CardioPatternFormer: Pattern-Guided Attention for Interpretable ECG Classification with Transformer Architecture

1st Berat Kutay Uğraş
Department of Electrical and Electronics Engineering
Eskisehir Technical University
Eskisehir, Turkey
ORCID: 0009-0006-3760-4870

2nd İbrahim Talha Saygı
Department of Electrical and Electronics Engineering
Eskisehir Technical University
Eskisehir, Turkey
email address or ORCID

3rd Ömer Nezih Gerek
Department of Electrical and Electronics Engineering
Eskisehir Technical University
Eskisehir, Turkey
ORCID: 0000-0001-8183-1356

***Abstract*—Electrocardiogram (ECG) interpretation is fundamental to cardiac diagnosis, but deep learning models often lack transparency, hindering clinical trust. We introduce CardioPatternFormer, a novel transformer-based architecture that reframes ECG interpretation through the lens of pattern recognition, treating cardiac patterns as a vocabulary learned from data. CardioPatternFormer integrates several innovations: (1) a Cardiac Pattern Tokenizer that decomposes ECG signals into learned, multi-scale patterns; (2) Physiologically Guided Attention mechanisms incorporating adaptable, domain-specific constraints based on cardiac electrophysiology; (3) Multi-Resolution Temporal Encoding to capture diverse temporal dynamics; and (4) specialized classification heads providing class-specific attention visualizations for detailed diagnostic explanations. Evaluated on the Chapman-Shaoxing dataset across major diagnostic categories, CardioPatternFormer demonstrates strong classification performance, particularly for rhythm disorders, with results aligning with clinical experience regarding diagnostic difficulty gradients. More significantly, CardioPatternFormer enhances interpretability by visualizing physiologically relevant ECG regions influencing each diagnosis, bridging automated analysis and clinical reasoning. This pattern-centric approach advances ECG classification and establishes a foundation for more transparent and clinically integrated cardiac signal analysis.**

***Keywords*—ECG classification, Transformers, Deep learning, Medical signal processing, Physiological Attention, Cardiac Pattern Tokenization, Interpretability**

I. INTRODUCTION

The electrocardiogram (ECG) stands as a cornerstone of cardiovascular diagnostics, offering a non-invasive window into the heart's electrical activity. Annually, hundreds of millions of ECGs are performed worldwide, underscoring its fundamental role in identifying a wide spectrum of cardiac conditions [20]. Despite its ubiquity and the wealth of information it provides, accurate ECG interpretation demands considerable expertise, typically cultivated over years of rigorous training and clinical practice. Even amongst seasoned cardiologists, inter-reader variability remains a notable challenge, with studies reporting significant discrepancies in the identification of specific cardiac abnormalities. This variability highlights not only the inherent complexities of ECG analysis but also the persistent need for advanced computational tools that can deliver consistent, precise, and interpretable analyses.

Deep learning (DL) methodologies have demonstrated remarkable potential in automating ECG interpretation, with several studies showcasing their ability to achieve cardiologist-level performance for specific diagnostic tasks. However, the clinical adoption of these powerful algorithms has been hampered by two primary limitations. Firstly, many DL models operate as "black boxes," offering limited insight into their decision-making processes, which can erode clinical trust and hinder their integration into routine practice. Secondly, these models often distill the rich physiological data embedded within ECG signals into simplified binary or categorical outputs, thereby losing the nuanced patterns and temporal relationships that clinicians utilize for comprehensive diagnostic reasoning.

We propose that ECG signals can be conceptualized as a complex "language" of cardiac electrophysiology. This language comprises discrete, recognizable elements (the "vocabulary") which combine in specific temporal sequences (the "grammar") to form a complete diagnostic picture. Just as human languages convey meaning through words and syntax, cardiac pathologies manifest through distinct patterns and their temporal interplay within the ECG. This perspective suggests that the challenge of ECG interpretation can be effectively addressed through advanced pattern recognition and contextual understanding, areas where transformer architectures, initially popularized in natural language processing [2], have shown

exceptional promise.

In this paper, we introduce CardioPatternFormer, a novel framework for ECG interpretation that leverages a physiologically-informed pattern tokenization approach coupled with specialized attention mechanisms. Our primary contributions are:

- A CardiacPatternTokenizer that transforms continuous ECG signals into a vocabulary of learned cardiac patterns, capturing salient electrophysiological features across multiple time scales.
- Physiologically Guided Attention, a novel mechanism that refines standard attention scores by incorporating adaptable biases derived from established principles of cardiac electrophysiology, such as signal locality and rhythmic characteristics. The influence of these biases is governed by learnable weights, thereby guiding the model to focus more effectively on diagnostically relevant signal features.
- A Multi-Resolution Temporal Encoding scheme designed to capture both localized morphological abnormalities and broader rhythm patterns, which are critical for comprehensive ECG interpretation.
- Class-specific attention visualization derived from specialized classification heads, providing detailed, interpretable explanations for each diagnostic prediction, thereby addressing the critical need for explainability in clinical applications.

CardioPatternFormer demonstrates strong classification performance on the publicly available Chapman-Shaoxing ECG dataset [3] across six major diagnostic categories, with notable strength in identifying complex rhythm disorders and conduction abnormalities. More significantly, it offers enhanced transparency in its decision-making process through visualizations that highlight the specific ECG regions influencing each diagnosis, thereby creating a valuable bridge between automated analysis and clinical reasoning. Furthermore, we evaluate the model's adaptability to varying lead availability through a comprehensive lead ablation study.

While current approaches to automated ECG interpretation primarily frame the task as classification, the pattern vocabulary and attention mechanisms developed in this work also establish a foundation for more sophisticated diagnostic understanding. The foundational pattern recognition capabilities established here may support future work exploring automated report generation or other advanced diagnostic communication methods.

The remainder of this paper is organized as follows: Section II discusses related work in deep learning for ECG interpretation and attention-based explainability. Section III details the CardioPatternFormer architecture and its components. Section IV presents our experimental setup and results. Section V discusses the implications of our approach, its clinical potential, limitations, and future directions, followed by the Conclusion in Section VI.

