed Model (Multi modal Contrastive Learning)	97.8	97.1	94.4	95.8	98.2

The proposed contrastive learning approach got 97.8% in accuracy, 95.8% in F1 Score, and 98.2% AUC-ROC, this result highlighting the benefits of combining ECG and textual features. Figure 6 indicates the various model performances.

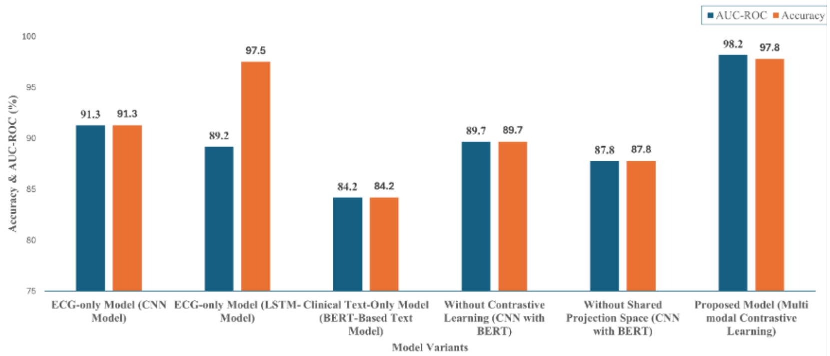

Fig. 6: Different Arrhythmia Model Performances.

The proposed model performance has better results comparing to the other model variants, we have achieved the result by running the proposed model with 50 epochs and its training and testing results are shown in the Figure7.

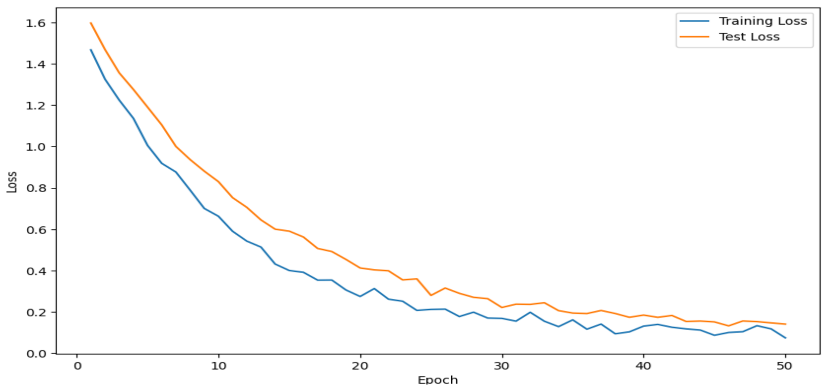

Fig. 7: Loss Plot During Training and Validation.

5.2. Interpretability enhancements

To promote clinical adoption of the AI-based ECG diagnostic model, enhancing interpretability is crucial. This section outlines two core approaches using attention score analysis and feature importance explanation (SHAP values). These techniques help clinicians understand why the model made a specific decision

5.2.1. Attention-based interpretation of clinical notes

For transformer-based models, interpretability is enhanced by extracting and presenting attention weights that highlight important clinical tokens from the input text. the clinical text is tokenized and passed through the transformer model, then Attention weights are extracted from the last layer's attention matrices. Finally, the average attention across all heads and normalized per token gives a Normalized Attention Score. Table 5 indicates the sample clinical notes' attention weights. The model focuses primarily on chest pain and irregular heartbeat, suggesting these were key indicators in its decision-making process

Table 5: Attention Weights from Clinical Notes

Token	Raw Attention Score	Normalized Attention Score
chest pain	2.5	0.32
irregular heartbeat	2.1	0.27
shortness of breath	1.8	0.23
beta blockers	0.9	0.12
stable vitals	0.5	0.06
Total	7.8	1

5.2.2. SHAP-based interpretation of ECG feature contribution

For structured or signal-based ECG features, SHAP (Shapley Additive Explanations) helps explain the contribution of each feature to the final prediction. A SHAP explainer is used depending on the model type, then SHAP values quantify the impact of each input feature on the model's prediction, finally, the base value is adjusted by the SHAP sum to yield the model's predicted probability. Table 6 has the patient's SHAP values for ECG Features.

