ECG Arrhythmia Classification

Inter-Patient Generalization in Heartbeat Classification

Capstone Project Report

A 1D convolutional neural network that classifies individual heartbeats into four clinical categories, and a rigorous study of the inter-patient generalization gap in which ~96% validation performance falls to 44% on previously unseen patients.

Arel · RLcapstone.ai

1. Summary

This project classifies individual heartbeats from an electrocardiogram (ECG) into four AAMI categories — Normal, Supraventricular, Ventricular, and Fusion — using a **1D convolutional neural network**. Its central contribution is methodological: a rigorous demonstration of the **inter-patient generalization gap**.

During training, validation macro-recall climbs above **96%** and appears excellent. Evaluated honestly on patients the model has **never seen**, however, it attains **70% accuracy** and only **44% macro-recall** — strong on ventricular beats (86%) but nearly blind to supraventricular beats (9%). Explaining that gap is the point of the study.

> Not a medical device. This is an educational study of biomedical signal classification on a public research dataset. It does not diagnose cardiac conditions and must not inform any medical decision.

2. Motivation

Arrhythmia classification from the ECG is a canonical biomedical machine-learning problem. The objective here was not only to build a classifier but to evaluate it **correctly** — a great deal of published and hobbyist work on this dataset inadvertently leaks patient identity across the train/test split and reports inflated performance as a result.

3. Dataset and preprocessing

The project uses the **MIT-BIH Arrhythmia Database**: 48 half-hour, two-channel records with a cardiologist's annotation on every beat (~100,000 beats total). The pipeline reads the standard MLII lead, extracts a 260-sample (~0.7 s) window centered on each beat, normalizes it, and maps the fine-grained annotation symbols into the four AAMI super-classes.

> The single most consequential design decision is the train/test split, which is performed **by patient, not by beat**. A by-beat split lets one patient's heartbeats appear in both partitions, so the model learns to recognize individuals rather than arrhythmias — precisely the failure mode examined here.

4. Model and metric

The classifier is a compact **1D CNN** (four convolutional blocks, 77k parameters, 312 KB). Because roughly 90% of beats are normal, a trivial "always normal" model attains 90% accuracy while detecting none of the abnormal beats that matter. The primary metric is therefore **macro-recall** — the unweighted mean of the per-class recall rates — which improves only when the model handles the rare, clinically important classes. Class weighting was applied during training for the same reason.

5. Results

Performance on the 22 entirely unseen test patients:

Beat class	Recall	Note
Normal	72%	The dominant class
Supraventricular	9%	Almost entirely missed
Ventricular	86%	Detected reliably
Fusion	9%	Rare and difficult

This corresponds to 70% overall accuracy but only **44% macro-recall** once all four classes are weighted equally. The gap between the ~96% validation curve and this honest figure is the central finding of the project.

Why the model fails on supraventricular beats

The per-class breakdown is diagnostic. A ventricular beat has a distinctly abnormal waveform **morphology**, which a shape-sensitive model detects readily. A supraventricular beat is nearly identical in shape to a normal beat; its defining feature is **timing** — it occurs prematurely. Because the model processes each beat in isolation, with no representation of rhythm, it is structurally blind to the one feature that distinguishes the class it fails on. This is an honest limitation of the chosen design, not a bug.

6. Deployment

The model was exported to ONNX (a 28 KB file) and runs directly in the browser demonstration, with a matching TensorFlow Lite build in the Android application. Both reproduce the original model's predictions. In the browser demonstration the user can select a real heartbeat and watch it classified on their own device.

7. Limitations

- Not a medical device — an educational study on a research dataset.
- 44% macro-recall is not usable performance — demonstrating this honestly is the purpose of the project.
- Real-world ECGs are noisier and more varied than the clean research recordings used here.

8. Future work (v2)

The remedy follows directly from the diagnosis: provide the model with timing information. Adding **RR-interval features** — the intervals to the preceding and following beats — would let the model perceive prematurity, the hallmark of supraventricular beats. A full confusion matrix will accompany the v2 results.