II. RELATED WORK

A. Deep Learning for ECG Interpretation

The application of deep learning to ECG interpretation has witnessed substantial growth and evolution over the past decade. Initial forays predominantly relied on Convolutional Neural Networks (CNNs) to automatically extract morphological features from ECG waveforms. Deep CNN architectures demonstrated notable efficacy in arrhythmia detection, particularly from single-lead ECGs, with some studies reporting performance comparable to that of experienced cardiologists for specific rhythm disorders [1]. This paradigm was subsequently extended to 12-lead ECG analysis, often employing advanced CNN structures like Residual Networks (ResNets) [4]. Furthermore, researchers have successfully utilized CNNs to identify subtle ECG indicators of conditions such as left ventricular dysfunction, uncovering previously unrecognized signal patterns with significant diagnostic value [5].

To address the temporal dependencies inherent in ECG data, Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, have been extensively applied. These models are well-suited for capturing sequential information and have been used for tasks ranging from arrhythmia classification to the prediction of adverse cardiac events [6], [7]. Hybrid models, which integrate the feature extraction strengths of CNNs with the sequential modeling capabilities of RNNs (often termed CNN-RNN architectures), have also emerged as a popular approach, demonstrating improved performance in complex tasks like multi-label arrhythmia detection from extended ECG recordings [8]. More recently, self-supervised pre-training on large, unlabeled ECG datasets has shown promise in learning robust and generalizable feature representations, which can then be fine-tuned for specific diagnostic tasks with enhanced data efficiency [9]. Despite these significant advances, a persistent challenge across many of these models is their limited interpretability, often treating ECG analysis as a complex pattern-matching problem without explicitly modeling underlying physiological mechanisms.

B. Transformer Architectures for Physiological Signals

Transformer architectures, originally introduced for natural language processing tasks [2], have recently garnered significant attention for their potential in analyzing physiological time series, including ECGs. Early applications of transformers to ECG interpretation demonstrated their ability to effectively model long-range dependencies within the signal, leading to improved performance in arrhythmia detection compared to conventional CNN and RNN approaches [10]. Subsequent research has focused on adapting transformer architectures more specifically for ECG data, for instance, by incorporating relative positional encodings to better capture the temporal relationships crucial for cardiac signal analysis. General-purpose transformer-based architectures for time series, such as TimesNet, have also shown strong performance on ECG classification benchmarks [11]. However, many existing methods apply transformer architectures with

minimal domain-specific modifications, often treating ECG signals as generic time-series data rather than as reflections of complex cardiac electrophysiology. A key limitation in these approaches often lies in the tokenization strategy, where individual time points or fixed-size windows are used as tokens, which may not align optimally with physiologically meaningful segments or events within the ECG.

C. Explainability in Cardiac Models

The integration of deep learning models into clinical decision-making processes necessitates a high degree of transparency and interpretability. This has spurred significant research into eXplainable Artificial Intelligence (XAI) for cardiac applications. In ECG interpretation, several techniques have been explored to elucidate model predictions. Gradient-based visualization methods, such as Grad-CAM [12] and its variants, have been widely applied to generate saliency maps that highlight the regions of the ECG signal most influential in a model's decision [13]. Such visualizations can aid clinicians in understanding the basis of a model's output and in identifying potential biases or reliance on spurious features [14].

Attention mechanisms, which are integral to transformer architectures but can also be incorporated into other neural network designs, provide an alternative pathway to explainability. By examining attention weights, it is possible to infer which segments of the input ECG signal the model prioritizes when forming a prediction [15]. More sophisticated approaches have explored hierarchical attention mechanisms that aim to distinguish between features at different scales, such as beat-level morphology versus rhythm-level patterns, thereby offering explanations that are more aligned with clinical reasoning [13], [16]. Nevertheless, many of these methods provide post-hoc explanations or rely on generic attention patterns, rather than integrating domain-specific cardiac knowledge directly into the attention computation process, which could lead to more inherently interpretable and clinically relevant models.

D. Pattern Recognition in Cardiac Signals

The concept of recognizing discrete, meaningful patterns within ECG signals is fundamental to clinical cardiology. For decades, clinicians have relied on identifying specific waveforms (e.g., P wave, QRS complex, T wave), intervals (e.g., PR, QT intervals), and morphological characteristics to diagnose cardiac conditions [18]. Early computational approaches to ECG analysis attempted to codify this expert knowledge through rule-based systems and explicit feature engineering. However, these traditional methods often struggled with the inherent variability and noise present in real-world ECG data.

More recently, studies have shown that deep learning models, even when trained end-to-end, can implicitly learn to identify patterns that correspond to known ECG abnormalities, though these learned features often remain latent within the network's complex architecture [4]. The idea of learned feature dictionaries or "electrophysiological motifs" has been explored, demonstrating that models can discover representations that align with clinically defined patterns, even without explicit guidance towards a predefined pattern vocabulary [18]. Some research has proposed pattern-based approaches using learned templates for tasks like arrhythmia classification [19]. While transformer-based feature extraction has also been investigated for ECG signals [10], a comprehensive framework that explicitly tokenizes ECGs into a vocabulary of physiological patterns and integrates this with domain-guided attention mechanisms remains an area ripe for exploration. Our work seeks to bridge this gap by explicitly framing ECG interpretation through a pattern vocabulary lens, aiming to enhance both performance and the clinical relevance of explanations.

III. Methods

A. CardioPatternFormer Architecture Overview

The CardioPatternFormer is a novel transformer-based architecture specifically designed for interpretable multi-label ECG classification. Components are carefully engineered to incorporate cardiological domain knowledge while leveraging the representational power of transformers. Fig. 1 presents an overview of the revised architecture.

The CardioPatternFormer processes 12-lead ECG signals through several key stages:

1) **Input Processing:** A `CardiacPatternTokenizer` transforms continuous signals into a sequence of embedded representations based on learned, multi-scale cardiac patterns. This is augmented by a `Multi-Resolution Temporal Encoding` layer that captures temporal relationships across various time scales.
2) **Encoding:** A stack of Transformer encoder layers, utilizing specialized `Physiologically-Guided Attention` mechanisms (detailed in Sec III-C), processes the embedded sequence to capture complex dependencies and contextual information.
3) **Classification Heads:** The output of the encoder is fed into two parallel, complementary classification heads:
   - An `ExplainableDiagnosticHead` generates class-specific attention maps alongside diagnostic predictions, providing interpretability.
   - An `AdaptiveDiagnosticPooling` head uses a learnable relevance mechanism and uncertainty estimation for robust classification.
4) **Output Generation:** The logits from the two classification heads are fused using learnable weights (`diagnostic_fusion`) to produce the final multi-label classification predictions. Corresponding attention maps and relevance scores from the heads provide insights into the classification decisions.

