

A DUAL-STREAM DEEP TRANSFORMER NETWORK FOR ROBUST MULTI-LEAD ECG INTERPRETATION: CROSS-PATIENT ECG REPRESENTATION LEARNING FOR ARRHYTHMIA DETECTION USING A SELF-SUPERVISED CONTRASTIVE METHOD

Yeligeti Raju¹, R. Ramakrishna², Mrs.Mursubai Sandhya Rani³, B Saritha⁴, Veeramachaneni Dhanasree⁵

¹Professor, Department of Computer Science and Engineering (Data Science), Vignana Bharathi Institute of Technology, Aushapur, Hyderabad

²Professor, Department of Mathematics, Vidya Jyothi Institute of Technology, Aziznagar gate, CB Post, Hyderabad, Telangana, India 500075

³Assistant professor, Department of Computer Science and Engineering, School of Engineering, Anurag University, Hyderabad

⁴ Professor, Department of CSE, Sreenidhi university, Yamnampet, Hyderabad, Charishma Ambati, Software Engineer

⁵Assistant professor, Department of CSE-Cyber Security, Geethanjali College of Engineering and Technology, Keesara, Hyderabad

ABSTRACT

Despite its importance in the non-invasive diagnosis of cardiac arrhythmia, the ECG is still the major tool for routine clinical diagnosis which relies on large manual artifact-based interpretation of 12 simultaneous leads and the ability to resolve a detailed morphology from a short-term rhythm behaviour usually lasting no more than 5 seconds. The efforts for this process are labour intensive, inter-observer dependent and not suitable for the amount of data collected while subjects are in continuous ambulatory and wearable monitoring. Most of the models proposed for automatic arrhythmia detection exploit beat-level annotations and are fully fed supervisedly (intra-patient partitions); this translates into a high convincing factor for the clinical performance, since most of the beats in the test set are the same subject's every beat, appearing in both training and test data. These models can be transferred to different subjects, but since the subject-specific idiosyncrasy of waveforms is captured in the model, it fails to accurately model the pathology when the other patient comes into view. In this paper, we present a dual-stream deep transformer network with the aim of achieving robust insight on multiple leads as to improve the interpretation of ECG in most cases. Through proposed dual-stream deep transformer, this paper presents a strategy to learn a cross-patient representation using self-supervised contrastive training, with an ultimate goal of robust multi-lead ECG interpretation. The architecture features a two-stream processing paradigm – one stream of the architecture is called the morphology stream, where a one-dimensional (1D) convolutional tokenizer followed by a local-window transformer encoder captures the shape of P-QRS-T. The other stream is referred to as a rhythm stream in the architecture, where a global transformer encoder processes an RR-interval token sequence to capture long-range temporal dependence. A cross-lead attention fusion module merges the two streams, giving different weight to the twelve different leads based on the information they hold, instead of giving them an equal weight. The encoder is trained end-to-end for an NT-Xent contrastive objective where positives are views of the same cardiac segment across the same subject, but negatives are across patients (see S. Chandra et al., 2021) to order the embedding space by the rhythm class of the training subject instead of by the identity of the subject. After the lightweight classification head is trained on a small number of labelled beats, the head is further fine-tuned to be tuned to create an arrhythmia class and a calibrated confidence value, and attention-rollout and Grad-CAM visualisations shed light on the signal intervals and leads on which each classification was based. It formulates the problem, designates the system architecture, mathematically describes each processing stage and proposes an evaluation protocol in terms of macro-averaged F1 score of a strictly separated inter-patient partition, silhouette coefficient of an embedding space learned by the neural network, and localisation accuracy against the intervals annotated by cardiologists. We propose to develop an alternative to the traditional supervised ECG classification that is patient-independent and clinically interpretable and can be labelled efficiently.

KEYWORD: Electrocardiogram (ECG); Arrhythmia Detection; Deep Transformer Network; Dual-Stream Architecture; Self-Supervised Learning; Contrastive Representation Learning; NT-Xent Loss; Cross-Patient Generalisation; Cross-Lead Attention; Explainable AI; Grad-CAM; Biomedical Signal Processing

1 INTRODUCTION

1.1 Background and Motivation

Any irregularity or abnormality in the frequency, rhythm or source of the electric signal that propels the heart muscles is considered cardiac arrhythmia; going from harmless ectopic beats to instantly fatal attempts of ventricular fibrillation. This electrical activity is recorded from 12 different orientations from a twelve-lead electrocardiogram; a disturbed electrical signal that isn't visible in one line can easily be observed in another line. For clinical interpretation, the cardiologist must attend carefully to morphology and rhythm simultaneously in all the leads, which takes a lot of expertise, and isn't easily scalable. However, the proliferation of Holter monitoring,

implantable devices and consumer wearable devices, which continuously record for an increasing number of days or even weeks, have led to ever larger problems -- a single ambulatory study might yield more than one hundred thousand beats, with only a fraction of that being abnormal.

The technology of deep learning has come down to earth to the point that the accurate identification of arrhythmia is now possible. Raw ECG waveforms have been used in convolutional networks, recurrent networks and most recently, transformer encoders, achieving results of close to that of a practising cardiologist on carefully crafted benchmarks. The setting is particularly appealing, since the self-attention mechanism in the transformer attends to every part of a sequence to all the other parts; but conversely, there's no distance-dependent decays in the transformer that recurrent models would suffer from; and the arrhythmia is by nature a phenomenon over a sequence of beats, rather than a beat in isolation. However, the success in practice found in the published record has not been portable to deployment and the difficulty is not so much architectural as it is in the training and evaluation of these systems.

