Leveraging ECG Foundation Models in Critical Care for Sinus Rhythm and Atrial Fibrillation Classification

Maria Galanty[1,2] Björn van der Ster[2] Alexander P. Vlaar[2] Clara I. Sánchez[1,2]

[1]University of Amsterdam; [2]Amsterdam UMC location University of Amsterdam

Motivation

Atrial fibrillation (AF) is a common arrhythmia marked by irregular heartbeats and is frequent among Intensive Care Unit (ICU) patients, where it is associated with increased risks of stroke, heart failure, and mortality [1]. Detecting AF in the ICU is challenging due to unstable physiology and coexisting conditions, but automatic detection could enable earlier intervention and improved outcomes.

However, reliable automated AF detection in the ICU remains challenging. Developing robust models requires labelled data, but manual annotation is costly and public ECG datasets often contain noisy, inconsistent labels [2]. Many deep learning models for arrhythmia detection perform poorly outside their training domains [2] and most prior AF studies focus on ambulatory (non-ICU) or machine-labelled ICU data [3, 4, 5, 6]. Recently, foundation models have shown new promises by learning generalizable representations from large-scale datasets.

Research question

Can ECG foundation models maintain robust performance on external data coming from the ICU population despite the domain shift and varying lead configurations?

Method

- **Data sources:** We used publicly available PhysioNet Challenge 2021 dataset[2] and an in-house Amsterdam University Medical Center (AUMC) dataset, which comprised two datasets: *AUMC Expert* - a high-quality expert-annotated test set (600 samples) and *AUMC-weak* - weakly labelled training set derived from routine nursing annotations (8,000 samples).
- **Task:** Multi-label classification AF and Sinus Rhythm (SR)
- **Model:** ECG-FM [7], a transformer-based architecture pretrained on 1.5 million ECGs.
- **Experiments:**
  1. **Label validation:** Assessed consistency and reliability of AF/SR annotations across PhysioNet Challenge 2021 datasets
  2. **Baseline evaluation:** Tested publicly available ECG-FM [7] on AUMC Expert.
  3. **Fine-tuning:** Investigated whether fine-tuning on selected public datasets, with or without weakly labelled in-house data, improves model results.
- **Evaluation metrics:** recall, precision, F1 score, and accuracy