B. Cardiac Pattern Tokenizer

Traditional approaches to ECG processing typically treat signals as continuous waveforms, applying generic convolutional operations without explicit consideration of underlying cardiac patterns. In contrast, our `CardiacPatternTokenizer` is inspired by the way

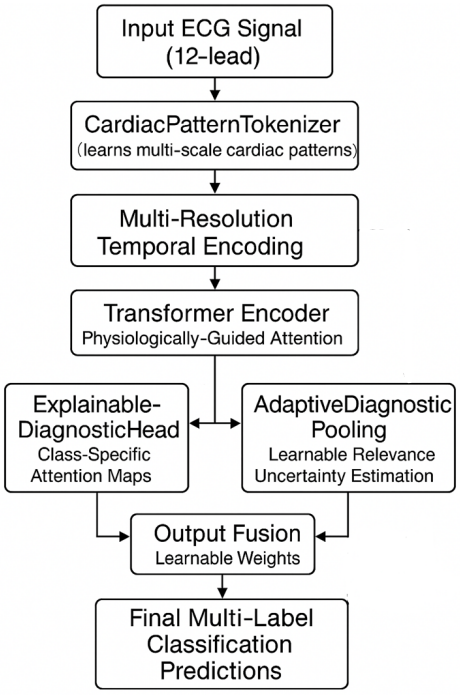

Fig. 1. Overview of the revised CardioPatternFormer architecture, illustrating the flow from input processing through encoding to the dual classification heads and output fusion.

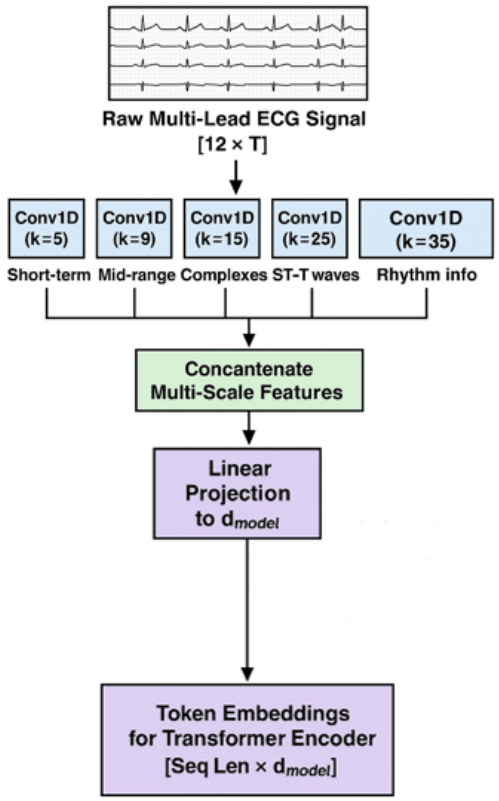

Fig. 2. Illustration of the multi-scale convolutional processing within the Cardiac Pattern Tokenizer.

cardiologists recognize distinct, meaningful patterns in ECG signals (e.g., P-waves, QRS complexes, ST-segments). The tokenizer employs multi-scale 1D convolutional layers (`nn.Conv1d`) with varying kernel sizes (e.g., 5, 9, 15, 25, 35) to detect patterns at different temporal resolutions, capturing both fine-grained morphological features and broader waveform characteristics. This multi-scale approach is critical for ECG interpretation, as diagnostic information exists at various temporal scales from narrow deflections to wider complexes and rhythm patterns spanning multiple heartbeats.

The features extracted from each convolutional pathway are concatenated and projected to the model's embedding dimension (d_{model}). Unlike initializing with predefined waveforms, the convolutional kernels are initialized using standard methods (e.g., Kaiming normal) and learn relevant patterns directly from the data. A standard learned positional embedding is added to the resulting sequence of token embeddings to retain temporal order information. The tokenizer effectively transforms the raw multi-lead ECG into a sequence of rich feature vectors ready for the transformer encoder.

C. Physiologically-Guided Attention

Standard transformer self-attention mechanisms treat all input tokens uniformly, potentially overlooking known physiological relationships within ECG signals. To address this, our `PhysiologicallyGuidedAttention` mechanism integrates cardiac domain knowledge directly into the attention score calculation, enhancing the model's focus on clinically relevant patterns.

Instead of imposing fixed biases, this mechanism incorporates several physiological priors whose influence is controlled by learnable parameters, allowing the model to adapt the importance of each constraint during training. These adaptable constraints include:

1) **Local Context Emphasis:** A relative positional bias is scaled by a learnable parameter (`local_bias_strength`). This encourages attention between temporally nearby tokens, reflecting the clinical principle that the interpretation of certain waveform segments (e.g., the ST segment or T-wave) often depends critically on the characteristics of the immediately preceding waveforms (e.g., the QRS complex), making local temporal context vital for morphological assessment.
2) **Rhythm-Aware Weighting:** A periodic bias (e.g., cosine function based on relative position) is scaled by learnable, per-head weights (`rhythm_constraint_weight`). This allows different attention heads to potentially focus on

different rhythmic patterns or periodicities relevant to arrhythmias or normal rhythms. Incorporating rhythm awareness is crucial as many cardiac conditions manifest primarily as deviations from normal sinus rhythm, requiring the model to effectively assess regularity and rate.

3) **Cardiac Cycle Awareness:** A small convolutional network (`cycle_detector`) processes the input embeddings to predict weights indicating the likelihood of specific cardiac cycle phases (e.g., QRS complex, T wave). These weights modulate the attention scores, scaled by another learnable parameter (`cycle_attn_strength`), guiding attention towards relevant phases. This allows the model to differentially weight information based on the specific phase of the P-QRS-T cycle, mimicking how clinicians focus on different waveform components to identify distinct abnormalities (e.g., P-waves for atrial issues, QRS for ventricular issues).
4) **Beat-to-Beat Consistency:** A convolutional layer (`conv_constraint`) operates on the attention patterns across the sequence length dimension, encouraging smoother or more consistent attention patterns between adjacent beats, scaled by a learnable `beat_weight_strength`. This encourages the model to recognize that for many stable conditions, waveform morphology is expected to be relatively consistent across consecutive beats, while abrupt changes in attention might signify transient events or noise.

These learnable physiological constraints are added to the standard scaled dot-product attention scores before the softmax operation. This allows the model to leverage prior knowledge about ECG structure and dynamics while retaining the flexibility to adjust the influence of these priors based on the data and the specific diagnostic task. Fig. 3 may illustrate the effect of these combined, adaptable constraints.