1.2 Problem Statement

Existing automated ECG interpretation systems have two drawbacks, which prevent them from being used in the clinic. The first has to do with subject-to-subject generalisation. Many publications validate the data using an intra-patient partition where the beats are randomly scrambled prior to partitioning and consequently half the “beats” of a single patient appear in the training set while the remaining half are removed and appear in the test set. There are very high amplitude, baseline and morphological correlation between consecutive beats from the same individual, which can be exploited to yield very high reported accuracy – accuracy drops when the same model presented with the same patient, but one that it did not test. The second restriction that has to be considered is the cost of annotations. For supervised training, beat-level labels need to be produced by expert annotators and the not inconsiderable effort required has limited the number of large, diverse, well-balanced and labelled corpora; rare but clinically significant rhythm classes have only been poorly represented and this inevitably biases the classifier towards labelling the vast majority of beats as belonging to the dominant normal class. Another practical problem, obvious from the models that most achieve such good performance that they show no clues or interpretable evidence for their decisions – a clinician cannot decide whether a decision was made due to a true pathological interval or due to a motion artefact. What is needed is a framework that uses largely unlabelled cardiac electrical signals to learn a patient-independent representation of the electrical activity in the heart, that makes use of the complementary information contained in both the morphological (signal) and the rhythmic (timing) aspects of multiple leads and communicates its conclusions in a way that can be checked against the input signal.

1.3 Research Objectives

Objective 1: To develop a multi-lead ECG dual-stream deep transformer encoder which emulates the morphology and rhythm dynamics of the ECG signal in two different streams that are coupled in a non-linear fashion.

Objective 2: To propose a cross-patient unlabelled ECG representation learning via a self-supervised contrastive pre-training strategy, where positive pairs consist of augmented views of the same cardiac segments and negative samples consist of other patient cardiac segments.

Objective 3: To model a cross-lead adaptive attention mechanism, where each lead is given a data-driven importance weight instead of hand-specified ones, and to assign a weight to each representation descriptor associated with each of these leads, again based on the data.

Objective 4: To test and validate the label-efficient framework on arrhythmia classification across different patients following a stringent inter-patient protocol while providing explainable-AI visualisation of the evidence for each prediction.

1.4 Contributions

A brief summary of the main achievements of this work is given below.

A multi-lead ECG is achieved by using a dual stream transformer encoding – isolating “morphology” analysis from “rhythm” analysis instead of fitting a single transformer encoding to fit both. A global transformer encoder is used to model the beat-to-beat irregularity that is present in most arrhythmia classes as a token sequence over an RR interval, while a local-window transformer is applied to a convolutional tokenizer to capture the shape of the P wave, QRS complex and T wave.

Cross patient contrastive representation learning: Here the encoder is pre-trained in an unlabelled fashion using an objective function known as NT-Xent where the negatives are sampled from patients instead of being sampled from other patients in the same rhythm class. This specifically addresses the so-called inter-patient generalisation failure that the intra-patient evaluation herein hides.

So-called cross-lead attention fusion module: In contrast to considering the twelve leads as a single information channel and a single input tensor, the framework is designed to learn a lead specific attention weight (A) for each segment, as clinically, diagnostic information for a specific arrhythmia is compressed into a smaller number of leads depending on the specific cardiac condition.

In addition to using the traditional temporal and morphological descriptors a novel descriptor "Inter-Lead Coherence Index" is defined that quantifies the consistency of a candidate abnormality across leads and a "Lead Dispersion Index" which quantifies the spatial spread of the expression of a candidate abnormality, neither of which is deformed by one lead descriptors.

Label-efficient and auditable operation: as it is a supervised learning, only a small fraction of labelled samples is needed to tune the classification head and attention-rollout, and Grad-CAM provides interval and leads explanations when inspecting the predications for clinical use.

