ProtoECGNet: Case-Based Interpretable Deep Learning for Multi-Label ECG Classification with Contrastive Learning

THE UNIVERSITY OF CHICAGO
UChicago Medicine
Pritzker School of Medicine

THE UNIVERSITY OF CHICAGO
CENTER FOR COMPUTATIONAL MEDICINE & CLINICAL ARTIFICIAL INTELLIGENCE

ML4HC MACHINE LEARNING FOR HEALTH CARE

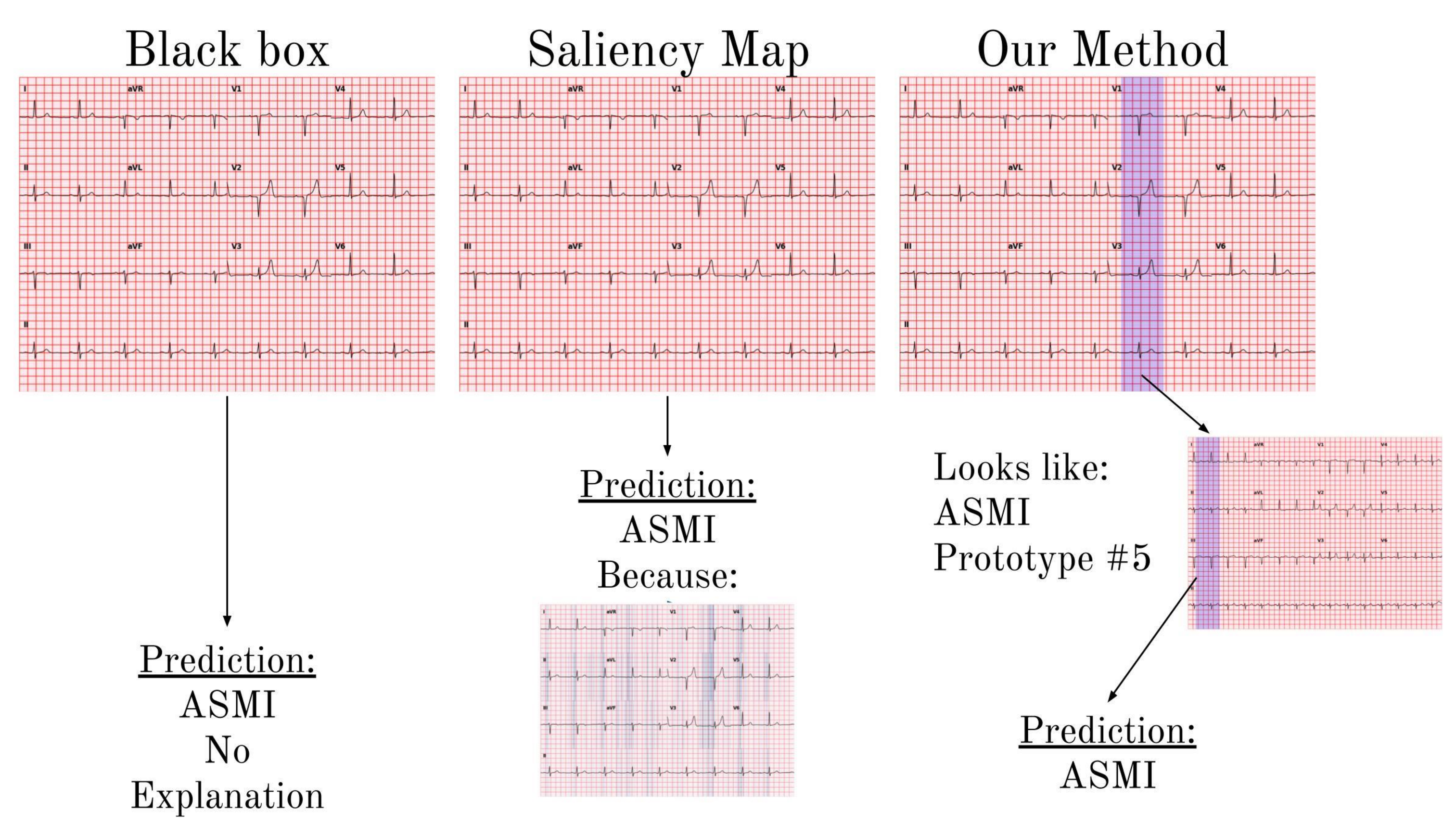
Sahil Sethi[1,2], David Chen[2], Thomas Statchen[1,2], Michael C. Burkhart[2], Nipun Bhandari[3], Bashar Ramadan[4], Brett Beaulieu-Jones[2]

