Safe and Explainable AI-Enabled Decision Making for Personalized Treatment

Rajeev Alur, PhD

Rajat Deo, MD
Sameed Khatana, MD
Qi Long, PhD
Mayur Naik, PhD
Ravi Parikh, MD
Payal Shah, MD
Gary Weissman, MD
Eric Wong, PhD

July 10, 2025

Team Members

- Haideliza Soto-Calderon; Project manager
- Nicholas Bishop; Data engineer
- Benjamin Schmidt; Clinical research coordinator
- Penn Engineering PhD students
  - Inyoung Choi
  - Seewon Choi
  - Cassandra Goldberg
  - Helen Jin
  - Mayank Keoliya
  - Chaehyeon Kim
  - Alaia Solko-Breslin
  - Jiayi Xin
- Penn Medicine Research Fellows
  - Alireza Oraii
  - Claire Zhang

Cardiology

Agenda

- Feedback on Prev. Quarterly Report
- State of the Practice in Cardiology by Dr. Deo
- Literature Survey, Challenges & Opportunities
- Prelim. Results & Experiments
- Next Steps
- Q & A (20 mins)

Feedback on Q3 Report

Any comments / questions are appreciated!

Recap: Workflow for Each Clinical Use Case

1. Data Collection

2. Literature Review

3. Problem Formulation

4. Model Building

5. Explainability

6. Clinical Evaluation

Recap: Workflow for Each Clinical Use Case

1. Data Collection

2. Literature Review

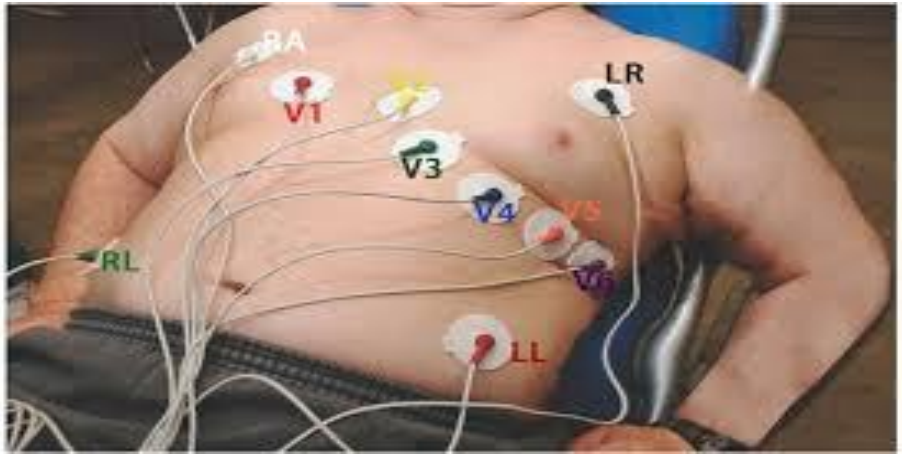
3. Problem Formulation

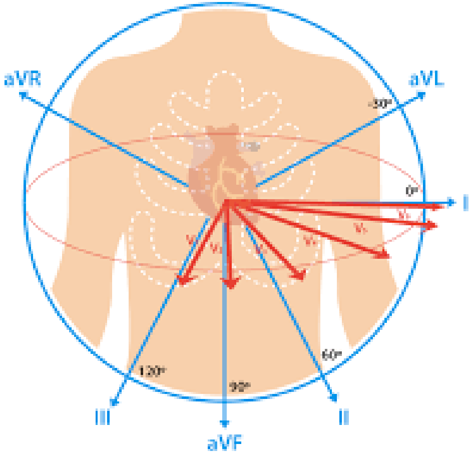
4. Model Building

5. Explainability

6. Clinical Evaluation

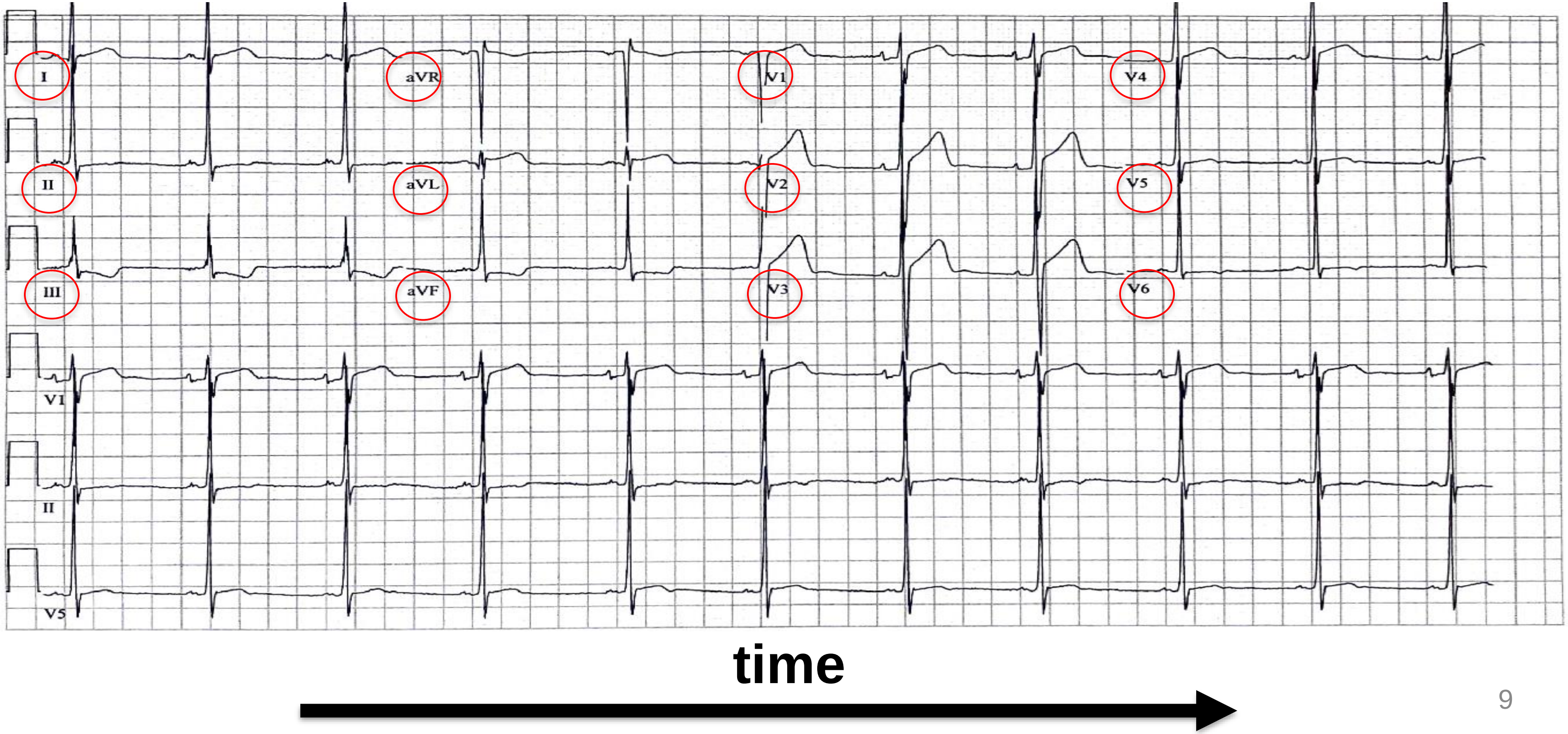
Background: Standard 12-lead Electrocardiogram (ECG)