2 LITERATURE REVIEW

The literature on robust multi-lead ECG interpretation shows a clear progression from attention-based sequence modeling and self-supervised representation learning toward patient-independent arrhythmia detection. Vaswani et al. (2017) introduced the Transformer architecture based on self-attention, establishing an effective framework for capturing long-range dependencies in sequential data [1], while Chen et al. (2020) demonstrated the effectiveness of contrastive learning for learning discriminative representations without relying entirely on labeled data [2]. Building on these foundations, Kiyasseh et al. (2021) proposed CLOCS, which specifically applied contrastive learning to cardiac signals across spatial, temporal, and patient dimensions, providing a strong basis for cross-patient ECG representation learning [3]. Deep-learning approaches have also demonstrated strong potential for automated arrhythmia analysis, with Hannun et al. (2019) achieving cardiologist-level arrhythmia detection from ambulatory ECG recordings [4], while Moody and Mark (2001) established the importance of the MIT-BIH Arrhythmia Database as a benchmark resource for ECG research [5]. Wagner et al. (2020) further contributed the PTB-XL dataset, supporting large-scale and standardized ECG learning experiments [6]. For interpretability, Selvaraju et al. (2017) introduced Grad-CAM, enabling visualization of regions contributing to deep-network predictions [7]. Hybrid architectures combining convolutional and recurrent networks have also been explored for arrhythmia classification, particularly for variable-length heartbeats, demonstrating the value of jointly modeling local morphological and temporal information (Oh et al., 2018) [8]. More directly aligned with Transformer-based ECG analysis, Che et al. (2021) developed a constrained Transformer network for ECG signal processing and arrhythmia classification, demonstrating the applicability of attention mechanisms to cardiac time-series analysis [9]. Eldele et al. (2021) investigated temporal and contextual contrasting for time-series representation learning, further supporting contrastive objectives for extracting meaningful temporal features from sequential signals [10]. Within the broader biomedical domain, Kumar et al. (2023) applied optimized deep learning to IoT-based arrhythmia classification, highlighting the potential of intelligent connected healthcare systems for cardiac monitoring [11], while Kailasanathan et al. (2025) developed an interpretable framework for heart-disease prediction using IoMT sensor data [12]. Shitharth et al. (2022) investigated biomedical signal representation using wireless sensors and swarm-intelligence optimization, emphasizing computational processing of physiological signals [13], and Selvarajan et al. (2022) explored artificial intelligence and fuzzy-logic interpretation for biomedical signals in healthcare environments [14]. Deep-learning approaches for other physiological signals, including EEG-based seizure detection and emotion recognition, have further demonstrated the applicability of optimized neural architectures to complex biomedical time-series data (Kantipudi et al., 2024; Karthiga et al., 2024) [15], [16]. Patient-response prediction using multi-scale dilated ensemble networks has also shown the relevance of multi-scale temporal feature extraction in healthcare prediction tasks (Manogaran et al., 2025) [17], while collaborative learning and AI-IoMT frameworks have been investigated for disease prognosis and distributed healthcare intelligence (Nandagopal et al., 2024) [18]. IoT-driven healthcare monitoring combined with evolutionary optimization and game-theoretic mechanisms further demonstrates the growing integration of intelligent algorithms into continuous medical monitoring systems (Selvarajan et al., 2025) [19]. Attention-based deep convolutional learning has also been applied to healthcare security, indicating the broader usefulness of attention mechanisms in complex healthcare data-processing environments (Maruthupandi et al., 2025) [20]. Similarly, optimized deep-learning approaches have been proposed for automated disease classification (Kumar et al., 2025) [21], while attention-based data-fusion techniques have been investigated for IoMT healthcare applications (Khadidos et al., 2023) [22]. Artificial neural networks have additionally been explored for recognition tasks in medical applications, demonstrating the continued role of learned representations in biomedical intelligence (Selvarajan et al., 2023) [23]. Quantum-assisted machine-learning models have been examined for elderly health monitoring, extending intelligent healthcare analytics toward distributed and service-oriented architectures (Bhuvaneswari et al., 2024) [24], whereas digital transformation approaches using audio and virtual-reality procedures illustrate the increasing diversity of AI-enabled medical applications (Shitharth et al., 2023) [25]. Privacy-preserving authentication mechanisms for wearable wireless medical sensor networks address the security requirements associated with continuous physiological-data collection (Rabie et al., 2023) [26], while blockchain-integrated IoT and machine-learning approaches further emphasize secure healthcare data management (Manoharan et al., 2023) [27]. Deep-learning optimization has also been applied to bioinformatics pathway enrichment and data integration, demonstrating the broader potential of representation learning for healthcare analytics (Manoharan & Selvarajan, 2025) [28]. Time-series learning using liquid neural networks has been investigated for brain-tumor disease prediction, reinforcing the importance of temporal modeling in medical prediction tasks (Khan et al., 2025) [29], and lightweight authentication and privacy-preservation techniques have been developed for IoMT environments to support secure medical data exchange (Kumar et al., 2025) [30]. Collectively, these studies establish the importance of Transformer-based attention, contrastive representation learning, temporal modeling, multi-scale feature extraction, interpretability, and secure IoMT integration, providing a literature foundation for developing

a dual-stream deep Transformer framework capable of learning robust cross-patient representations from multi-lead ECG signals for arrhythmia detection.

3 Proposed Methodology

Combining multiple-lead ECG data from a single patient, this study introduces a learning framework that makes use of self-supervised contrastive learning to learn a learner-independent representation of multi-lead ECG, an adaptive cross-lead attention to take advantage of morphological and rhythmic information to learn the multi-lead ECG feature, and an explainable-AI technique to test the learned multi-lead ECG feature to validate the results of the learning framework. The method is implemented in seven steps, such as problem formulation, input preprocessing of dual stream, design of dual stream encoder, self-supervised contrastive input pre-training, feature extraction of representation, adaptive cross-limbard feature learning, explainability validation of classification results, which effectively meets the four research goals mentioned in Section 3.

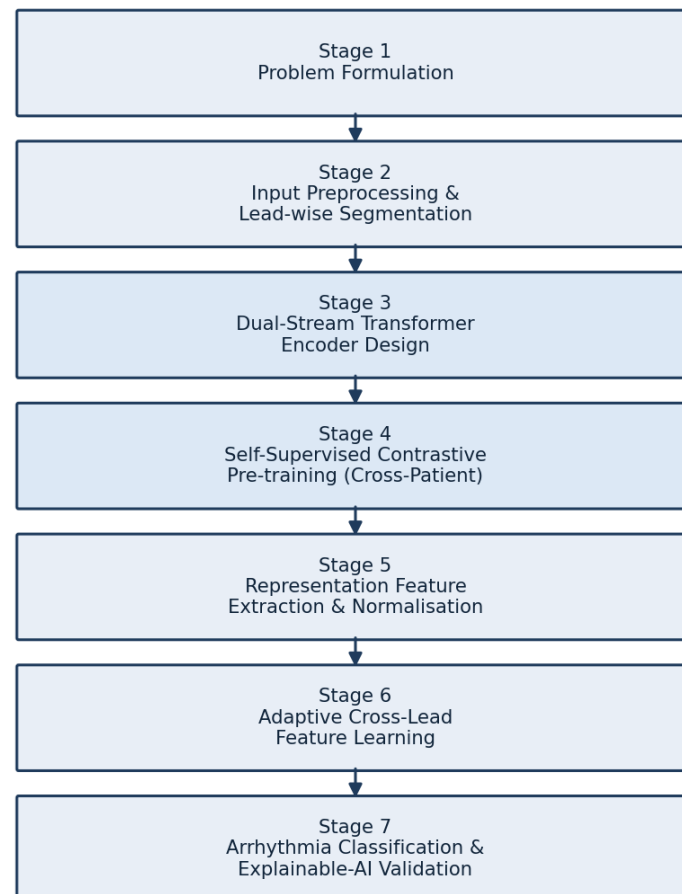


Figure 1. Overview of the seven-stage methodology, from problem formulation through preprocessing, dual-stream encoder design, contrastive pre-training, adaptive cross-lead feature learning, and explainable-AI validation.