[1]UChicago Pritzker School of Medicine, [2]UChicago Center for Computational Medicine & Clinical AI, [3]UC Davis Division of Cardiovascular Medicine, [4]UChicago Medicine Section of Hospital Medicine

Motivation

- Transparency in model predictions is essential for clinical adoption
- Post-hoc explainability methods are often not faithful to a model's reasoning
- For ECG model explanations to be useful, they should be aligned with how clinicians reason, and cover the full diagnostic spectrum of ECG interpretation

ProtoECGNet provides case-based explanations, tailored to the type of visual reasoning used by clinicians for each diagnosis, that are faithful to its internal reasoning process

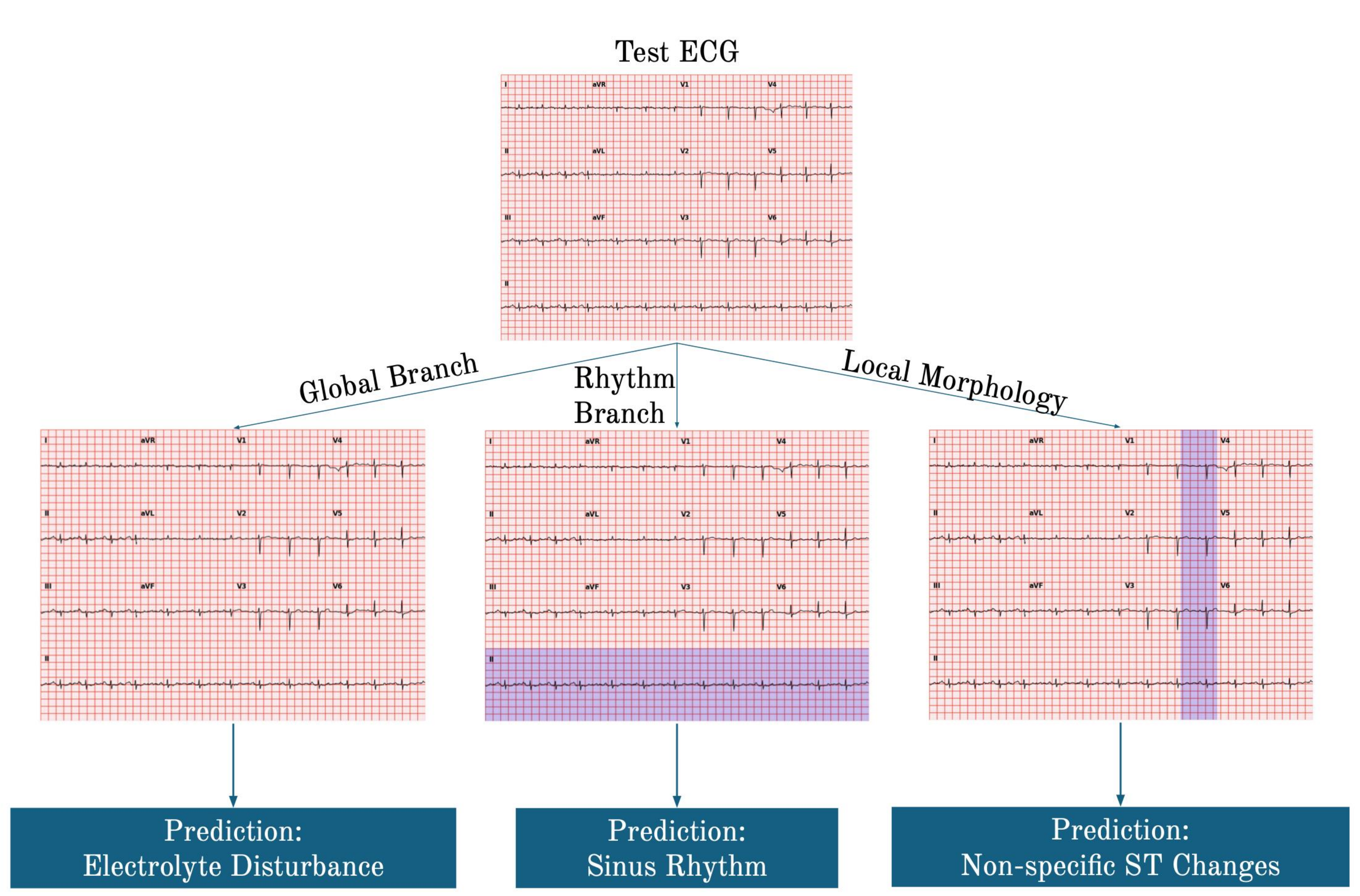


ASMI = anteroseptal myocardial infarction

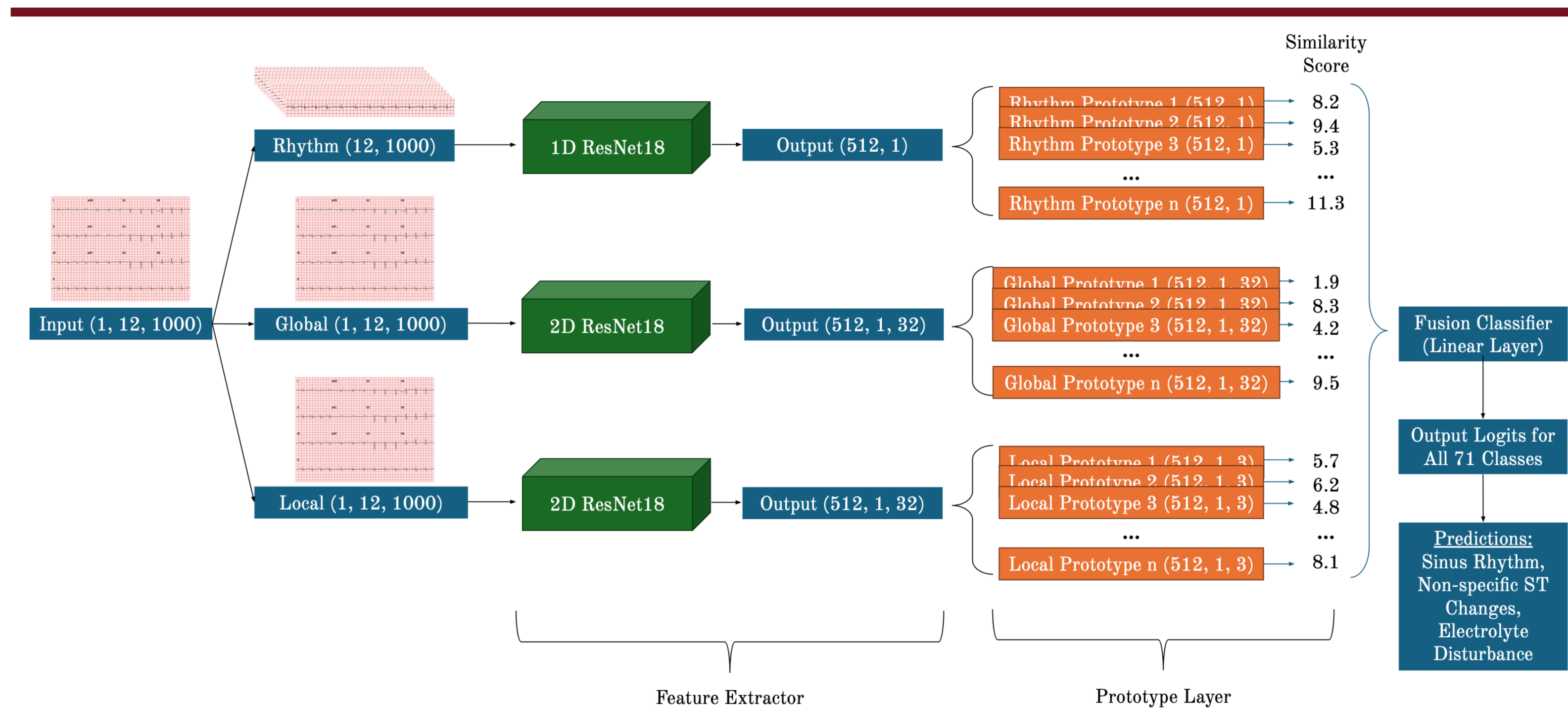
Multi-Branch Approach

We grouped the 71 labels from PTB-XL into three prototype categories based on the type of visual reasoning required for diagnosis:

1. **Rhythm-based diagnoses**—require temporal pattern analysis across full-length ECG signals, often discernible from a single lead
2. **Morphology-based diagnoses**—require localized waveform shape or inter-lead comparisons over short time intervals
3. **Global diagnoses**—require full-lead patterns spanning the full ECG duration



Internal Architecture



Contrastive Loss

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \lambda_{\text{clst}} \cdot \mathcal{L}_{\text{clst}} + \lambda_{\text{sep}} \cdot \mathcal{L}_{\text{sep}} + \lambda_{\text{div}} \cdot \mathcal{L}_{\text{div}} + \lambda_{\text{cntrst}} \cdot \mathcal{L}_{\text{cntrst}}$$

$$\mathcal{L}_{\text{cntrst}} = -\frac{1}{\sqrt{P}} \left(\frac{\sum_{i,j} C_{ij} \cdot \mathrm{S}(p_i, p_j)}{\sum_{i,j} C_{ij}} - \frac{\sum_{i,j} (1 - C_{ij}) \cdot \mathrm{S}(p_i, p_j)}{\sum_{i,j} (1 - C_{ij})} \right)$$

Performance on PTB-XL (71 labels)

The final multi-branch, contrastive model outperforms non-contrastive and single-branch variants, and matches the best black box benchmark

Table 1: Macro-AUROC across branch-specific, single-branch, and multi-branch settings for ProtoECGNet.

Setting	Model (Label Set)	Black-box	No Contrastive	w/ Contrastive
Branch-Specific Labels	Rhythm Branch (16 labels)	0.9403	0.8903	0.9064
	Morphology Branch (52 labels)	0.8872	0.8533	0.9051
	Global Branch (3 labels)	0.8649	0.8362	0.8667
Full 71-Label, Single Branch	1D Prototype Model	**0.9250**	0.8646	0.8977
	2D Partial Prototype Model	N/A	0.8873	0.9091
	2D Global Prototype Model	0.8990	0.8681	0.9074
Full 71-Label, Multi-Branch	Macro Aggregation	0.8982	0.8609	0.9038
	Fusion Classifier	N/A	0.8855	**0.9132**

Table 2: Weighted AUROC with bootstrapped 95% confidence intervals across branch-specific, single-branch, and multi-branch settings.