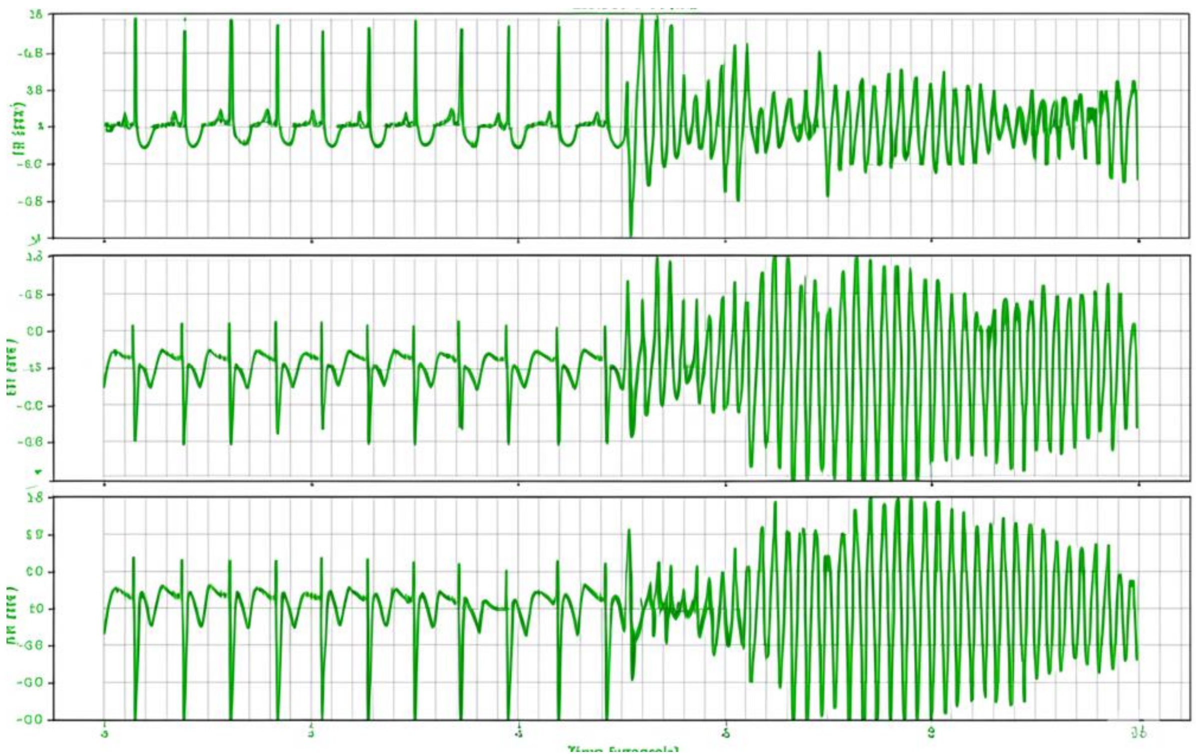


Full ECG Check-up = 12 Leads
ICU = 8 Leads
Fitbits = 1 Lead

10 second ECG tracing (in office)



An Example of Cardiac Arrest (VFib)