3.1 Problem Formulation

A sequence classification problem with a patient-independent representation space is formulated to detect arrhythmia. Multi-lead recording is represented in the form of a matrix of 12 lead-wise recordings sampled within a fixed observation window within which cardiac segments are detected around the R peak. The goal is to learn an encoder that maps each segment to an embedding vector where segments learned from the same patient that are recorded in the same rhythm class are near each other regardless of their superficial amplitude or baseline difference, while those registered in different rhythm class are far away regardless of their superficial amplitude or baseline difference. The training of encoders consists of two steps. The first scheme involves no labels, and the encoder is designed such that two augmented views of the same segment will have higher attraction and the segments coming from other patients in the batch have higher repulsions. The second one is a light-weighted-classification head which is more finely tuned on a small size of labelled data to figure out the learned embedding and also a classification-confidence score. Evaluation is based on an inter-patient partition used during the evaluation process, where no recording shares segments in any two separate splits to leave no trace of memorised characteristics for the specific subjects during the evaluation process.

3.2 System Architecture Overview

The overall system architecture is depicted by the following; Figure 2. The first step in the pre-processing of the raw twelve-lead recording includes the removal of baseline wander, power line interference, resampling to a common sample rate and the segmentation (into fixed windows) around detected R peaks (R wave), and an

amplitude normalisation of each lead. The dual-stream transformer encoder transforms each pre-processed segment in a different dual stream, with the first processing the shape of the waveforms in a morphological stream comprised of a convolutional tokenizer and a local-window transformer, and the second processing the rhythmicity of the beats in a temporal stream comprised of a global transformer operating on an RR-interval based token sequence. The cross-lead attention fusion module combines the two streams by learning the informativeness of each stream for each segment analysed and using this as a weight to contribute to the overall representation of the analysed segment in a shared embedding space. This representation increases, optimised in a pre-training stage, using the contrastive objective, but without labels. During fine tuning, the same representation is sent to a classification head, which outputs data comprising the class of given arrhythmia and confidence range; then, the explainable-AI module will display the temporal intervals and leads, which were used with that classification, for fine tuning the model, so that each prediction can be checked against the data.

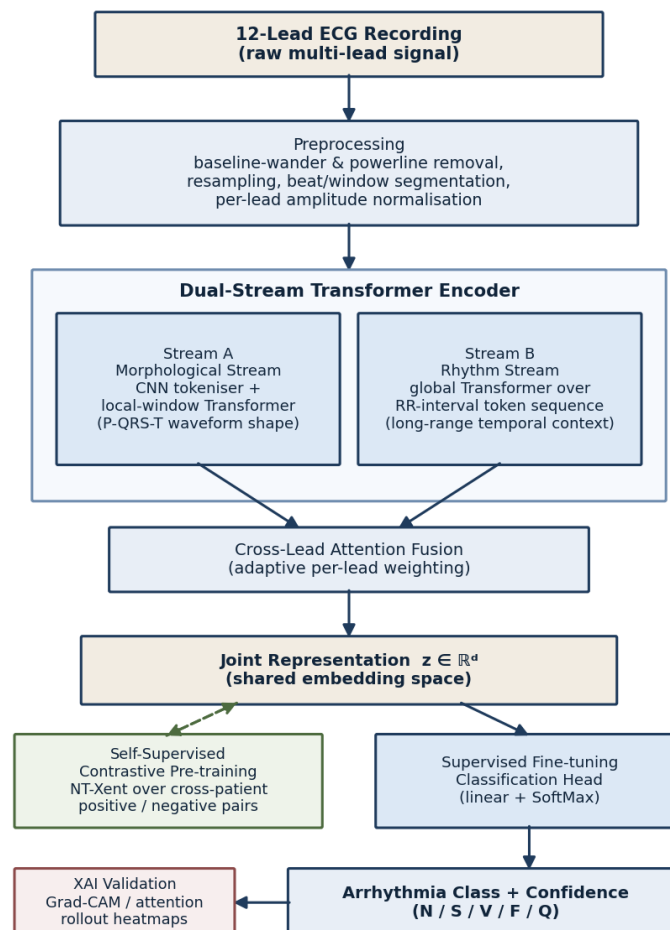


Figure 2. Detailed architecture of the proposed framework, from the raw twelve-lead input through preprocessing, the dual-stream transformer encoder, cross-lead attention fusion, contrastive pre-training and supervised fine-tuning, to arrhythmia classification and the XAI validation module.

3.3 Input Preprocessing

Preprocessing standardises recordings which vary in sampling rate, amplitude calibration, noise level and electrode placement to ensure that this variation does not continue throughout the learned representation, otherwise it will become subject specific and be one of the ways that a model will become subject specific. High pass filtering with a cut-off frequency sufficiently low to retain the ST segment eliminates baseline wander due to respiration and electrode movements while a narrow notch filter at the mains frequency also removes power line interference. All recordings are resampled to a common rate, in order to have a fixed number of samples that has a fixed physical duration in the datasets. Then, R peaks will be identified, and every segment will be chopped within a window that contains the beat of interest, in addition to a certain number of beats before and after the chosen beat, providing the rhythm stream with the temporal context necessary for irregularity characterization. Every lead is normalised independently and given a zero mean and unit variance in the segment, to normalize the amplitude of the signals caused by differences in the electrode contact. In the contrastive pretraining, two stochastically generated versions of each segment are shown by augmenting its signal with amplitude scaling,

temporal masking, Gaussian noise injection and random lead dropout (where appropriate) to the multi-lead setting; the latter helps to keep the encoder from relying on a particular lead.

