Explainable Deep Learning for ECG Arrhythmia Classification

Abstract

This paper presents an end-to-end Explainable Artificial Intelligence (XAI) framework for ECG arrhythmia classification. The system integrates signal preprocessing, beat segmentation, convolutional neural network (CNN) classification, and Integrated Gradients explainability. The proposed framework provides interpretable predictions and highlights clinically relevant waveform regions such as the QRS complex.

I. Introduction

Electrocardiogram (ECG) analysis plays a vital role in cardiovascular disease detection. Deep learning models achieve high classification accuracy but are often criticized for their black-box nature. This work addresses interpretability by integrating attribution-based explanation techniques into ECG classification.

II. Dataset Description

The system utilizes the MIT-BIH Arrhythmia Database. Each record contains raw ECG signals sampled at 360 Hz. Beat annotations provide R-peak locations and beat types. A total of 48 records were processed, resulting in over 111,000 segmented beats.

III. Methodology

A. Beat Segmentation: Each beat is extracted around the R-peak using a fixed window. Start = R_peak – 72, End = R_peak + 144, producing 216-sample beats.

B. Normalization: Each beat is standardized using z-score normalization: $x_{norm} = (x - \mu) / \sigma$.

C. CNN Architecture: The model consists of two Conv1D layers with ReLU activation and MaxPooling, followed by fully connected layers for binary classification.

Architecture Flow: Input → Conv1D → ReLU → MaxPool → Conv1D → ReLU → MaxPool → Fully Connected → Softmax.

IV. Mathematical Formulation

Loss Function: Cross-Entropy Loss is defined as $L = - \Sigma\, y_i \log(p_i)$.

Integrated Gradients Attribution: $IG_i(x) = (x_i - x'_i) \times \int_{\blacksquare}^{1} \partial F(x' + \alpha(x - x')) / \partial x_i \, d\alpha$.

V. System Flowchart

Raw ECG → Preprocessing → Beat Segmentation → CNN Classification → Integrated Gradients → Visualization.

VI. Performance Metrics

The model performance is evaluated using Accuracy, Precision, Recall, and F1-score. Training loss decreased steadily across epochs, demonstrating stable convergence.

Accuracy = (TP + TN) / (TP + TN + FP + FN).

F1-score = 2 × (Precision × Recall) / (Precision + Recall).

VII. Explainability Analysis

Integrated Gradients highlights waveform regions contributing most to predictions. Attribution maps show increased importance around QRS complexes in abnormal classifications, aligning with clinical interpretation.

VIII. Conclusion

The proposed ECG-XAI framework provides both high-performance arrhythmia detection and clinically interpretable explanations. The integration of CNN classification and Integrated Gradients offers a transparent and trustworthy diagnostic support system.