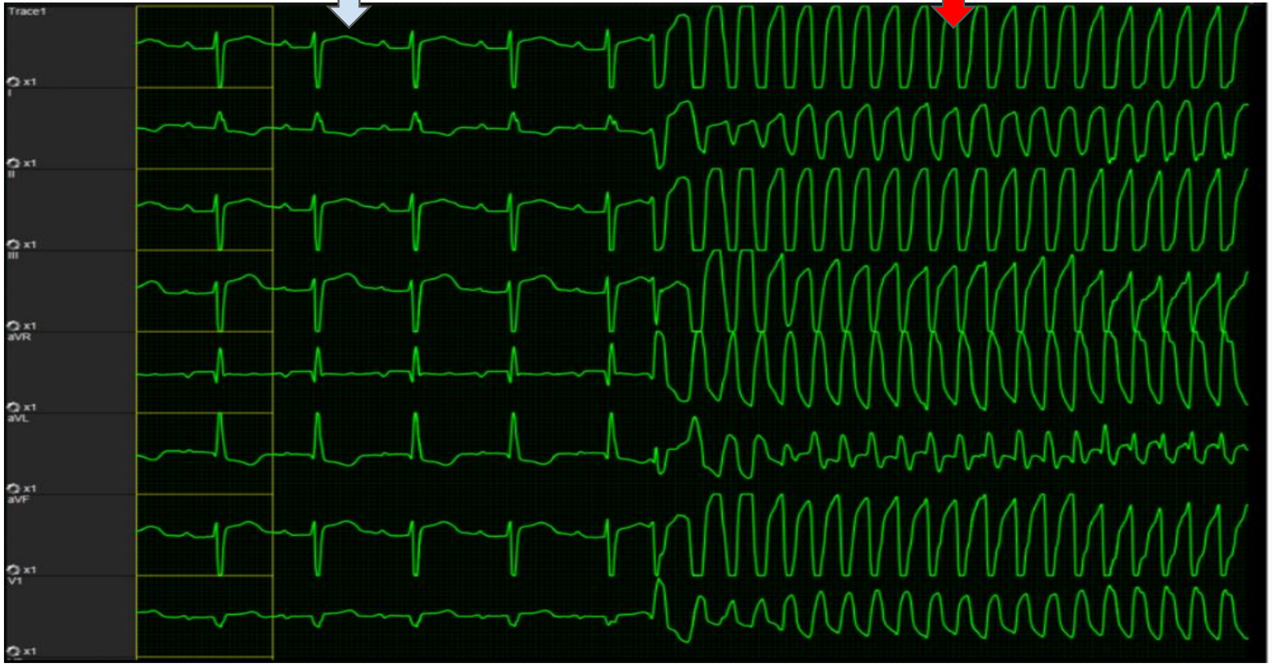


Another Example of Cardiac Arrest (VTach)

Goal: *Predicting* CA

How do we use this

to predict this?



Cardiac Arrest: Risk & Prevention

- ~300,000 adults per year (in-hospital) → Major cause of morbidity & mortality
- Only ~25% of patients who suffer in-hospital cardiac arrests survive to discharge[1]
- Retrospective studies have shown that a primary cause of preventable CA deaths is poor clinical monitoring[2]

1. *In-Hospital Cardiac Arrest, A Review.* Andersen et al. JAMA 2019.
2. *Preventable deaths due to problems in care in English acute hospitals. Hogan et al, BMJ 2012.*

Cardiac Arrest: Risk & Prevention

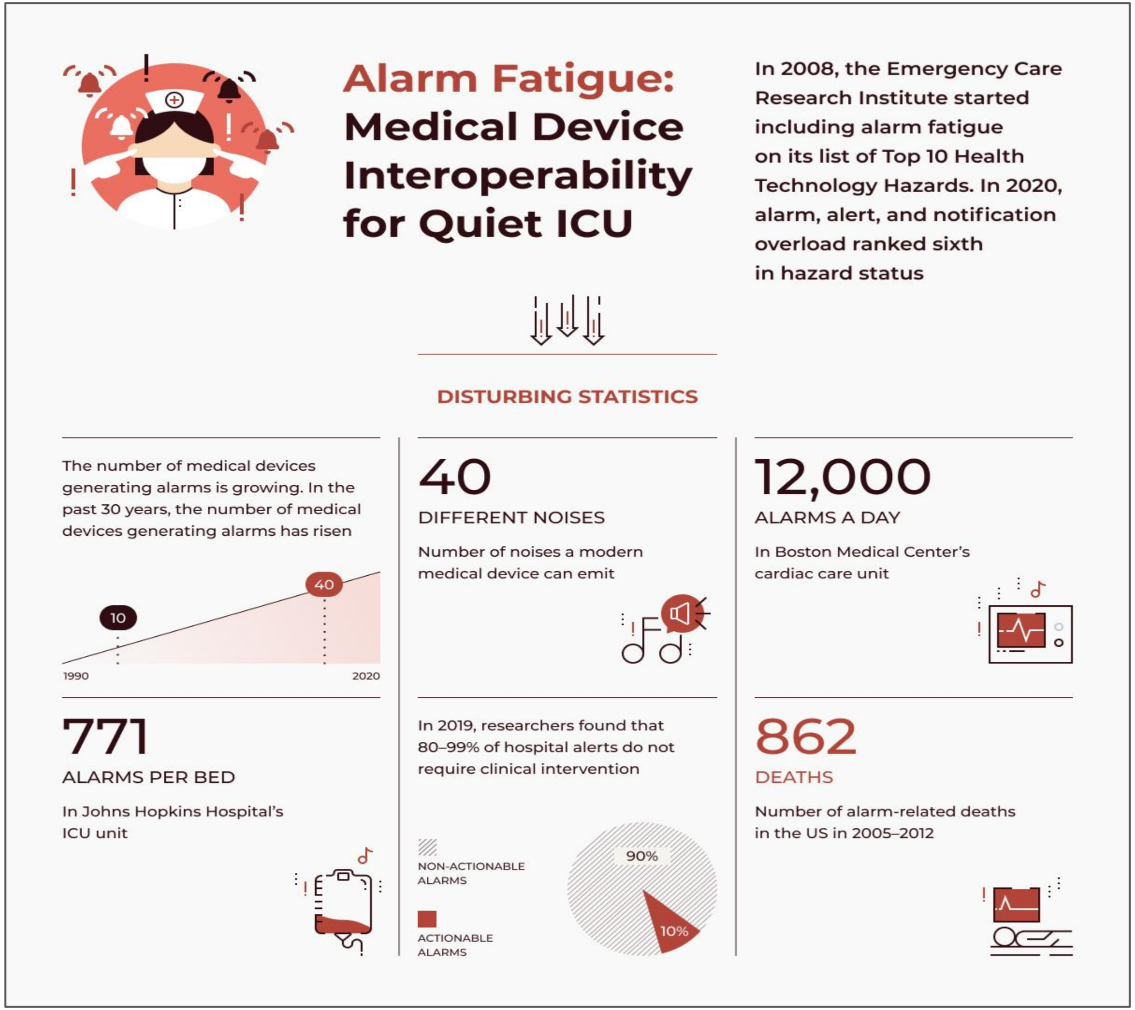
- Only ~25% of patients who suffer in-hospital cardiac arrests survive to discharge[1]
- Retrospective studies have shown that a primary cause of preventable CA deaths is poor clinical monitoring[2]
- At Penn Medicine → we observe that **post-hoc analysis of ECG by cardiologists can often reveal high risk of impending CA**, but
  - Currently deployed systems have a high false alarm rate ❌
  - Academic models *also* have a high false alarm rate ❌

1. *In-Hospital Cardiac Arrest, A Review.* Andersen et al. JAMA 2019.
2. *Preventable deaths due to problems in care in English acute hospitals. Hogan et al, BMJ 2012.*

Current Practice: High False Alarm Rate in ICUs[1,2]

>85% false alarm rate for ECGs

Strong correlation b/w alarm fatigue and medical errors[3]



Infographic from Auriga Inc.