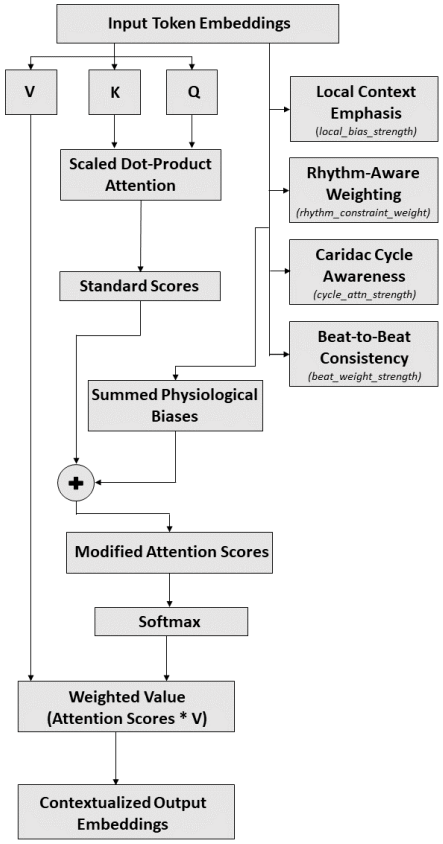

Fig. 3. Conceptual illustration of Physiologically-Guided Attention incorporating adaptable constraints, or an example of resulting attention patterns.

D. Multi-Resolution Temporal Encoding

ECG interpretation requires understanding temporal relationships at multiple scales, from millisecond-level wave morphologies to second-level rhythm patterns. Standard positional embeddings capture order but not necessarily scale. Our `MultiResolutionTemporalEncoding` addresses this by learning distinct temporal embeddings at multiple resolutions simultaneously (e.g., using scales of 1, 2, 5, 10, 20 relative to the input sequence length).

For each resolution scale, a separate learnable embedding parameter (`nn.Parameter`) is defined. During the forward pass, the appropriate length of each resolution's embedding is selected and then interpolated (using `F.interpolate` with linear mode) to match the actual input sequence length. These interpolated embeddings, representing temporal information at different scales, are concatenated along the feature dimension. A final linear projection layer (`self.projection`) maps the concatenated multi-resolution temporal features back to the model's embedding dimension. This projected temporal encoding is then added to the input token embeddings (after the `CardiacPatternTokenizer`), providing the transformer layers with rich, multi-scale temporal context. This allows the model to adaptively leverage fine-grained temporal information for morphological analysis and broader temporal information for rhythm analysis.

E. Explainable Diagnostic Classification

Following the transformer encoder stack, CardioPatternFormer employs two distinct classification heads in parallel, whose outputs are ultimately fused:

1) **`ExplainableDiagnosticHead`:** This head aims to provide interpretability alongside classification. For each of the `num_classes` diagnostic categories, it uses a sep- arate small attention network (`class_attentions`) to compute attention scores over the encoder's output sequence. These scores generate class-specific attention maps, indicating which parts of the ECG signal were most influential for predicting that spe- cific class. The weighted context vector for each class is then passed through a dedicated predictor (`class_predictors`) to generate the class logit.

This head returns both the logits and the corresponding attention maps (`explanation_maps`).

2) **`AdaptiveDiagnosticPooling`:** This head offers an alternative classification approach focused on relevance and uncertainty. It first uses a `relevance_detector` (a linear layer followed by a sigmoid) to calculate a relevance score for each time step in the encoder output sequence. These scores weight the importance of each time step's features. A weighted average pooling is performed, using the relevance scores as weights, to obtain a single context vector representing the most relevant parts of the signal. This pooled vector is fed through a `classifier` to produce diagnostic logits. Additionally, the pooled vector is passed to an `uncertainty_estimator` head to predict an uncertainty score (between 0 and 1) for each class prediction. This head returns logits, relevance maps (`relevance`), and uncertainty scores (`uncertainty`).

The final classification logits are obtained by combining the logits from the `ExplainableDiagnosticHead` and `AdaptiveDiagnosticPooling` head using learned scalar weights (`diagnostic_fusion`).

F. Loss Function

To train the model effectively for the multi-label classification task, we utilize a specialized loss function strategy that incorporates domain knowledge and addresses common challenges in medical datasets. The primary component is an **Enhanced Physiological Focal Loss** function. This builds upon the standard Focal Loss concept by focusing the training process on more challenging diagnostic examples that the model initially classifies poorly. Furthermore, it mitigates the issue of class imbalance, which is common in medical datasets, by applying weights that give more importance to underrepresented diagnostic categories. This classification loss can also optionally incorporate prior knowledge about the physiological relationships between different cardiac conditions, encouraging the model to make predictions consistent with known clinical co-occurrences.

In addition to the primary classification loss, we incorporate an optional **Attention Diversity Loss**. When used (controlled by a hyperparameter weight), this auxiliary loss encourages the explanation mechanisms within the model (specifically, the class-specific attention maps) to focus on distinct regions of the ECG signal for different diagnoses, where appropriate. This promotes more varied and potentially more insightful explanations for each predicted condition.

The overall training objective combines the primary classification loss and the optional attention diversity loss, using hyperparameters to balance their respective contributions. This carefully designed loss strategy guides the model towards learning accurate, robust, and interpretable ECG classifications.

IV. Experiments and Results

A. Dataset and Preprocessing

We evaluated CardioPatternFormer on the publicly available Chapman-Shaoxing ECG dataset, which contains 12-lead ECG recordings from 10,247 patients. Each recording is accompanied by cardiologist-annotated diagnostic labels conforming to the SCP-ECG standard. For model training and evaluation, we followed common practice by mapping the numerous specific diagnostic codes into six clinically relevant, broader categories:

1) Sinus Bradycardia
2) Sinus Rhythm and Tachycardia
3) Supraventricular Arrhythmias
4) Ventricular Arrhythmias and Conduction Blocks
5) ST-T Changes and Ischemic Changes
6) Structural Abnormalities and Miscellaneous

This grouping results in a multi-label classification task where each ECG can belong to one or more categories. The dataset exhibits a natural imbalance across these categories, as detailed in Table I, motivating the use of specialized loss functions described in Section III-F.