Table 6: SHAP Values for ECG Features

ECG Feature	SHAP Value	Impact on Prediction
RR interval	0.19	Increased arrhythmia likelihood
PR interval	0.12	Moderate positive contribution
QRS amplitude	-0.04	Slightly reduced arrhythmia likelihood
T wave abnormality	0.18	Strong positive influence
Heart rate variability	0.06	Minor positive contribution
Base value	0.5	-
Prediction	0.91	(Base + SHAP sum = 0.91)

The model identified “T wave abnormality” and “RR interval” as key drivers of the positive diagnosis, closely aligning with clinical intuition.

5.2.3. Clinical relevance and trust

Based on the Table 5 and 6 interpretable values, clinicians can Validate model reasoning against known pathophysiological markers, build trust in model decisions through transparent contribution breakdowns, and engage in human-AI collaboration using model outputs to support, not replace, decision-making

6. Discussion

The results confirm that our proposed multimodal contrastive learning approach significantly enhances arrhythmia classification performance by leveraging complementary information from ECG and textual data. The study highlights the following key findings:

- Contrastive learning improves feature alignment, leading to better generalization.
- Multimodal integration enhances classification accuracy, surpassing unimodal models.
- The model remains computationally efficient and robust, making it suitable for real-world clinical applications.

Despite its strong performance, the proposed framework has some limitations. The contrastive learning framework requires significantly longer training times due to the additional computational burden of learning joint representations. Multi-GPU frameworks like Distributed Data Parallel (DDP) optimize contrastive learning by parallelizing batch processing, minimizing communication latency, and enabling efficient gradient aggregation, thus reducing wall-clock training time. Its effectiveness is also constrained by the availability of high-quality paired ECG and clinical text datasets, which limits the generalizability of the model. ECG-text alignment improves via metadata filtering (time and demographics), synthetic identifiers enforcing explicit linkages, and contrastive learning with hard negatives, minimizing spurious status-based correlations in multimodal clinical data. Our study used arrhythmia status as a proxy for ECG-text alignment due to the lack of patient identifiers in MIMIC-III. While this approach achieved high accuracy, it may introduce noise if clinical notes describe conditions not fully reflected in the paired ECG. Potential improvements include leveraging temporal metadata or probabilistic record linkage. Future work could test these methods in datasets with explicit pairings. While contrastive learning enhances feature alignment, the interpretability of deep multimodal models remains an area for further exploration, requiring additional efforts to make the decision-making process more transparent and explainable. These findings suggest that multimodal deep learning approaches hold great promise for advancing AI-driven cardiac diagnostics. Future work will focus on extending the model to handle additional modalities such as patient demographics and historical medical records to further improve predictive performance and validating the model on additional datasets like PTB-XL and expanding arrhythmia categories to enhance generalizability and robustness across diverse clinical scenarios.

7. Conclusion

This study proposed a multimodal contrastive learning framework integrating ECG signals and clinical text for arrhythmia classification. The experimental results demonstrated that the fusion of multimodal features significantly improves classification accuracy compared to unimodal models. The introduction of contrastive learning effectively aligned heterogeneous representations, leading to enhanced generalization and robustness. Additionally, the model exhibited computational efficiency, making it suitable for real-world clinical applications. However, challenges such as training time, data availability, hardware constraints, and model interpretability remain areas for future research. Future work will focus on expanding the dataset, improving model interpretability, and optimizing the framework for deployment in resource-limited environments. The findings from this study highlight the potential of AI-driven multimodal learning in advancing cardiac diagnostics and improving patient outcomes

Acknowledgement

All authors are grateful to the Department of Artificial Intelligence and Machine Learning, Sri Ramachandra Faculty of Engineering and Technology, for providing Computing Resources to conduct this Research.

Conflict of interest

The authors have no conflicts of interest, either financial or commercial.

Ethics Approval

Ethical approval is not required for this review paper, as it does not involve direct experimentation with human or animal subjects.

Funding

The authors did not receive support from any organization for the submitted work.

