Problem 9: ECG Arrhythmia Multi-Label Classification

Dataset Information

• **Name:** CPSC 2018 Dataset (China Physiological Signal Challenge) • **Download Link:** https://physionet.org/content/challenge-2020/1.0.1/ • **Primary Repository:** https://github.com/physionetchallenges/python-classifier-2020 • **Paper:** https://www.frontiersin.org/articles/10.3389/fphys.2021.678597/full • **Size:** 6,877 clinical 12-lead ECG recordings

- **Patients:** Multi-institutional dataset from China
- **Sampling Rate:** 500 Hz with 16-bit resolution
- **Duration:** Variable length recordings (6 seconds to 60 seconds)
- **Labels:** 9 distinct cardiac arrhythmia classes with multi-label annotations • **Multi-label Cases:** 476 recordings have 2-3 simultaneous cardiac conditions • **Release Date:** 2018 (PhysioNet Challenge 2020)

Preprocessed Dataset Information

- **Preprocessed Version:** Available on Figshare
- **Download Link:** https://figshare.com/articles/dataset/ECG_data/
- **Preprocessing Repository:** https://github.com/antonior92/physionet-12ecg-classification
- **Resampling:** Standardized to 250 Hz sampling rate
- **Duration:** Normalized to 60 seconds (T=15,000 data points)
- **Format:** Ready-to-use preprocessed signals for deep learning

Problem Statement

Title: Multi-Label Cardiac Arrhythmia Classification from Variable-Length 12-Lead ECG Signals

Background: Cardiac arrhythmias often co-occur in clinical practice, making multi-label classification essential for comprehensive ECG interpretation. The CPSC 2018 dataset represents real-world clinical scenarios where patients may exhibit multiple simultaneous cardiac conditions, challenging traditional single-label classification approaches.

Dataset: 6,877 clinical 12-lead ECG recordings with 9 cardiac arrhythmia classes including Atrial Fibrillation (AF), Left Bundle Branch Block (LBBB), Normal Heartbeat, Premature Atrial Contraction (PAC), Premature Ventricular Contraction (PVC), Right Bundle Branch Block (RBBB), ST-segment Depression (STD), ST-segment Elevation (STE), and additional cardiac states. Multi-label annotations capture the clinical reality of co-occurring conditions.

Task: Develop a robust deep learning model capable of handling variable-length ECG signals (6-60 seconds) and predicting multiple simultaneous cardiac arrhythmias. The model
must address temporal variability, class imbalance, and multi-label dependencies while

maintaining clinical interpretability.

Input Data Format:

- **12-lead ECG signals:** (I, II, III, AVL, AVR, AVF, V1-V6)
- **Duration:** Variable (6-60 seconds) or preprocessed (60 seconds)
- **Sampling rate:** 500 Hz (original) or 250 Hz (preprocessed)
- **Format:** WFDB (WaveForm DataBase) format
- **Labels:** Multi-label with 9 cardiac arrhythmia classes
- **Arrhythmia Classes:** AF (Atrial Fibrillation), LBBB (Left Bundle Branch Block), Normal, PAC (Premature Atrial Contraction), PVC (Premature Ventricular Contraction), RBBB (Right Bundle Branch Block), STD (ST-segment Depression), STE (ST-segment Elevation), Others
- **Multi-label Statistics:** 476/6,877 recordings (6.9%) have multiple simultaneous labels

Evaluation Criteria

ECG Signal Data Understanding - EDA with comprehensive signal visualization across all 12 leads, Variable-length signal analysis and duration distribution, Lead-wise correlation and morphology analysis, Multi-label co-occurrence pattern analysis (476 multi-label cases), Class distribution and imbalance assessment across 9 arrhythmia types

Signal Preprocessing & Augmentation - Proper handling of variable-length signals (padding/truncation strategies), Patient-wise data splitting to prevent data leakage, ECG-specific preprocessing (baseline correction, noise filtering), Advanced augmentation techniques for multi-label scenarios, Comparison between original (500Hz) and preprocessed (250Hz) versions

Deep Learning Architecture & Performance - Multi-label classification metrics (Hamming Loss, Subset Accuracy, F1-macro/micro), Per-class AUC-ROC for all 9 arrhythmia types, Variable-length sequence handling (RNN/LSTM, attention mechanisms), 1D CNN architectures optimized for multi-label ECG classification, Ensemble methods for improved multi-label performance

Multi-Label Learning & Validation - Multi-label loss functions (Binary Cross-Entropy, Focal Loss), Class imbalance handling for minority arrhythmia classes, Multi-label stratification for train/validation/test splits, Cross-validation strategy accounting for multi-label dependencies, Threshold optimization for binary predictions per class

Detailed Presentation Approach & Technical Deep Dive - Final presentation

Optional (if time permits) - Real-time ECG monitoring dashboard with multi-label predictions, Integration with clinical decision support systems
Bonus - Advanced architectures (Transformers, self-attention), Clinical validation, Real-time ECG monitoring systems