TABLE I
Dataset Characteristics and Diagnostic Category Distribution

Diagnostic Category	Prevalence (% or Count)
Sinus Bradycardia	33.07%
Sinus Rhythm and Tachycardia	33.12%
Supraventricular Arrhythmias	31.11%
Ventricular Arrhythmias and Conduction Blocks	14.24%
ST-T Changes and Ischemic Changes	29.29%
Structural Abnormalities and Miscellaneous	11.70%

The original ECG signals were recorded at 500 Hz, typically spanning 10 seconds (≈5,000 time points). Our preprocessing pipeline involved several steps applied to each recording:

- **Filtering:** A fourth-order Butterworth bandpass filter between 0.5 Hz and 45 Hz was applied to remove baseline wander and high-frequency noise.
- **Resampling:** Signals were downsampled from 500 Hz to a target frequency of 100 Hz, reducing the sequence length to approximately 1,000 time points while preserving clinically relevant frequencies.
- **Normalization:** Each of the 12 leads was independently normalized to have zero mean and unit variance.

This preprocessing resulted in 12-lead ECG tensors of shape (12, ≈1000) along with corresponding multi-label diagnostic vectors and calculated parameter vectors for input into the CardioPatternFormer model.

B. Experimental Setup

We implemented the CardioPatternFormer model using PyTorch. The model architecture utilized an embedding dimension (d_{model}) of 256, 8 attention heads

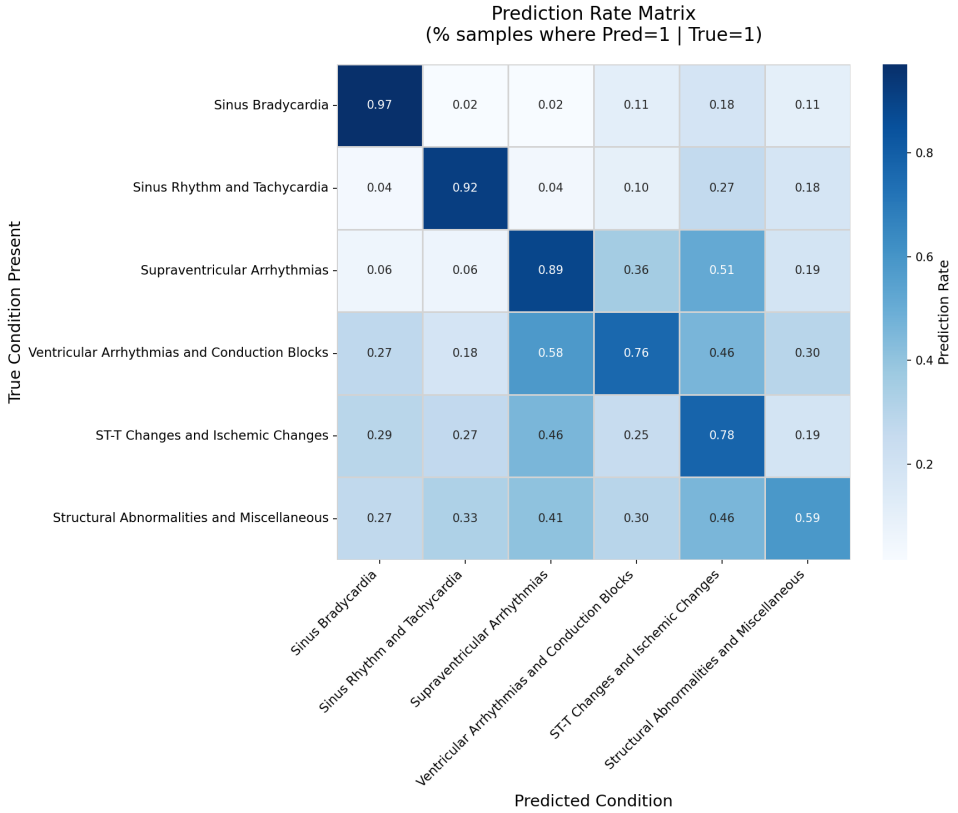

Fig. 4. Heatmap illustrating the Prediction Rate Matrix on the test set. Rows indicate the true condition present, columns indicate the predicted condition, and cell values represent the rate P(Pred=1 — True=1). The diagonal elements highlight the recall (sensitivity) achieved for each diagnostic category.

within the `PhysiologicallyGuidedAttention` layers, and a stack of 4 transformer encoder layers. The `CardiacPatternTokenizer` used a pattern vocabulary size of 16 (`num_patterns`).

The model was trained using the AdamW optimizer with an initial learning rate of 5×10^{-5} and a weight decay of 0.01. A `ReduceLROnPlateau` learning rate scheduler was employed, monitoring the validation macro F1-score, reducing the learning rate by a factor of 0.2 if no improvement was observed for 3 epochs. The primary classification loss function was the Enhanced Physiological Focal Loss (detailed in Section III-F), configured with parameters $\alpha = 0.5$ and $\gamma = 2.0$.

We utilized a 5-fold cross-validation strategy on approximately 85% of the full dataset designated for training and validation. The remaining ≈15% was held out as a final test set. Within each fold of the cross-validation, the data was split into training and validation sets. Training was performed for a maximum of 30 epochs, using a batch size of 16 and gradient accumulation over 2 steps, resulting in an effective batch size of 32. Automatic mixed-precision (AMP) training was enabled to optimize computational efficiency and memory usage. We implemented early stopping with a patience of 5 epochs based on the validation macro F1-score to prevent overfitting.

All reported performance metrics in Section IV-C onwards are based on the evaluation of the best model (selected based on the highest validation macro F1-score during cross-validation) on the held-out test set.

C. Classification Performance

CardioPatternFormer demonstrated strong multi-label classification performance on the held-out Chapman-Shaoxing test set. The model achieved an overall Hamming accuracy of 0.9184 and a macro F1-score of 0.8019. Table II summarizes the key macro-averaged performance metrics on the test set.

TABLE II
OVERALL PERFORMANCE ON CHAPMAN-SHAOXING TEST SET

Metric	Value
Accuracy (Hamming)	0.9184
Macro Precision	0.7896
Macro Recall	0.8184
Macro F1-Score	0.8019
Macro AUC	0.9437

The high macro AUC value (0.9437) indicates robust discriminative ability across all diagnostic categories, suggesting the model effectively distinguishes between positive and negative cases even with the dataset's inherent class imbalance.

Table III presents the detailed per-class performance metrics, evaluated using the optimal thresholds determined during

cross-validation. These results reveal variations in performance across the different diagnostic groups.