1. Kim et al, 2025. 2. Park et al, 2023. 3. Kataria et al.

Goal: Early Prediction of Shockable CA Using ECG *Only*

- Early prediction of shockable CA can be very useful
  - **Even 30 seconds of advance notice can improve mortality outcomes![1]**

- Benefits of ECG-driven models
  - **Low marginal cost:** ICUs already have an ECG monitor per bed
  - **Rapid response to alarms:** ECG is high-frequency and continuous
  - **Likely more generalizable:** ECG is a high-fidelity reflection of patient's true state, standardized across hospitals, and not confounded with treatment unlike EHR

1. *Trends in Survival after In-Hospital Cardiac Arrest* Girotra et al. NEJM 2012.

What Are Existing Studies Missing?

Literature Review of Datasets & Models

Most Public Datasets Only Have 10 second ECGs

JOURNAL ARTICLE

Sudden cardiac arrest prediction via deep learning electrocardiogram analysis

Matt T Oberdier , Luca Neri ✉ , Alessandro Orro , Richard T Carrick , Marco S Nobile , Sujai Jaipalli , Mariam Khan , Stefano Diciotti , Claudio Borghi , Henry R Halperin

Author Notes

European Heart Journal - Digital Health, Volume 6, Issue 2,

https://doi.org/10.1093/ehjdh/ztae088

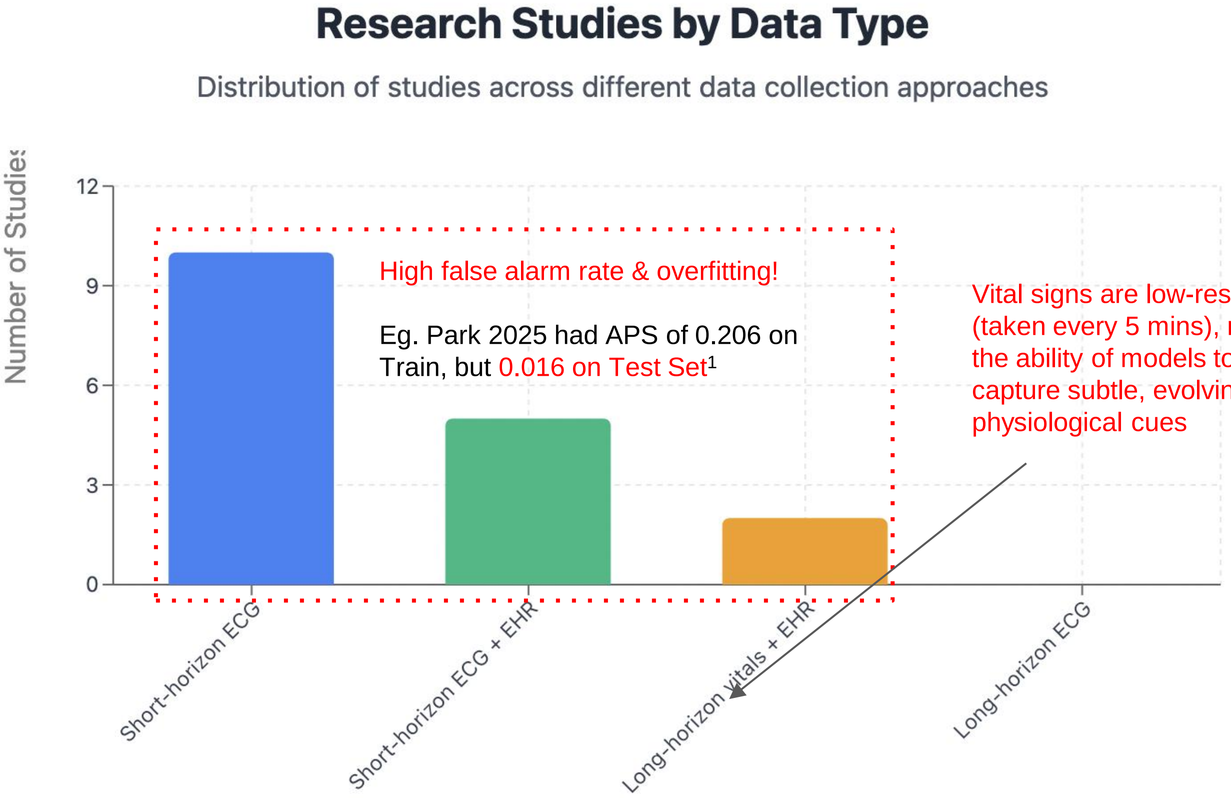
Published: 25 February 2025 **Article history** ▾

Methods and results

A publicly available data set containing 10 s of 12-lead ECGs from individuals who did and did not have an SCA, information about time from ECG to arrest, and age and sex was utilized for analysis to individually predict SCA or not using deep convolution neural network models. The base model that included age and sex,

"With sensitivity set at 95%, base model specificity was 31%, which is not clinically applicable"

Overview of Dataset Usage in Research Studies



Vital signs are low-resolution (taken every 5 mins), restrict the ability of models to capture subtle, evolving physiological cues

1. A Machine Learning Approach for Predicting In-Hospital Cardiac Arrest. *Park et al.* Ann Lab Med, 2025.

(Dr. Deo) Finer-Grained Information is Helpful!