3.4 Dual-Stream Transformer Encoder Module

The structure of the core component in the framework is encoder, and its internal structure is shown in Figure 3. A segment is first pre-processed and then output as a token sequence. Morphological stream: A one dimensional convolutional tokenizer is used to extract local features of the waveform to generate patch embeddings which are enhanced with positional embeddings, followed by a multi-head self-attention transformer operating in restricted local windows with each head learning a unique convolutional filter, between which only very short-range attention, such as that of the QRS complex, is calculated—computational effort concentrated on the fine-structured region of the P wave, QRS complex and T wave. The segment is coded as a stream of RR-interval tokens and a transformer trained by a global one attention head is used to hear the entire sequence to build irregularity models, that characterize the irregularity patterns observed during atrial fibrillation, ectopic rhythms, conduction defects. The operation for attention and both streams is given by

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T / \sqrt{d}) V \quad (1)$$

Q, K and V are the query matrix, key matrix and value matrix projected from the token embeddings, respectively and d is the key dimension for scaling the dot product result prior to applying SoftMax normalisation. It's intentional that it's split into two streams. It is important to note that one single encoder has to mind a certain attention budget for morphology material at millisecond scale, as well as for the material of structure of the multi-second rhythm; in practice, the latter has a higher energy level, and consequently a higher attention demand. The framework encourages the separation of each scale to a distinct stream with the appropriate attention span, which is then reconciled by the cross-attention instead of simply concatenating them together and shares the scales while allowing the intertwining of morphological information within the framework of the existing rhythm and vice-versa.

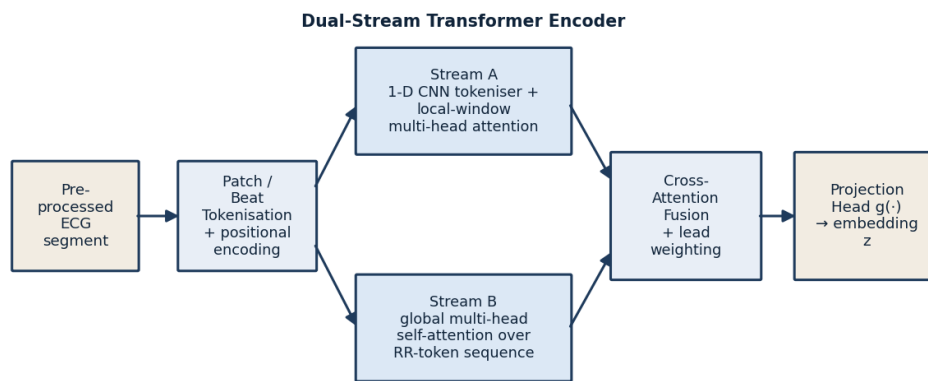


Figure 3 Internal structure of the dual-stream transformer encoder: patch and beat tokenisation with positional encoding, a morphological stream with local-window attention, a rhythm stream with global attention, cross-attention fusion with lead weighting, and the projection head producing the embedding.

3.5 Self-Supervised Contrastive Pre-training

The embedding space is pre-trained without using any labels to optimize the encoder in order to maximize the space of getting different cardiac states. Two augmented views are created for each segment in the batch and sent to the encoder, weighted by the non-linear projection head, to yield a pair of embeddings each: a positive pair. All other embeddings are gathered and pulled exclusively from other patients: all of these are negative embeddings in the batch. The loss function for the trainer is the normalised temperature-scaled cross-entropy loss commonly used for training an encoder [2] that is represented for a positive pair (i, j) as

$$L(i, j) = -\log \left[\exp(\text{sim}(z_i, z_j) / \tau) / \sum \exp(\text{sim}(z_i, z_k) / \tau) \right] \quad (2)$$

The function uses cosine similarity between two projected embeddings denoted as 'ou sim', a parameter 't' to control the sharpness of the similarity distribution, and the summation in the denominator is performed over all of the negative samples in the batch. One key part of the design is to have a cross-patient negative sample. If we drew back only within one recording, the encoder could attempt to reduce the loss by encoding only some characteristics of a subject (baseline offset or amplitude dependent on the type of electrode) and this would address this many of the cases of intra-patient accuracy failing in inter-patient evaluation. However, creating negatives across patients is unprofitable when drawing negatives, since it no longer distinguishes a positive from a negative, but pushes the encoder towards features that are common among patients, and discriminatory of the cardiac condition itself.

3.6 Representation Feature Extraction Module

The fused representation is further described by several quantitative analysis descriptors covering the temporal, morphological and inter-lead analysis aspects. The aforementioned descriptors give an interpretable summary of the evidence embedded, which are given after normalisation to the adaptive feature-learning module. These are presented in Table 1 below.

Feature	Description	Analysis View
RR-Interval Variability	Dispersion of successive RR intervals within the segment; elevated variability indicates irregular rhythm such as atrial fibrillation.	Temporal Analysis
QRS Morphology Descriptor	Width, amplitude and shape deviation of the QRS complex relative to the patient-normalised template; captures ventricular ectopy and conduction delay.	Morphological Analysis
ST-Segment Deviation	Signed displacement of the ST segment from the isoelectric baseline, quantifying ischaemic change accompanying the rhythm abnormality.	Morphological Analysis
P-Wave Presence Score	Confidence that a discernible P wave precedes the QRS complex; its absence is a primary discriminator of supraventricular arrhythmia.	Morphological Analysis
Inter-Lead Coherence Index (novel)	Measures whether a candidate abnormality is expressed consistently across leads or appears in isolation, separating genuine pathology from lead-specific artefact.	Inter-Lead Analysis
Lead Dispersion Index (novel)	Quantifies the spatial spread of abnormal activity across the twelve lead axes, distinguishing localised from widely distributed involvement.	Inter-Lead Analysis

Table 1. Quantitative representation descriptors extracted from the fused dual-stream embedding.