TABLE III
PER-CLASS PERFORMANCE METRICS ON TEST SET

Condition	Precision	Recall	F1 Score	AUC	Threshold
Sinus Bradycardia	0.9475	0.9705	0.9588	0.9950	0.396
Sinus Rhythm and Tachycardia	0.9639	0.9179	0.9404	0.9869	0.554
Supraventricular Arrhythmias	0.9408	0.8919	0.9157	0.9719	0.446
Ventricular Arrhythmias and Conduction Blocks	0.6691	0.7647	0.7137	0.9414	0.386
ST-T Changes and Ischemic Changes	0.7585	0.7785	0.7684	0.9190	0.416
Structural Abnormalities and Miscellaneous	0.4576	0.5870	0.5143	0.8481	0.406

Consistent with clinical intuition, the model performs exceptionally well on distinct rhythm categories like Sinus Bradycardia (F1=0.9588) and differentiating normal Sinus Rhythm/Tachycardia (F1=0.9404). Performance remains strong for Supraventricular Arrhythmias (F1=0.9157). As expected, performance is moderately lower for the more heterogeneous and potentially subtler categories of Ventricular Arrhythmias/Conduction Blocks (F1=0.7137) and ST-T Changes/Ischemic Changes (F1=0.7684). The lowest F1-score is observed for Structural Abnormalities/Miscellaneous (F1=0.5143), likely reflecting the combined challenge of diagnostic subtlety from ECG alone and lower prevalence in the dataset. Despite these variations, the per-class AUC values remain high ($\geq$0.84 for all classes), indicating good underlying discriminative capability across all conditions. This performance gradient generally aligns with clinical experience regarding the relative difficulty of diagnosing these conditions solely from ECG signals.

To contextualize CardioPatternFormer's performance, Table IV compares its results with selected studies utilizing the Chapman-Shaoxing dataset. However, readers should note that direct comparison is inherently challenging due to significant variations in task definitions (e.g., specific arrhythmias vs. broad categories), reported metrics, and evaluation protocols across different publications. Therefore, the table serves primarily as a contextual reference rather than a direct leaderboard.

D. Interpretability Analysis

A key advantage of the CardioPatternFormer architecture is its capacity for interpretability, primarily facilitated by the `ExplainableDiagnosticHead`. This component generates class-specific attention maps (`explanation_maps`) alongside its predictions. These maps are designed to highlight the temporal regions within the input ECG signal that were most influential in reaching a specific diagnostic conclusion.

Fig. 5 provides an example of these class-specific attention maps.

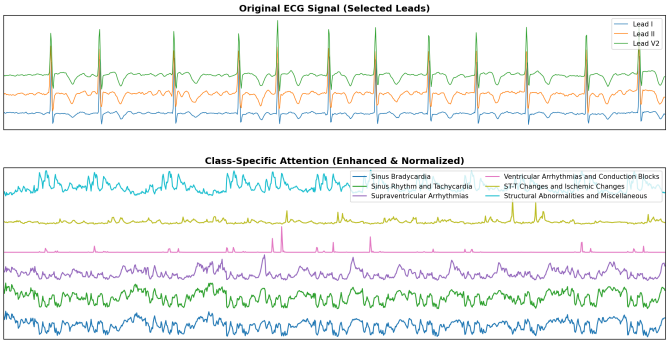

Fig. 5. Example class-specific attention maps from the `ExplainableDiagnosticHead` for a representative test sample. Top panel: Selected input ECG leads (e.g., I, II, V2). Bottom panel: Corresponding normalized attention weights over time for different diagnostic classes, illustrating the model's temporal focus for each potential diagnosis.

By visualizing these attention scores, typically overlaid on the ECG signal (as shown in Fig. 5), it is possible to gain insights into the model's decision-making process. The goal is for these highlighted regions to correspond to physiologically relevant waveforms or patterns associated with the predicted condition. This mechanism aims to enhance the model's clinical utility by providing explanations that can be evaluated against established electrophysiological knowledge, thereby fostering trust and potentially aiding clinical review.

Furthermore, comparing attention patterns across different predicted diagnoses can illustrate how the model potentially uses distinct ECG features to differentiate between various conditions, which is essential for accurate multi-label classification.

E. Analysis of Complex Cases

Real-world ECG interpretation often involves patients with multiple simultaneous cardiac conditions. To assess CardioPatternFormer's performance in such challenging scenarios, we examined its predictions and explanations on test cases exhibiting multiple ground-truth diagnoses. Fig. 6 presents a representative 12-lead ECG example from such a complex case. The detailed multi-label classification results for this specific case are provided in Table V, while the corresponding average attention map, highlighting the most influential signal regions for the model's overall prediction, is shown in Fig. 7.

In representative examples, the model demonstrated the ability to identify several co-occurring conditions correctly. Examination of the corresponding class-specific attention or relevance maps suggests the model often leverages distinct signal features or time windows when predicting different concurrent diagnoses, aligning with the expectation that different conditions may have unique ECG manifestations even when

TABLE IV
SIMPLIFIED PERFORMANCE COMPARISON WITH SELECTED RELATED WORKS (CHAPMAN-SHAOXING DATASET CONTEXT).

Model	Architecture (Brief)	Task Focus (Leads, Classes / Type)	Accuracy Results (% or Score)
CardioPatternFormer	**Transformer (Pattern Tok. + Guided Attn)**	**12L / 6 Broad Multi-label Cats.**	**91.84 (Hamming Acc.)**
Bimodal CNN	Dual Inception CNN (Img + Scalogram)	12L / 11 Rhythms	≈95.7 (Accuracy)
CNN-BiLSTM-BiGRU	CNN+BiLSTM+BiGRU+Attn	Lead II / 7 Rhythms	≈98.6 (Accuracy)
Evo CNN Trees	Evo. CNN-Tree Fusion	12L / 11 Rhythms	≈97.6 (Avg. Accuracy)
CNN Teacher+Student	CNN (Knowledge Distill.)	12L / 11 Rhythms	97.55 (Accuracy - Student)
CMA Classifier	Custom Feature-Image Classif.	12L / 5 Rhythms	99.76 (Accuracy)
GCN-WMI	Graph ConvNet (15-layer)	12L / Mult. Rhythms	99.82 (Accuracy)

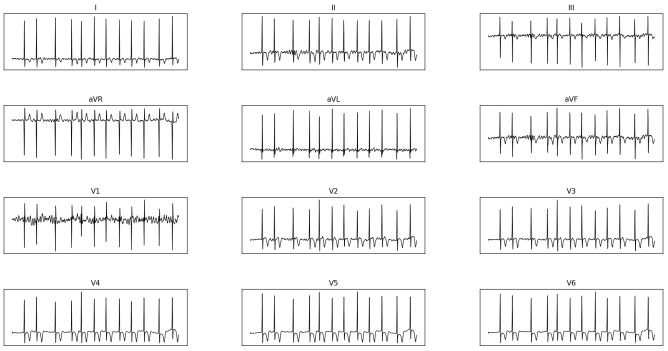

Fig. 6. A representative 12-lead ECG signal from a patient presenting with multiple simultaneous cardiac conditions. The signal spans 10 seconds and is sampled at 100 Hz.