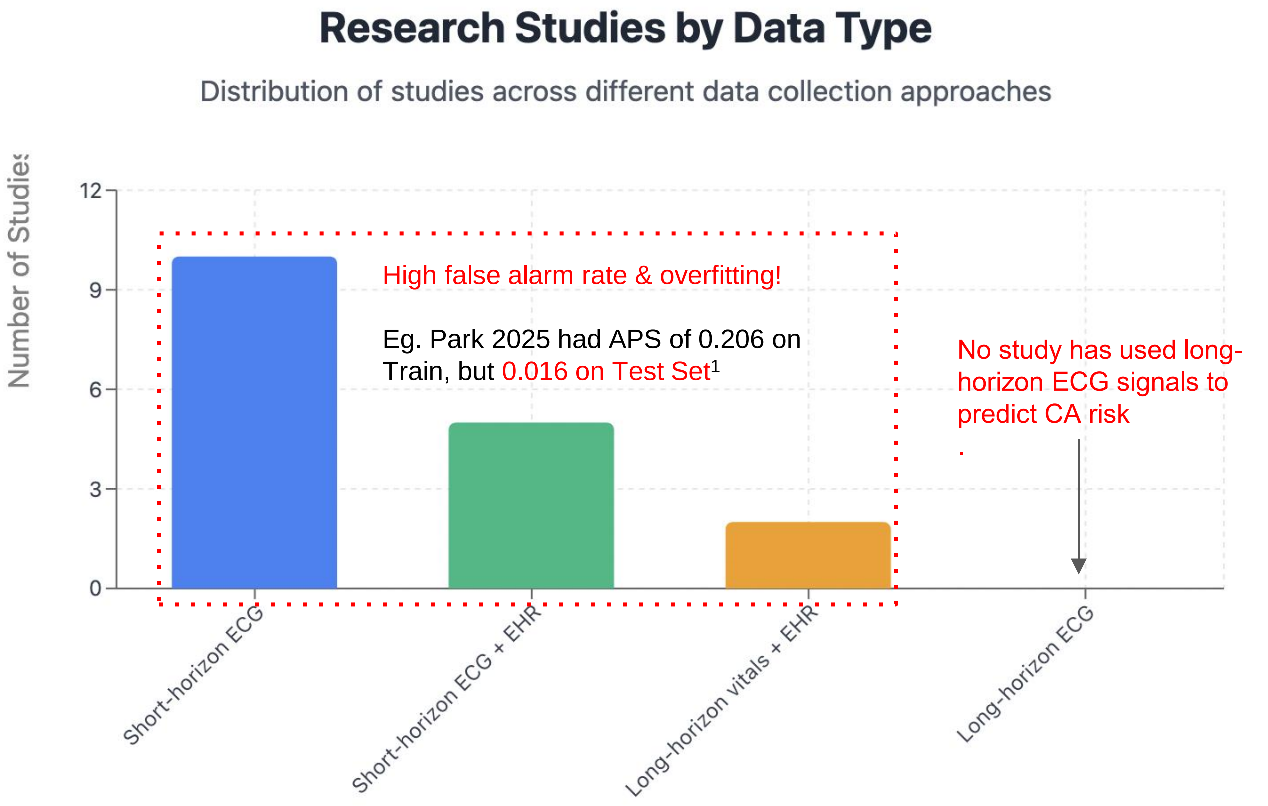
High-Res ECG

(250Hz)

Low-Res Vitals
(1 per 5 mins)

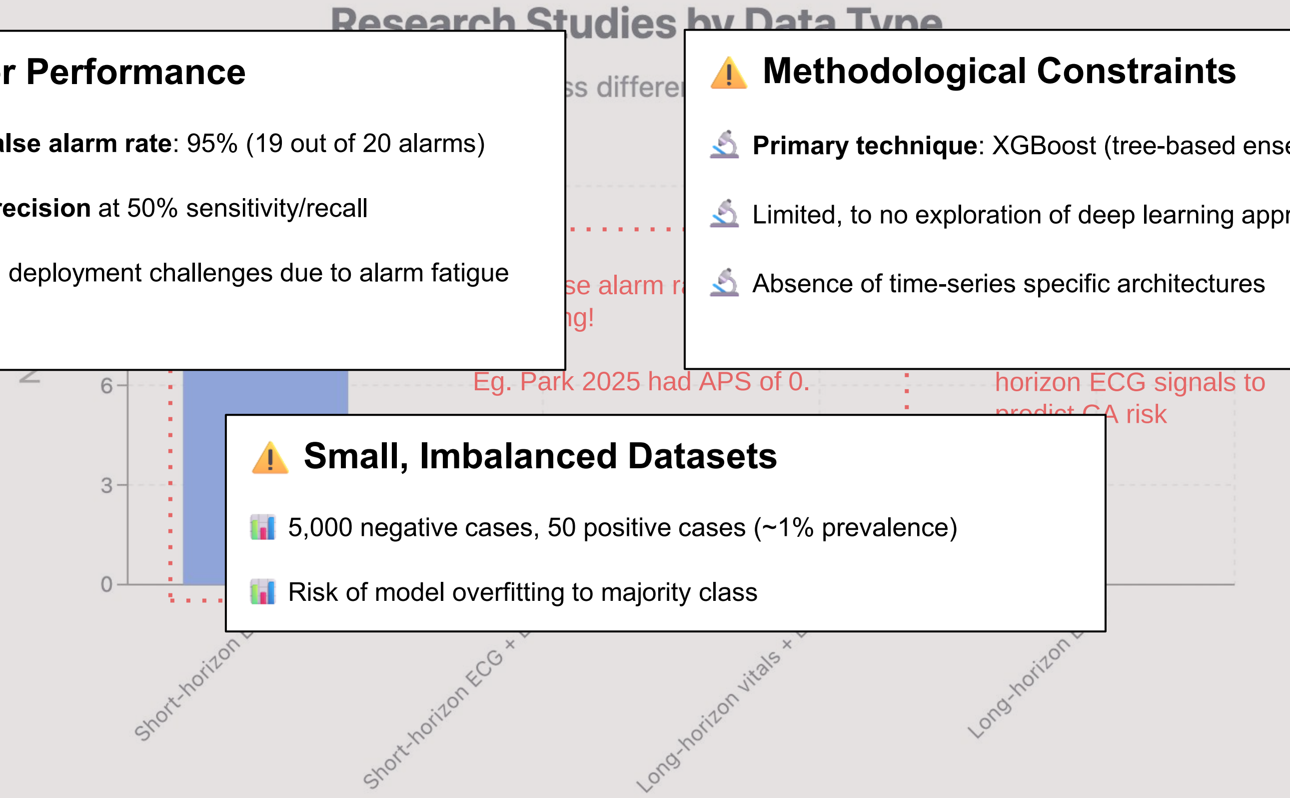
HR: 70, SDNN = ~27ms, pNN50 < 5% **(constant)**