3.7 Feature Normalization

Note that each of the six descriptors has a different set of units, along with a different range of numbers; thus min–max normalisation is used before the descriptors are fused, so that they are put on a common scale:

$$X_{norm} = (X - X_{min}) / (X_{max} - X_{mi}) \quad (3)$$

so that all the normalised descriptors are within this range [0, 1]. If this is not the case, a descriptor with the dimension of milliseconds would be the only difference (besides its meaning) between a dimensionless ratio; and the learned attention weights would tell us more about the units used than about the diagnostic informativeness of the descriptor. Normalisation statistics are calculated for a single training split and used for other splits without any changes; that is, no information about unseen patients is used to affect the scaling.

3.8 Adaptive Cross-Lead Feature Learning

The conventional multi-lead models stack the twelve leads as channels in a 1 input tensor, so they automatically give the same weights to all leads. This is not a clinically realistic study; there is a focus on inferior ischaemia on the inferior leads, on anterior on the precordial leads, and on informative leads depends on the pathology and on the quality of the connections made by the electrodes in a particular recording. The adaptive feature-learning module thus learns an attention score for each lead and each normalised descriptor, avoiding the use of a per-expert instruction on how to weight them. The scores are normalised by so-called softmax function, so that the weights will add up to one and the output is a sort of representation vector for the output, where weight of each lead and weight of each descriptor correspond to the estimated informativeness of the corresponding lead and descriptor for the analysed segment. Weighing a particular recording segment high if it was recorded with high-quality equipment and low if it was badly recorded with shoddy gear, is a characteristic a fixed weighting scheme can't offer: a corrupted recording will be given a low weight in the case of motion artefact, but no fixed weight scheme could say that it was permanently of poor quality.

3.9 Arrhythmia Classification and Confidence Scoring

A light backbone classification head, consisting of a linear projection layer and SoftMax normalisation, is fine-tuned by gradient descent on a small, labelled portion of the data with the learning rate of the pre-trained encoder either being frozen or being decayed by a small factor. A weighted feature vector where the components are f_1, f_6 and the learned weights are w_1, w_6 , and whose (fused) representation is fed to the head is equal to

$$F = \sum_{i=1}^6 w_i \cdot f_{i, orm} \quad (4)$$

The output of the SoftMax is regarded as a posterior distribution of the arrhythmia classes, and the maximum posterior is used as the confidence value with the predicted class. The classes are aligned with how the beats are annotated in the reference corpora that were provided by the AAMI group (Table 1.2's grouping). In addition to reporting the class they are confident of they report their level of confidence, which allows for silent rejection but can alternatively alerted to manual verification, and the confidence level can be adjusted for each deployment depending on the relative risk of missing an arrhythmia compared to a false positive alarm.

Class	AAMI Label	Interpretation
Normal	N	Normal sinus beats and bundle-branch block beats originating in the sinoatrial node.
Supraventricular	S	Supraventricular ectopic beats, including atrial and nodal premature beats; typically marked by an absent or abnormal P wave.

Class	AAMI Label	Interpretation
Ventricular	V	Ventricular ectopic beats, including premature ventricular contractions; characterised by a widened QRS complex without a preceding P wave.
Fusion	F	Fusion of a ventricular and a normal beat, producing intermediate morphology.
Unknown	Q	Paced beats, fusion of paced and normal beats, and beats that cannot be classified.

Table 2. Arrhythmia classes following the AAMI grouping used for classification and reporting.

3.10 Explainable AI Validation Module

It helps make the evidence behind each prediction visible and clinically reliable by presenting it in two forms of visualisation: a traditional one and one in the Pie Chart format. The framework presents evidence behind each prediction in two different yet complementary ways so as to allow transparency and clinical reliability. For Gradient-weighted Class Activation Mapping [7] the gradient of the class to be predicted is back-propagated to the last set of convolutional feature maps of the morphological stream and results in a one-dimensional heatmap over the segment, with high activations pointing to the time periods that mostly contributed to the prediction. The attention stream can be summarized by a summarization group, whose attention matrix is the average of all the attention matrices in the rhythm stream, to mean the beats that had the highest impact on the classification in the sequence, and the cross-lead attention matrix corresponds to the leads during which the decision was made, where the class tends to be the most closely matched. These outputs enable a clinician to be certain that a prediction was based on a clinically important finding – a lack of P wave or a wide QRS – and not due to a motion artefact or a saturated electrode, a prerequisite for which trust in an automated interpretation system can be placed.

4 Evaluation Protocol

Three complimentary indicators are suggested as measures for different phases of the framework (Table 3). The macro-averaged F1 score, calculated according to a rigorous inter-patient partition (no segment from any recording belong to more than one partition), is used to assess classification performance. It is preferable to overall accuracy to use macro averaging because of the uneven distribution of the data for the arrhythmia classes, where the majority class comprises the majority of the data, and thus a model that always predicted the majority class would report a falsely high accuracy, while predicting no important events. To evaluate the quality of the representation learned during the contrastive pre-training, the embedding is first evaluated via a silhouette score on the embedding instead of using a classifier, and the evaluation is performed on the embedding divided into a grouping variable (here patient identity). This gives a direct test as to whether a subject-specific representation has been suppressed and has been computed on the embedding that can be trained on the data before a classifier is fit on it. The explainability layer is judged with localisation accuracy that is calculated as the overlap of the intervals marked by the output of the Grad-CAM along with that of the attention-rollout layers with the intervals marked by a cardiologist. They also report the efficiency of the label by fine-tuning with decreasing fractions of labels and saving the performance level (macro F1 score) in each step, which makes it clear how much label was needed to obtain a certain level of performance.

Metric	Purpose	Applied To
Macro-Averaged F1 Score	Quantifies per-class detection performance under a strict inter-patient partition, without allowing the majority class to mask minority-class failure.	Classification Module
Silhouette Score	Assesses cohesion and separation of segment clusters in the learned embedding space, by rhythm class and by patient identity.	Contrastive Pre-training Module
Localisation Accuracy	Measures whether Grad-CAM and attention-rollout outputs correctly highlight cardiologist-annotated intervals and leads.	XAI Validation Module

Table 3. Proposed evaluation metrics and the framework module each target.