Setting	Model (Label Set)	Black-box	No Contrastive	w/ Contrastive
Branch-Specific	Rhythm Branch (16 labels)	0.8919 (0.8757, 0.9071)	0.8762 (0.8616, 0.8901)	0.8853 (0.8708, 0.8991)
	Morphology Branch (52 labels)	0.8996 (0.8930, 0.9057)	0.8791 (0.8727, 0.8855)	0.8996 (0.8931, 0.9059)
	Global Branch (3 labels)	0.8307 (0.8137, 0.8470)	0.6981 (0.6767, 0.7197)	0.9039 (0.8906, 0.9164)
Single-Branch	1D Prototype Model	**0.9081 (0.9012, 0.9147)**	0.8108 (0.8025, 0.8188)	0.8857 (0.8782, 0.8930)
	2D Partial Prototype Model	N/A	0.8605 (0.8526, 0.8684)	0.8743 (0.8666, 0.8819)
	2D Global Prototype Model	0.8932 (0.8859, 0.9002)	0.8589 (0.8505, 0.8669)	0.8916 (0.8841, 0.8992)
Multi-Branch	Macro Aggregation	0.8855 (0.8779, 0.8926)	0.8486 (0.8413, 0.8557)	0.8950 (0.8882, 0.9016)
	Fusion Classifier	N/A	0.8800 (0.8728, 0.8872)	**0.9066 (0.9000, 0.9128)**

Clinician Evaluation of Prototypes

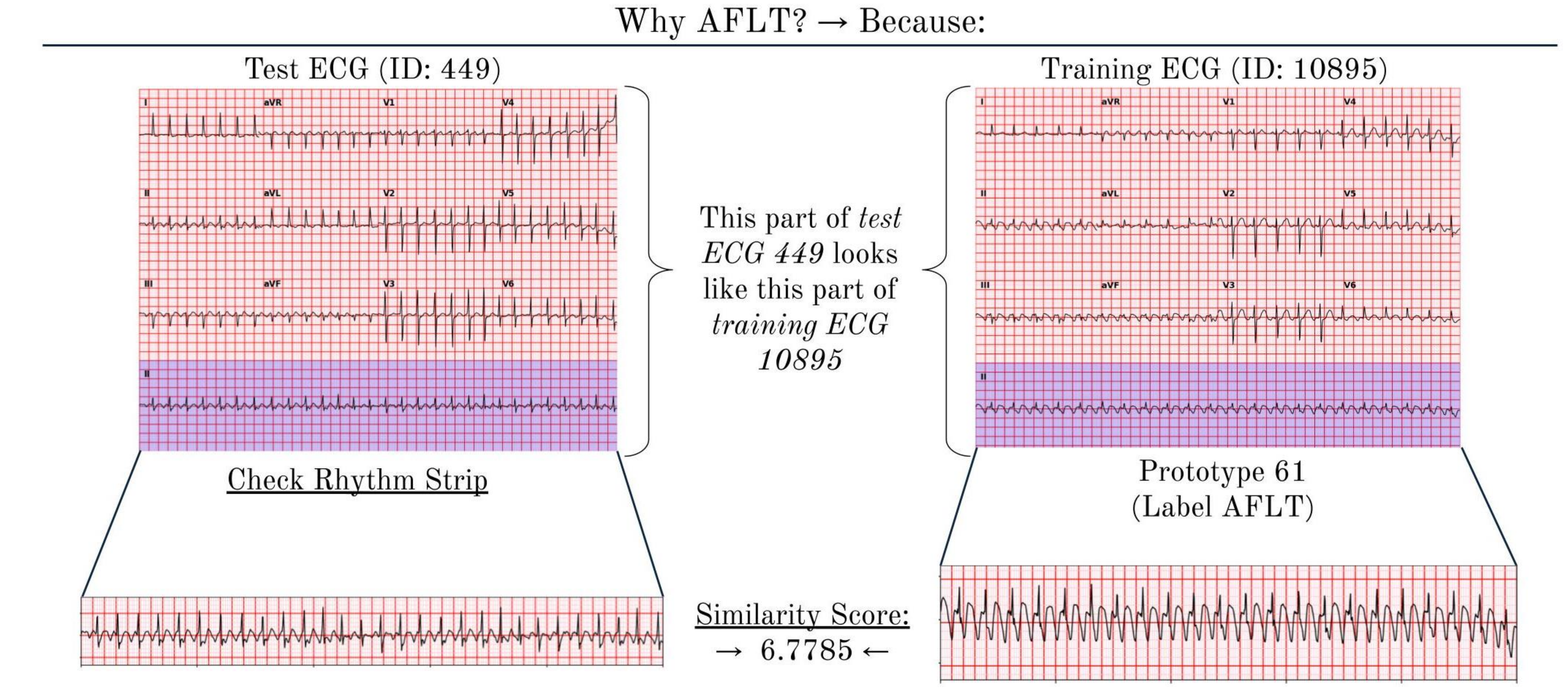
Table 3: Average prototype quality scores from structured clinician review (1–5 scale).