Overview of Dataset Usage in Research Studies



1. A Machine Learning Approach for Predicting In-Hospital Cardiac Arrest. *Park et al.* Ann Lab Med, 2025.

ML-Driven CA Prediction is Lacking: Limitations



⚠️ Poor Performance

- 📈 **High false alarm rate**: 95% (19 out of 20 alarms)
- 📈 **Low precision** at 50% sensitivity/recall
- 📈 Clinical deployment challenges due to alarm fatigue

⚠️ Methodological Constraints

- 🔬 **Primary technique**: XGBoost (tree-based ensemble)
- 🔬 Limited, to no exploration of deep learning approaches
- 🔬 Absence of time-series specific architectures

⚠️ Small, Imbalanced Datasets

- 📊 5,000 negative cases, 50 positive cases (~1% prevalence)
- 📊 Risk of model overfitting to majority class

Dataset Collection

Comparison of Existing Public Datasets with Penn

DATASET	DURATION	SIZE	LEADS	LABELLED?	STORAGE
MIMIC-IV	10 seconds	800K	12	NO	7.5GB
NTUH	10 seconds	10K	12	NO	1.27GB
BWH/ Berkeley	10 seconds	61K	12	NO	5GB
CCHS/ Berkeley	10 seconds	43K	12	NO	6.71GB
MC-MED	2 hours	20K	2	NO	300GB
Icentia11k	2 weeks	11K	1	NO	1.1TB
Penn	~days	10K	8	YES	2.3TB

Data Collection

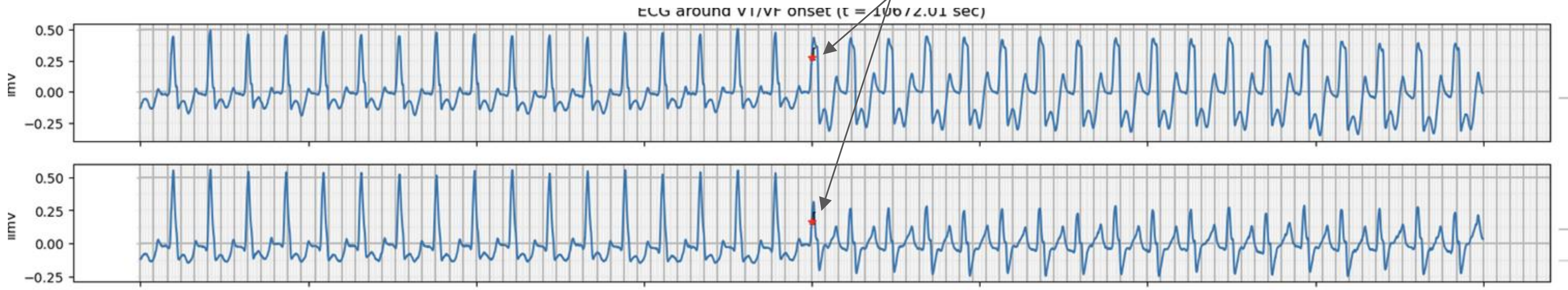
- **Total dataset:** 10K patients (unlabelled for now)

- **Labelled so far: 31 cardiac arrest patients** with precisely annotated onset times by physicians (pending: 10K)
    - Already more positive patients than vital-based datasets (JHU ‘23 & SNU ‘25)

- **High Resolution Sampling**
    - 3 hours of continuous ECG per patient (for now)
    - ~300 MB per patient (high-resolution sampling)
    - ~9.3 GB of ECG waveform data
    - **93 patient-hours** of annotated cardiac monitoring

Example from Penn Dataset

Physician-annotated time of CA



*Remaining 6 leads omitted in visualization

Preceding 3 hours is available!
(10 seconds shown here)

Modelling Efforts

Timeline, Challenges, Deliverables

Pre-processing

- Converted Penn data format (CSV) into **WFDB** (Waveform Database)
- Enables us to use off-the-shelf visualizers, noise cleaning, etc. like prior work
- Allows for loss-less compression relative to CSV
- Also lets us directly plug into other pre-trained models

Challenges & Timeline

Because we have lots of data → we have **two** novel challenges

1. **Rapidly labelling 10K+ patients** (classification): Physician-driven labelling is time-consuming and can't scale! → Ongoing

1. **Imbalanced, noisy long-horizon data** (forecasting): Difficult to distinguish noise (e.g. brushing teeth) from clinically-interesting ECG patterns (e.g. PVCs)