References

[1] Serha Serhal H, Abdallah N, Marion JM, Chauvet P, Oueidat M, Humeau-Heurtier A. Overview on prediction, detection, and classification of atrial fibrillation using wavelets and AI on ECG. Comput Biol Med. 2022;142:105168. https://doi.org/10.1016/j.compbiomed.2021.105168.
[2] Chung CT, Lee S, King E, Liu T, Armoundas AA, Bazoukis G, Tse G. Clinical significance, challenges and limitations in using artificial intelligence for electrocardiography-based diagnosis. Int J Arrhythmia. 2022;23(1):24. https://doi.org/10.1186/s42444-022-00075-x.
[3] Ramachandram D, Taylor GW. Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Process Mag. 2017;34(6):96-108. https://doi.org/10.1109/MSP.2017.2738401.
[4] Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. Adv Neural Inf Process Syst. 2020;33:18661-73.
[5] Le D, Truong S, Brijesh P, Adjeroh DA, Le N. scl-st: Supervised contrastive learning with semantic transformations for multiple lead ecg arrhythmia classification. IEEE J Biomed Health Inform. 2023;27(6):2818-28. https://doi.org/10.1109/JBHI.2023.3246241.
[6] Dey N, Borra S, Ashour AS, Shi F, editors. Machine learning in bio-signal analysis and diagnostic imaging. Academic Press; 2018.
[7] Skeppstedt M, Kvist M, Nilsson GH, Dalianis H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. J Biomed Inform. 2014;49:148-58. https://doi.org/10.1016/j.jbi.2014.01.012.
[8] Yuan X, Lin Z, Kuen J, Zhang J, Wang Y, Maire M, et al. Multimodal contrastive training for visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 6995-7004. https://doi.org/10.1109/CVPR46437.2021.00692.
[9] Celin S, Vasanth K. ECG signal classification using various machine learning techniques. J Med Syst. 2018;42(12):241. https://doi.org/10.1007/s10916-018-1083-6.
[10] Awasthi S, Sachdeva N, Gupta Y, Anto AG, Asfahan S, Abbou R, et al. Identification and risk stratification of coronary disease by artificial intelligence-enabled ECG. EClinicalMedicine. 2023;65. https://doi.org/10.1016/j.eclinm.2023.102259.
[11] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40. https://doi.org/10.1093/bioinformatics/btz682.
[12] Zhang C, Yang Z, He X, Deng L. Multimodal intelligence: Representation learning, information fusion, and applications. IEEE J Sel Top Signal Process. 2020;14(3):478-93. https://doi.org/10.1109/JSTSP.2020.2987728.
[13] aleb A, Lippert C, Klein T, Nabi M. Multimodal self-supervised learning for medical image analysis. In: International Conference on Information Processing in Medical Imaging; 2021. p. 661-73. https://doi.org/10.1007/978-3-030-78191-0_51.
[14] Hsu TMH, Weng WH, Boag W, McDermott M, Szolovits P. Unsupervised multimodal representation learning across medical images and reports. arXiv:1811.08615. 2018.
[15] Houssein EH, Kilany M, Hassanien AE. ECG signals classification: a review. Int J Intell Eng Inform. 2017;5(4):376-96. https://doi.org/10.1504/IJIEI.2017.087944.
[16] Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag. 2001;20(3):45-50. https://doi.org/10.1109/51.932724.
[17] Johnson AE, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035. https://doi.org/10.1038/sdata.2016.35
[18] Mesin L. Heartbeat monitoring from adaptively down-sampled electrocardiogram. Comput Biol Med. 2017;84:217-25. https://doi.org/10.1016/j.compbiomed.2017.03.023.
[19] Huang J, Osorio C, Sy LW. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. Comput Methods Programs Biomed. 2019;177:141-53. https://doi.org/10.1016/j.cmpb.2019.05.024.
[20] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019. p. 4171-86.
[21] Sohn K. Improved deep metric learning with multi-class n-pair loss objective. Adv Neural Inf Process Syst. 2016;29.
[22] Le-Khac PH, Healy G, Smeaton AF. Contrastive representation learning: A framework and review. IEEE Access. 2020;8:193907-34. https://doi.org/10.1109/ACCESS.2020.3031549.

[23] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980. 2014.
[24] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; 2020. p. 38-45. https://doi.org/10.18653/v1/2020.emnlp-demos.6.
[25] Naidu G, Zuva T, Sibanda EM. A review of evaluation metrics in machine learning algorithms. In: Computer Science On-line Conference; 2023. p. 15-25. https://doi.org/10.1007/978-3-031-35314-7_2.