TABLE V
MULTI-LABEL CLASSIFICATION RESULTS FOR A COMPLEX ECG CASE WITH MULTIPLE CO-OCCURRING CONDITIONS.

Cardiac Condition	Ground Truth	Prediction	Probability
Sinus Bradycardia			0.028
Sinus Rhythm and Tachycardia			0.066
Supraventricular Arrhythmias	✓	✓	0.924
Ventricular Arrhythmias and Conduction Blocks			0.407
ST-T Changes and Ischemic Changes	✓	✓	0.729
Structural Abnormalities and Miscellaneous			0.295

present simultaneously. This capability is crucial for providing comprehensive and clinically useful interpretations in complex cases.

F. Lead Ablation Study

To evaluate the model's robustness and the relative importance of different ECG leads for classification performance, we conducted a lead ablation study. Using the best-performing model checkpoint obtained from cross-validation, we systematically evaluated its performance on the test set while providing only specific subsets of the original 12 leads as input. Unavailable leads were masked by zero-padding their corresponding input channels before entering the `CardiacPatternTokenizer`. We tested various configurations, including all single leads, standard clinical subsets (e.g., limb leads only, precordial leads only), and other combinations. Performance was measured using macro F1-score and AUC for classification, applying the optimal thresholds determined previously.

The results demonstrate a clear dependency on the number of available leads for optimal classification performance. Performance degrades significantly when using fewer leads compared to the full 12-lead configuration. Table VI summarizes the macro F1-scores achieved for representative lead subsets.

TABLE VI
CARDIOPATTERNFORMER CLASSIFICATION PERFORMANCE (MACRO F1-SCORE) WITH REDUCED LEAD SETS

Lead Configuration	Leads Included	Macro F1-Score
Single Lead (Best)	II	0.3602
Single Lead (Worst)	V3	0.2828
Average Single Lead (approx)	*(Range: ≈0.28 - 0.36)*	*(≈0.34)*
Limb Leads Only	I, II, III, aVR, aVL, aVF	0.6122
Precordial Leads Only	V1, V2, V3, V4, V5, V6	0.5809
Example Reduced Set (4 Leads)	I, II, V1, V5	0.5639
Full 12-Lead Set	All	**0.8019**

As shown in Table VI, while single leads provide limited diagnostic information (macro F1 scores ranging from approximately 0.28 to 0.36), combining leads significantly improves performance. Standard 6-lead configurations (limb or precordial) achieve intermediate results, while the full 12-lead set provides substantially better overall classification accuracy. Among the single leads, Lead II achieved the highest individual macro F1-score in our test, aligning with its common use in rhythm assessment, although performance with any single lead was considerably lower than multi-lead configurations. Analysis also revealed variations in the importance of specific leads for classifying different diagnostic categories.

From this analysis, we observe the model's adaptability to varying lead inputs, confirming the expected performance

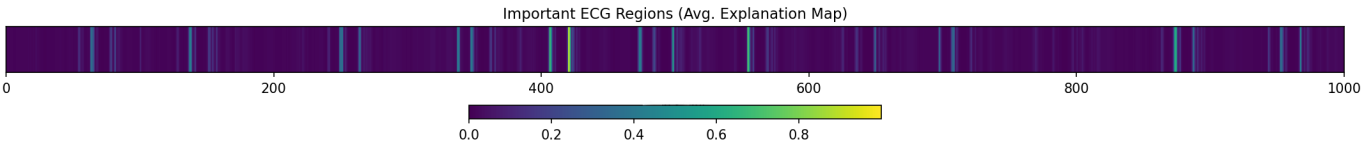

Fig. 7. Average attention map of CardioPatternFormer for the ECG signal in Fig. 6, highlighting the regions of increased model focus. The color intensity represents the average attention weight given by the model across all detected patterns, with brighter colors (yellow/green) indicating higher attention.

trade-offs, and gain quantitative metrics regarding the importance of individual leads for the multi-label classification task.

V. Discussion

CardioPatternFormer represents a significant step towards automated ECG interpretation systems that are not only accurate but also interpretable and aligned with clinical reasoning. By framing ECG analysis through the lens of pattern recognition and developing specialized architectural components like the `CardiacPatternTokenizer`, `MultiResolutionTemporalEncoding`, and particularly the `PhysiologicallyGuidedAttention` with learnable constraints, our approach aims to bridge the gap between black-box deep learning models and the nuanced process of clinical diagnosis. The model demonstrated strong classification performance on the challenging multi-label Chapman-Shaoxing dataset, achieving a macro F1-score of 0.8019 and a macro AUC of 0.9437 on the held-out test set. Analysis of per-class performance revealed a clinically plausible gradient: the model excelled at identifying conditions with distinct ECG patterns like Sinus Bradycardia (F1=0.9588) and Sinus Rhythm/Tachycardia (F1=0.9404), while performance was lower, yet still substantial (AUC $\geq$ 0.84), for more heterogeneous or subtle categories like Structural Abnormalities/Miscellaneous (F1=0.5143). This suggests the model learns meaningful diagnostic representations.

A key contribution of this work is the enhanced interpretability offered through class-specific attention maps generated by the `ExplainableDiagnosticHead`. Our qualitative analysis indicated that the model often focuses on physiologically relevant ECG regions corresponding to the predicted diagnosis, such as specific waveform segments or intervals pertinent to different conditions. This capability, visualized through attention maps, provides valuable transparency into the model's decision process, addressing a major limitation of previous deep learning approaches and potentially fostering greater clinical trust and utility. Furthermore, the inclusion of an `AdaptiveDiagnosticPooling` head provides complementary diagnostic information, including uncertainty estimates that correlate with prediction accuracy and could help prioritize challenging cases for expert review in clinical workflows. The overall pattern-based approach, combining multi-scale feature extraction with attention mechanisms that adaptively incorporate domain knowledge, appears effective for handling complex multi-label ECG classification.