5 Expected Outcomes and Discussion

The design of the framework is likely to have three main benefits to the existing practice. First, in order to narrow the gap between intra-patient and inter-patient performance typical of supervised ECG classifiers, one has to ensure that one trains the encoder to purposefully ignore the subject identity as discriminative feature: in practice, one will thus get a model whose reported accuracy will be closer to what it achieves on unseen patients. Secondly, for the rare classes in which it is most difficult to obtain and most useful to annotate, the amount of expert labelling required to achieve clinically useful performance, should be greatly reduced when learning the representation is done without labels. Third, with the dual stream (cross led) attention, there should be an improvement in the detection of arrhythmias whose evidence may be unevenly distributed across the leads or distributed across multiple beats as a single-stream encoder applied to stacked channels would tend to be dominated by the highest-

energy component of the signal. Otherwise, these expectations need to be validated with empirical data from cardiologists annotated with data.

Practicality in the computational and clinical settings also must be considered. The pre-training is the costliest component of the framework, as it needs large batches and long training timelines, but it is done offline and once. There are only few parameters in the lightweight classification head, with a forward pass of only a single pass through the encoder, and the global attention of the rhythm stream works on a short sequence of RR-integers, not on raw sounds, which only makes the quadratic cost of self-attention reasonably practicable. It is therefore also desirable to be able to deploy the framework in the near Realtime domain using standard clinical computing infrastructure and a reduced width configuration of the encoder is a viable path to deploying this framework on ambulatory/wearable device platforms although the relationship of the encoder widths to the inter-patient accuracy would have to be characterised.

6 CONCLUSION AND FUTURE WORK

This paper presents a dual-stream deep transformer network for strong multi-lead ECG interpretation, which tackles the inter-patient generalisation failure and annotation dependence of the current automated arrhythmia detection systems. The framework is composed of a stream of morphological evolution mapping (morphological stream), a stream of rhythm evolution mapping (rhythm stream) that model the shape of the recorded morphology and evolvability of the dynamics between beats, respectively, using separate attention streams, a cross-lead attention fusion module which learns the relative informativeness of each lead, a self-supervised contrastive pre-training phase which organises the embedding space by cardiac state instead of by the identity of the subject, and two novel representation descriptors (Inter-Lead Coherence Index, Lead Dispersion Index) capturing inter-lead properties not accessible to single-lead analysis. Grad-CAM and attention-rollout visualisations are integrated within the framework, and clinical evidence for each prediction can be seen.

The empirical test of this model described here is the next stage of this research. This starts with a pre-trained dual-stream encoder on unlabelled, yet twelve-lead records from PTB-XL [6] and the embedding that is learnt from the pre-training phase is subsequently evaluated with respect to the silhouette score by rhythm class and patient ID to demonstrate that subject-specific information has been suppressed. Then, fine-tuning and inter-patient evaluation will proceed with beats annotated from the MIT-BIH Arrhythmia Database [5] using the AAMI class grouping and the F1 macro-averaged across progressively smaller labelled fractions to quantify the degree of successful labelling, and the localisation in accordance with the ground truth [6] interval annotated by the cardiologists. Generalisation across different devices and electrode types, and across reduced-lead and single-lead wearable inputs where a cross-lead attention mechanism is required to handle incomplete spatial data will be explored. Extension to reduced-lead and single-lead wearable recording devices will be explored where low-confidence predictions will be automatically sent for expert review to create a clinical screening workflow will be addressed.

REFERENCES

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS 2017)*, 30, pp.5998–6008.
2. Chen, T., Kornblith, S., Norouzi, M. and Hinton, G., 2020. A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, PMLR 119, pp.1597–1607.
3. Kiyasseh, D., Zhu, T. and Clifton, D.A., 2021. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, PMLR 139, pp.5606–5615.
4. Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P. and Ng, A.Y., 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1), pp.65–69. DOI: 10.1038/s41591-018-0268-3.
5. Moody, G.B. and Mark, R.G., 2001. The impact of the MIT-BIH Arrhythmia Database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3), pp.45–50. DOI: 10.1109/51.932724.
6. Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F.I., Samek, W. and Schaeffter, T., 2020. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7, p.154. DOI: 10.1038/s41597-020-0495-6.
7. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017)*, pp.618–626. DOI: 10.1109/ICCV.2017.74.
8. Oh, S.L., Ng, E.Y.K., Tan, R.S. and Acharya, U.R., 2018. Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. *Computers in Biology and Medicine*, 102, pp.278–287. DOI: 10.1016/j.combiomed.2018.06.002.
9. Che, C., Zhang, P., Zhu, M., Qu, Y. and Jin, B., 2021. Constrained transformer network for ECG signal processing and arrhythmia classification. *BMC Medical Informatics and Decision Making*, 21, p.184. DOI: 10.1186/s12911-021-01546-2.