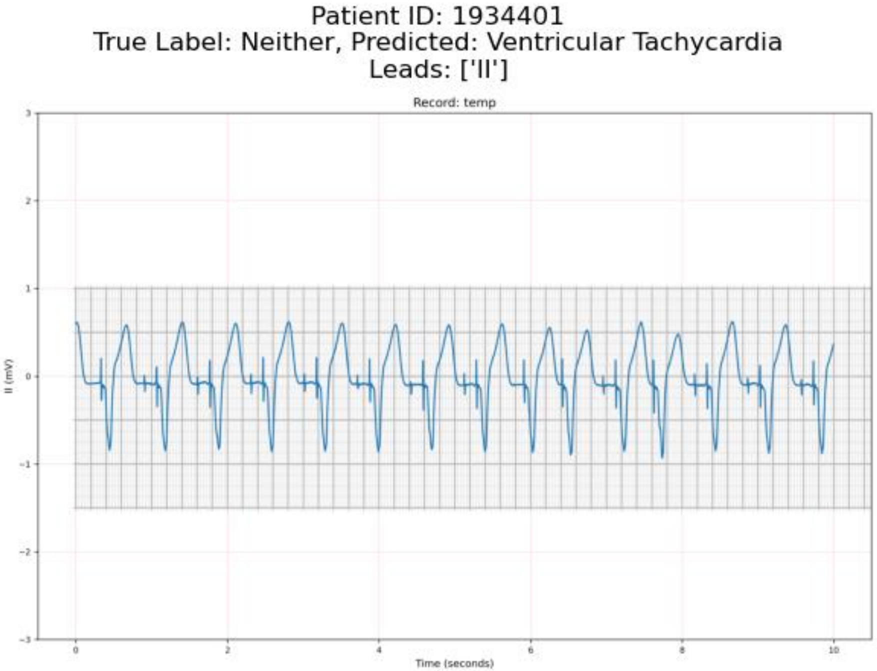
1. Data Annotation for 10K+ patients

We explored using LLMs (both ECG-specific and general) to label chunks of ECG which contain the first onset of CA

Leads/FP	Pulse-7B	gpt-4o	**o3**
1 Lead	0.97	0.80	**0.1**
3 Lead	0.97	1.0	**0.20**
8 Lead	0.97	1.0	**0.25**

All models have 0 False Negatives, but high False Positive (FP) rate

Example of False Positive (o3)



Incorrect Reasoning of False Positive (o3)

```
Model's Full Analysis:
================================================================================
1. Heart Rate and Rhythm Pattern:
   – Estimated heart rate: ~120 bpm (about 12 QRS complexes in 6 s → 120/min).
   – Rhythm: regular.
   – Organized QRS complexes are present throughout the strip.

2. QRS Analysis:
   – QRS duration: clearly prolonged, well over 120 ms (wide–complex).
   – Morphology: consistent from beat to beat (monomorphic).
   – No obvious P waves preceding each QRS and no fixed PR intervals—suggestive of AV dissociation.

3. Baseline Characteristics:
   – Baseline is clean, without fibrillatory or chaotic oscillations.
   – No continuous irregular waves interrupting the organized rhythm.
   – The pattern of wide complexes repeats uniformly across the 10 s.

4. Additional Relevant Features:
   – No clear fusion or capture beats are seen.
   – The electrical activity is organized and rapid, with identical QRS shapes.

DIAGNOSIS: Ventricular Tachycardia
```

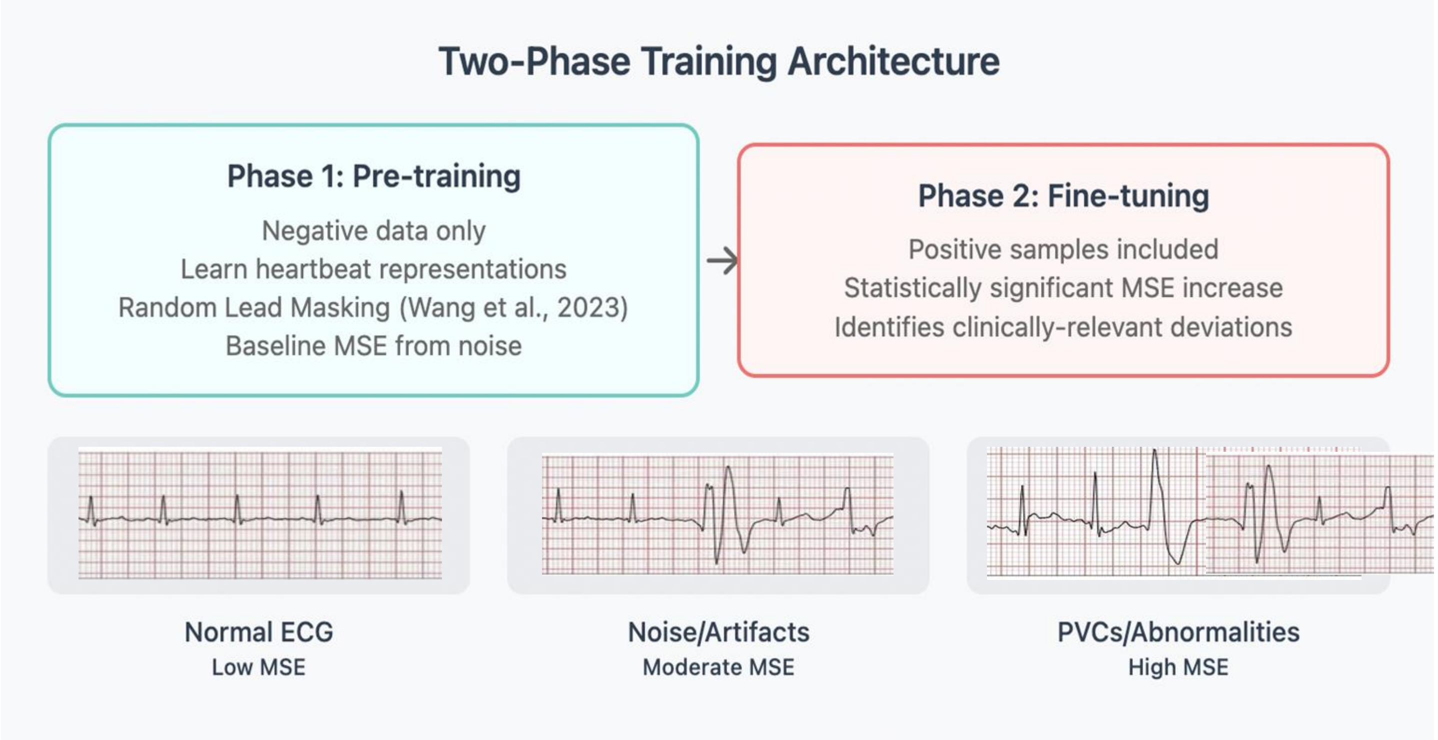
Counting error: should be 8

Pacemaker not identified!