Despite promising results, several limitations should be acknowledged. The current architecture relies on fixed-length inputs ($\approx$1000 time points after downsampling), which might limit its ability to capture very long-term dependencies relevant to certain intermittent arrhythmias. While the Chapman-Shaoxing dataset is relatively large, generalizability to patient populations with different demographics or ECG acquisition protocols requires further investigation. Class imbalance remains a challenge; although mitigated by the specialized loss function, performance on underrepresented conditions is still lower, highlighting the need for larger, more diverse datasets. Furthermore, a significant limitation in this study relates to the auxiliary task of physiological parameter prediction; the automated method used for calculating ground-truth parameters (particularly PR, QRS, QTc intervals) during data preprocessing was insufficient, leading to unreliable parameter prediction results. Future work necessitates implementing robust ECG delineation algorithms for accurate ground-truth parameter extraction before this component can be meaningfully evaluated. Our evaluation focused on classification accuracy and interpretability visualizations; comprehensive clinical validation, including usability studies and comparisons with existing decision support tools, is essential before deployment. Finally, while attention maps indicate *where* the model focuses, they don't fully elucidate the *physiological reasoning*, representing an ongoing challenge in deep learning explainability.

There are several interesting directions that could be explored in future work. One idea is to improve the `CardiacPatternTokenizer` so it can learn a more structured or layered vocabulary of cardiac patterns, which might help make the model both more accurate and easier to interpret. Another area worth looking into is how to better capture long-range temporal relationships, especially since this could be important for analyzing more complex heart rhythms. It's also important to address the current challenges in parameter prediction, possibly by developing better ways to calculate ground-truth values. Beyond technical improvements, running clinical evaluation studies would be a key step in understanding how CardioPatternFormer might actually perform in real-world healthcare settings. Lastly, combining ECG analysis with other types of clinical data like patient history, lab tests, or imaging could make the system more useful in supporting broader diagnostic decisions.

VI. Conclusion

CardioPatternFormer offers a promising approach to interpretable, automated ECG classification by leveraging a pattern-recognition perspective and integrating domain knowledge through specialized transformer components. Our re-

sults demonstrate strong classification performance across a diverse set of cardiac conditions on the Chapman-Shaoxing dataset, with enhanced transparency provided by class-specific attention mechanisms that highlight physiologically relevant signal regions. The ability to generate explanations aligned with clinical reasoning, alongside robust classification, marks progress towards AI tools that can function synergistically with clinicians. While further development and validation are essential, particularly concerning robust physiological parameter extraction and prospective clinical studies, CardioPatternFormer establishes a valuable framework for advancing ECG analysis. For AI tools like CardioPatternFormer to be successfully integrated into clinical practice, prioritizing transparency and alignment with clinical workflows, as we have aimed to do, is crucial. Such interpretable systems have the potential to foster clinician trust and enable effective human-AI collaboration, ultimately leading to improved patient care.

REFERENCES

[1] Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P., & Ng, A.Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1), 65-69.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.

[3] Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H., & Rakov, R. (2020). A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1), 48.

[4] Ribeiro, A.H., Ribeiro, M.H., Paixão, G.M.M., Oliveira, D.M., Gomes, P.R., Canazart, J.A., Ferreira, M.P.S., Pires, T.H.C., Samesima, N., Pastore, C.A., Celi, L.A., & Ribeiro, A.L.P. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1), 1760.

[5] Attia, Z.I., Kapa, S., Lopez-Jimenez, F., McKie, P.M., Ladewig, D.J., Satam, G., Pellikka, P.A., Enriquez-Sarano, M., Noseworthy, P.A., Munger, T.M., Asirvatham, S.J., Scott, C.G., Carter, R.E., & Friedman, P.A. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature Medicine*, 25(1), 70-74.

[6] Faust, O., Hagiwara, Y., Hong, T.J., Lih, O.S., & Acharya, U.R. (2018). Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine*, 161, 1-13.

[7] Murat, F., Yildirim, O., Talo, M., Baloglu, U.B., Demir, Y., & Acharya, U.R. (2020). Application of deep learning techniques for heartbeats detection using ECG signals. *Arabian Journal for Science and Engineering*, 45(12), 10413-10425.

[8] Schwab, P., Scebba, G., Zhang, J., Delai, M., & Karlen, W. (2017, September). Beat by beat: Classifying cardiac arrhythmias with recurrent neural networks. In 2017 *Computing in Cardiology (CinC)* (pp. 1-4). IEEE.

[9] Sarkar, I., & Etemad, A. (2020). Self-Supervised Learning for ECG Signal Analysis: A Review. *Frontiers in Digital Health*, 2, 13.

[10] Natarajan, A., Kannan, A., Vepakomma, P., & Raskar, R. (2020, September). ECG-transformer: A robust model for detecting cardiac arrhythmias from noisy ECGs. In 2020 *Computing in Cardiology (CinC)* (pp. 1-4). IEEE.

[11] Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., & Long, M. (2023). TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.

[12] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618-626).

[13] Yao, L., Sheng, X., Ouyang, C., Li, Z., Wang, X., & Zhang, S. (2020). Interpretable deep learning for automatic detection of atrial fibrillation from electrocardiograms. *Journal of Electrocardiology*, 60, 1-7.

[14] Shashikumar, S.P., Aneja, A., Nguyen, K., Clifford, G.D., & Nemati, S. (2020). Interpretable deep learning for diagnosis of 12-lead electrocardiograms. *Physiological Measurement*, 41(11), 115001.

[15] Hong, S., Zhou, Y., Shang, J., Xiao, C., & Sun, J. (2020). Opportunities and challenges in deep learning for health. *Frontiers in Big Data*, 3, 24.

[16] Choi, E., Bahadori, M.T., Searles, E., Coffey, C., Thompson, M., Bost, J., ... & Sun, J. (2016). RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems*, 29, 3504-3512.

[17] Hampton, J.R. (2013). *The ECG Made Easy* (8th ed.). Churchill Livingstone.

[18] Strodthoff, N., & Strodthoff, C. (2019). Detecting and interpreting myocardial infarction using fully convolutional neural networks. *Physiological Measurement*, 40(1), 015004.

[19] Zhang, Q., Li, Y., Sung, H.G., & Liu, C.C. (2019). Learning interpretable features for CPSC 2018 challenge: A case study on atrial fibrillation detection. *Frontiers in Physiology*, 10, 252.

[20] Benjamin, E.J., et al. (2023). Heart Disease and Stroke Statistics—2023 Update: A Report from the American Heart Association. Circulation, 147(5), e153-e639.