10. Eldele, E., Ragab, M., Chen, Z., Wu, M., Kwoh, C.K., Li, X. and Guan, C., 2021. Time-series representation learning via temporal and contextual contrasting. *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI 2021)*, pp.2352–2359. DOI: 10.24963/ijcai.2021/324.
11. Kumar, A., Kumar, S. A., Dutt, V., Shitharth, S., & Tripathi, E. (2023). IoT based arrhythmia classification using the enhanced hunt optimization-based deep learning. *Expert Systems*, e13298. <https://doi.org/10.1111/exsy.13298>.
12. Kailasanathan, N., Ezhilarasan, G., Selvarajan, S., Dhanaraj, R.K., Pamucar, D., & Shankar, N. (2025). Heart disease prediction with a feature-sensitized interpretable framework for the Internet of Medical Things sensors. *Frontiers in Digital Health*, 7, 1612915. doi:10.3389/fdgth.2025.1612915.
13. Shitharth, S., Yonbawi, S., Manoharan, H., Alahmari, S., Yafoz, A., & Mujlid, H. (2022). Physical Stint Virtual Representation of Biomedical Signals with Wireless Sensors Using Swarm Intelligence Optimization Algorithm. *IEEE Sensors Journal*. doi:10.1109/JSEN.2022.3233407.
14. Selvarajan, S., Manoharan, H., Hasanin, T., Alsini, R., Uddin, M., Shorfuzzaman, M., & Alsufyani, A. (2022). Biomedical Signals for Healthcare Using Hadoop Infrastructure with Artificial Intelligence and Fuzzy Logic Interpretation. *Applied Sciences*. <https://doi.org/10.3390/app12105097>.
15. Kantipudi, M.V.V.P., Kumar, N.S.P., Aluvalu, R. et al. (2024). An improved GBSO-TAENN-based EEG signal classification model for epileptic seizure detection. *Scientific Reports*, 14, 843. <https://doi.org/10.1038/s41598-024-51337-8>.
16. Karthiga, M., Suganya, E., Sountharajan, S. et al. (2024). EEG based smart emotion recognition using meta heuristic optimization and hybrid deep learning techniques. *Scientific Reports*, 14, 30251. <https://doi.org/10.1038/s41598-024-80448-5>.
17. Manogaran, N., Panabakam, N., Selvaraj, D. et al. (2025). An efficient patient's response predicting system using multi-scale dilated ensemble network framework with optimization strategy. *Scientific Reports*, 15, 15713. <https://doi.org/10.1038/s41598-025-00401-y>.
18. Nandagopal, M., Seerangan, K., Govindaraju, T. et al. (2024). A Deep Auto-Optimized Collaborative Learning (DACL) model for disease prognosis using AI-IoMT systems. *Scientific Reports*, 14, 10280. <https://doi.org/10.1038/s41598-024-59846-2>.
19. Selvarajan, S., Manoharan, H., Al-Shehari, T. et al. (2025). IoT driven healthcare monitoring with evolutionary optimization and game theory. *Scientific Reports*, 15, 15224. <https://doi.org/10.1038/s41598-025-99129-y>.
20. Maruthupandi, J., Sivakumar, S., Dhevi, B.L. et al. (2025). An intelligent attention based deep convoluted learning (IADCL) model for smart healthcare security. *Scientific Reports*, 15, 1363. <https://doi.org/10.1038/s41598-024-84691-8>.
21. Kumar, G.S., Suganya, E., Sountharajan, S. et al. (2025). SRADHO: statistical reduction approach with deep hyper optimization for disease classification using artificial intelligence. *Scientific Reports*, 15, 1245. <https://doi.org/10.1038/s41598-024-82838-1>.
22. Khadidos, A.O., Khadidos, A.O., Selvarajan, S., & Mirza, O.M. (2023). TasLA: An innovative Tasmanian and Lichtenberg optimized attention deep convolution based data fusion model for IoMT smart healthcare. *Alexandria Engineering Journal*, 79, 337–353. <https://doi.org/10.1016/j.aej.2023.08.010>.
23. Selvarajan, S., Manoharan, H., Iwendi, C., Alsowail, R.A., & Pandiaraj, S. (2023). A comparative recognition research on excretory organism in medical applications using artificial neural networks. *Frontiers in Bioengineering and Biotechnology*, 11, 1211143. doi:10.3389/fbioe.2023.1211143.
24. Bhuvaneswari, A., Srivel, R., Elamathi, S., Shitharth, S., & Sangeetha, K. (2024). Enhancing Elderly Health Monitoring Framework With Quantum-Assisted Machine Learning Models as Micro Services. *Quantum Innovations at the Nexus of Biomedical Intelligence*. doi:10.4018/979-8-3693-1479-1.ch002.
25. Shitharth, S., Manoharan, H., Nkechi, O., & Iwendi, C. (2023). Digital Transformations in Medical Applications Using Audio and Virtual Reality Procedures. In *Proceedings of ICACTCE'23, Lecture Notes in Networks and Systems*, vol. 735. https://doi.org/10.1007/978-3-031-37164-6_45.
26. Rabie, O.B.J., Selvarajan, S., Hasanin, T. et al. (2023). A full privacy-preserving distributed batch-based certificate-less aggregate signature authentication scheme for healthcare wearable wireless medical sensor networks (HWMSNs). *International Journal of Information Security*. <https://doi.org/10.1007/s10207-023-00748-1>.
27. Manoharan, H., Manoharan, A., Selvarajan, S., & Venkatachalam, K. (2023). Implementation of Internet of Things With Blockchain Using Machine Learning Algorithm: Enhancement of Security With Blockchain. In *Handbook of Research on Blockchain Technology and the Digitalization of the Supply Chain*. <https://doi.org/10.4018/978-1-6684-7455-6.ch019>.
28. Manoharan, H., Selvarajan, S. (2025). PEDI: Towards Efficient Pathway Enrichment and Data Integration in Bioinformatics for Healthcare Using Deep Learning Optimisation. *Biomedical Engineering and Computational Biology*, 16. doi:10.1177/11795972251321684.
29. Khan, F., Amanullah, S.I. & Selvarajan, S. (2025). Linear regressive weighted Gaussian kernel liquid neural network for brain tumor disease prediction using time series data. *Scientific Reports*, 15, 5912. <https://doi.org/10.1038/s41598-025-89249-w>.

30. Kumar, S., Abhishek, K. & Selvarajan, S. (2025). A lightweight and secure authentication and privacy protection scheme for internet of medical things. *Scientific Reports*, 15, 23876. <https://doi.org/10.1038/s41598-025-05910-4>