Hallucinated values, but can be computed symbolically!

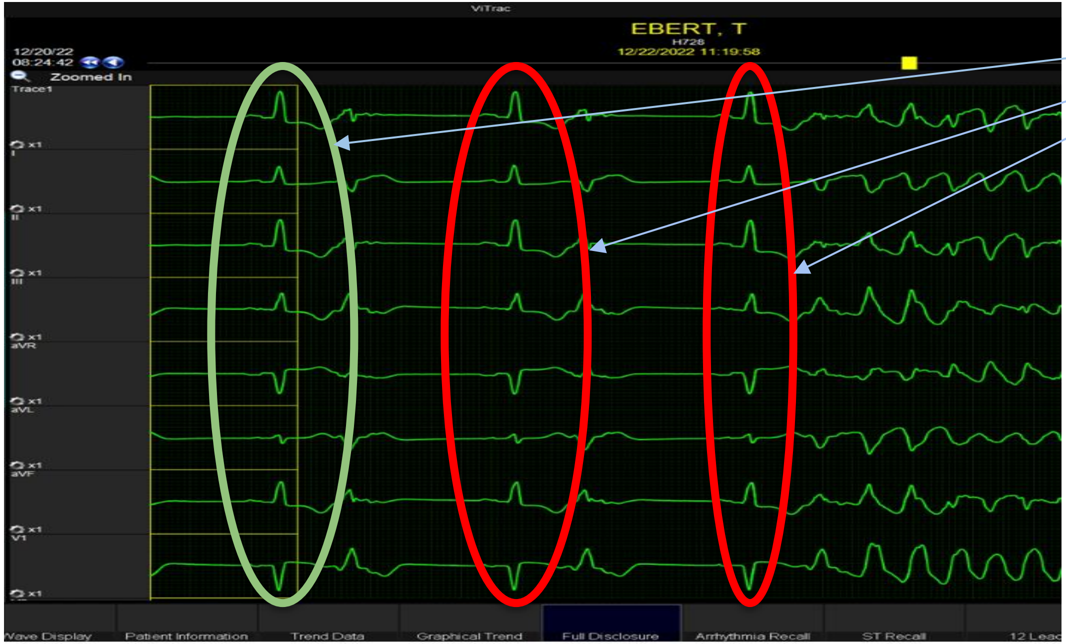
Next step: Computing ECG stats symbolically, supplying to LLMs (neurosymbolic)

2. Learning from Imbalanced, Noisy Long-Horizon Data



Explanations: Finding Relevant ECG Segments a la T-FIX[1]

Key challenge: Identifying segments of ECG which the model found to be indicative of high risk of CA



Assume the model flags 3 segments which have high deviation

1

R-on-T beat, due to wide …

Generate textual descriptions of ECG morphology using off-the-shelf ECG-text models

2

R-on-T

I prefer PVCs, R-on-T beats, with wide QRS

Compute alignment of descriptions with expert-annotated features

3

1. The FIX Benchmark: Extracting Features Interpretable to eXperts. Jin, Wong, et al 2024.

Deliverables & Criteria for Success

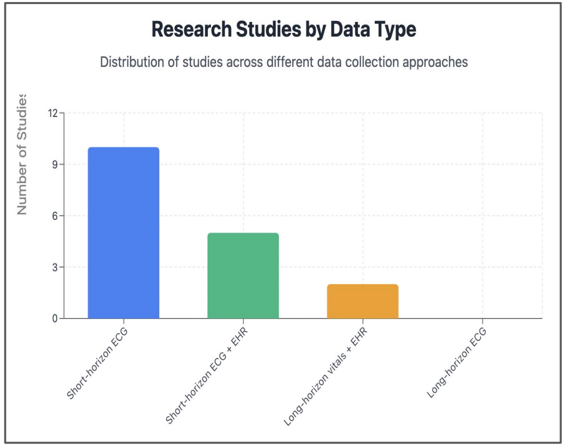
Sept: Long-horizon ECG dataset, ready for public release

Oct-Nov: Set of novel ECG patterns identified by model & clinically validated by Penn physicians

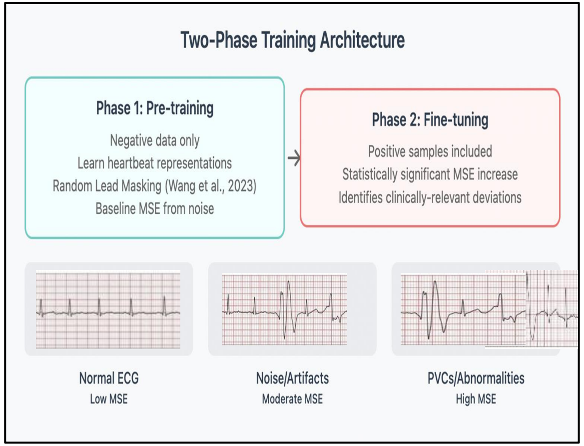
End-of-year: Model which has less than 1 false alarm *per 3 hours per patient* (with horizon in range [30s, 1hr]), as evaluated on Penn dataset

Future: Explorations of explainability, lead-agnostic ablations, generalizability.

Summary



DATASET	DURATION	SIZE	LEADS	LABELLED?	STORAGE
MIMIC-IV	10 seconds	800K	12	NO	7.5GB
NTUH	10 seconds	10K	12	NO	1.27GB
BWH/ Berkeley	10 seconds	61K	12	NO	5GB
CCHS/ Berkeley	10 seconds	43K	12	NO	6.71GB
MC-MED	2 hours	20K	2	NO	300GB
Icentia11k	2 weeks	11K	1	NO	1.1TB
Penn	~days	10K	8	YES	2.3TB



Existing models for CA prediction have high false alarm rates

We have collected a novel long-horizon ECG dataset at Penn

We will address novel challenges to build accurate & explainable models with low false alarm rates.