Reviewer	Representativeness (95% CI)	Clarity (95% CI)
Cardiologist	4.29 [4.22, 4.35]	4.48 [4.42, 4.54]
Internist	3.59 [3.52, 3.66]	4.73 [4.69, 4.77]

Case-Based Explanations

Case 1 – using 1D rhythm branch

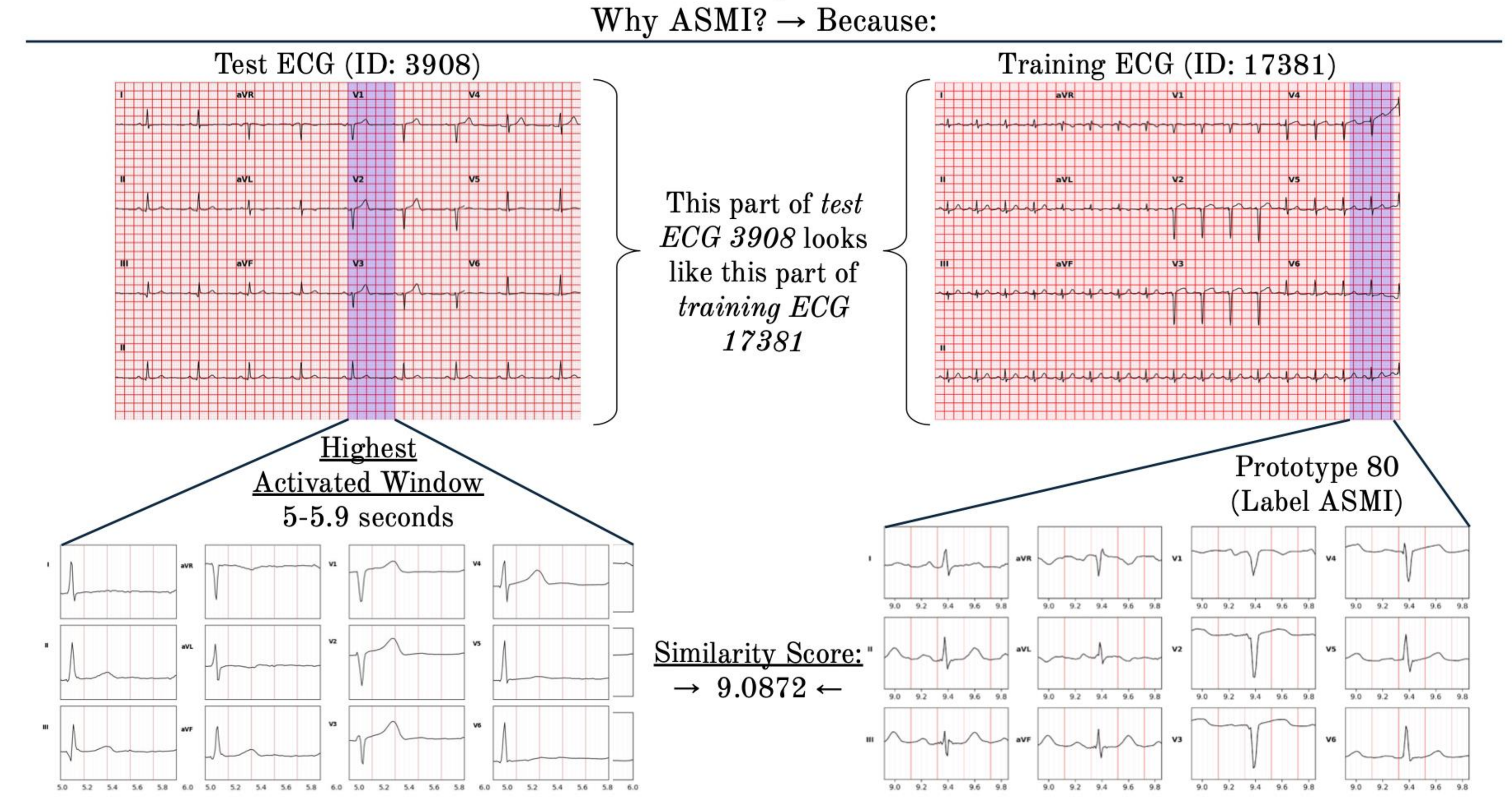
Prediction(s) From Fusion Classifier:
AFLT
Why AFLT? → Because:



AFLT = atrial flutter

Case 2 – using 2D partial prototype branch

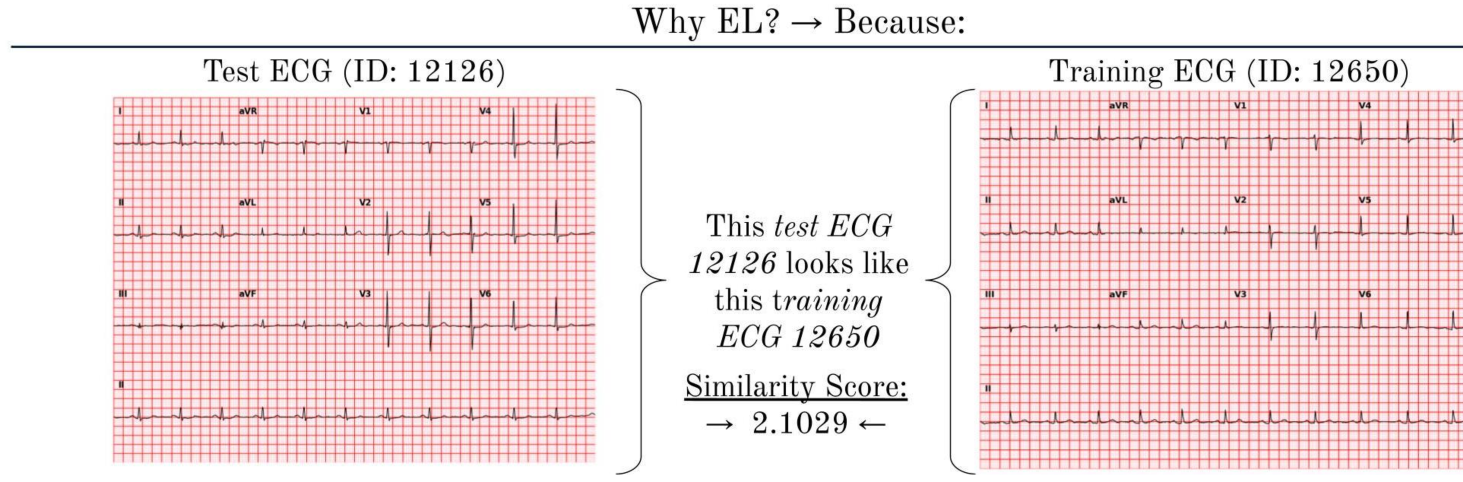
Prediction(s) From Fusion Classifier:
ASMI, SR
Why ASMI? → Because:



ASMI = anteroseptal myocardial infarction, SR = sinus rhythm

Case 3 – using 2D global prototype branch

Prediction(s) From Fusion Classifier:
EL, STD_, SR
Why EL? → Because:



EL = electrolyte abnormality, STD_ = ST depression

Paper, Code, and Contact Info