Data

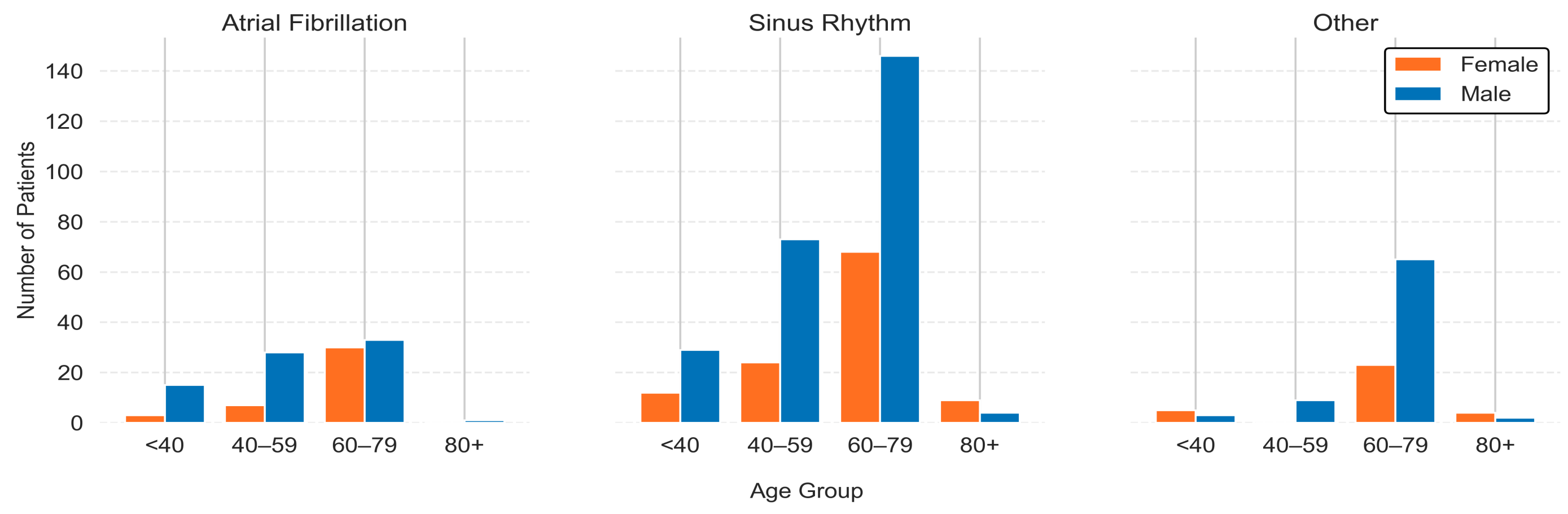

Figure 1: Patients distribution in the AUMC Expert set

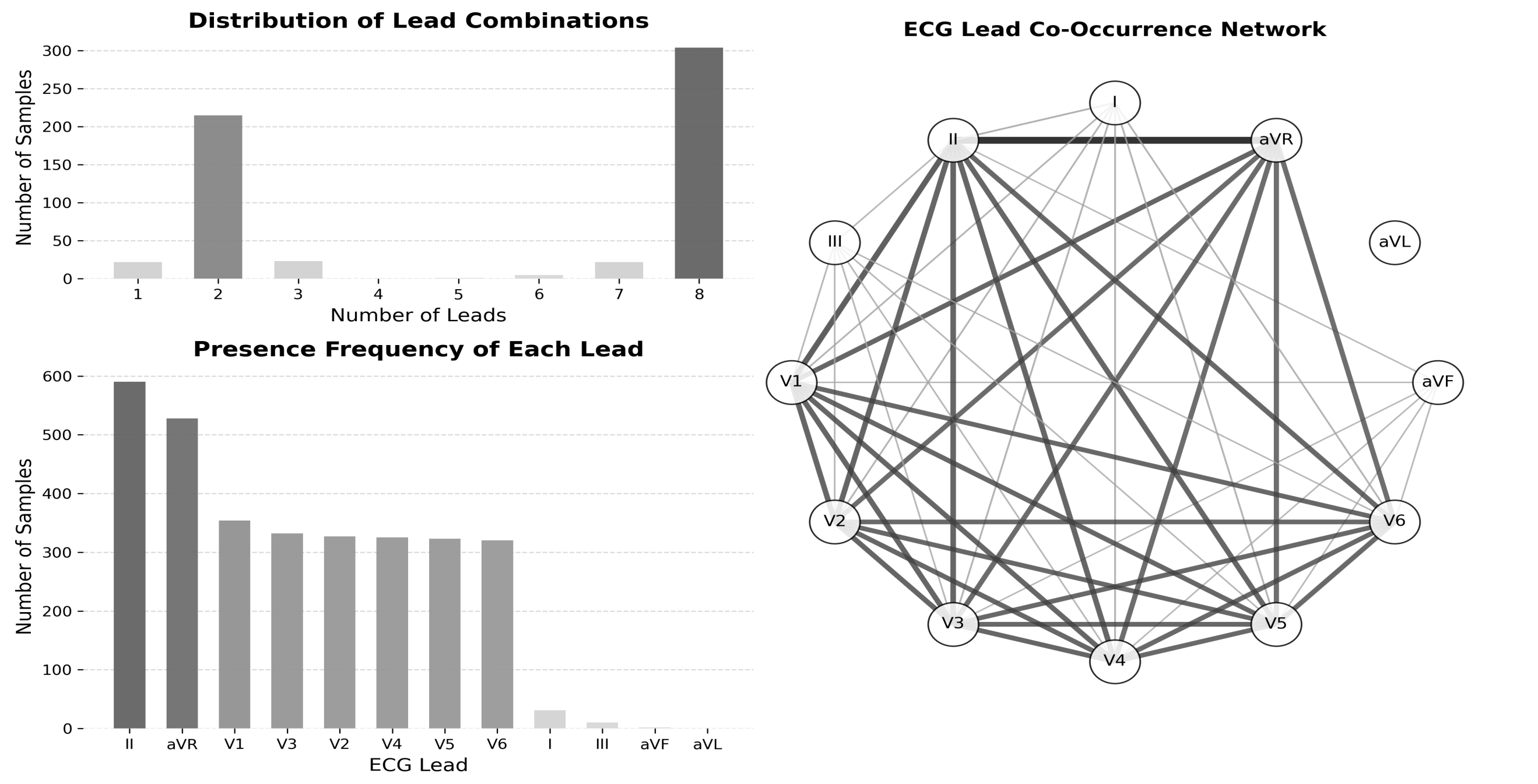

Figure 2: AUMC Expert: Left: lead availability per sample and frequency of each lead. Right: lead co-occurrence network (thicker edges indicate more frequent co-occurrence).

References

[1] S. Sibley et al., *Atrial fibrillation in critical illness: state of the art*, Intensive Care Medicine, pp. 1--13 (2025)

[2] M. A. Reyna et al., *Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021*, in *2021 computing in cardiology (CinC)*, volume 48, pp. 1--4, IEEE (2021)

[3] A. H. Ribeiro et al., *Automatic diagnosis of the 12-lead ecg using a deep neural network*, Nature communications, volume 11(1):p. 1760 (2020)

[4] A. A. Laghari et al., *Deep residual-dense network based on bidirectional recurrent neural network for atrial fibrillation detection*, Scientific reports, volume 13(1):p. 15109 (2023)

[5] W. Cai et al., *Accurate detection of atrial fibrillation from 12-lead ecg using deep neural network*, Computers in biology and medicine, volume 116:p. 103378 (2020)

[6] B. Gow et al., *Mimic-iv-ecg: Diagnostic electrocardiogram matched subset (version 1.0)* (2023), URL http://dx.doi.org/10.13026/4nqg-sb35, rRID:SCR_007345

[7] K. McKeen et al., *Ecg-fm: An open electrocardiogram foundation model*, arXiv preprint arXiv:2408.05178 (2024)

Results

Experiment 1

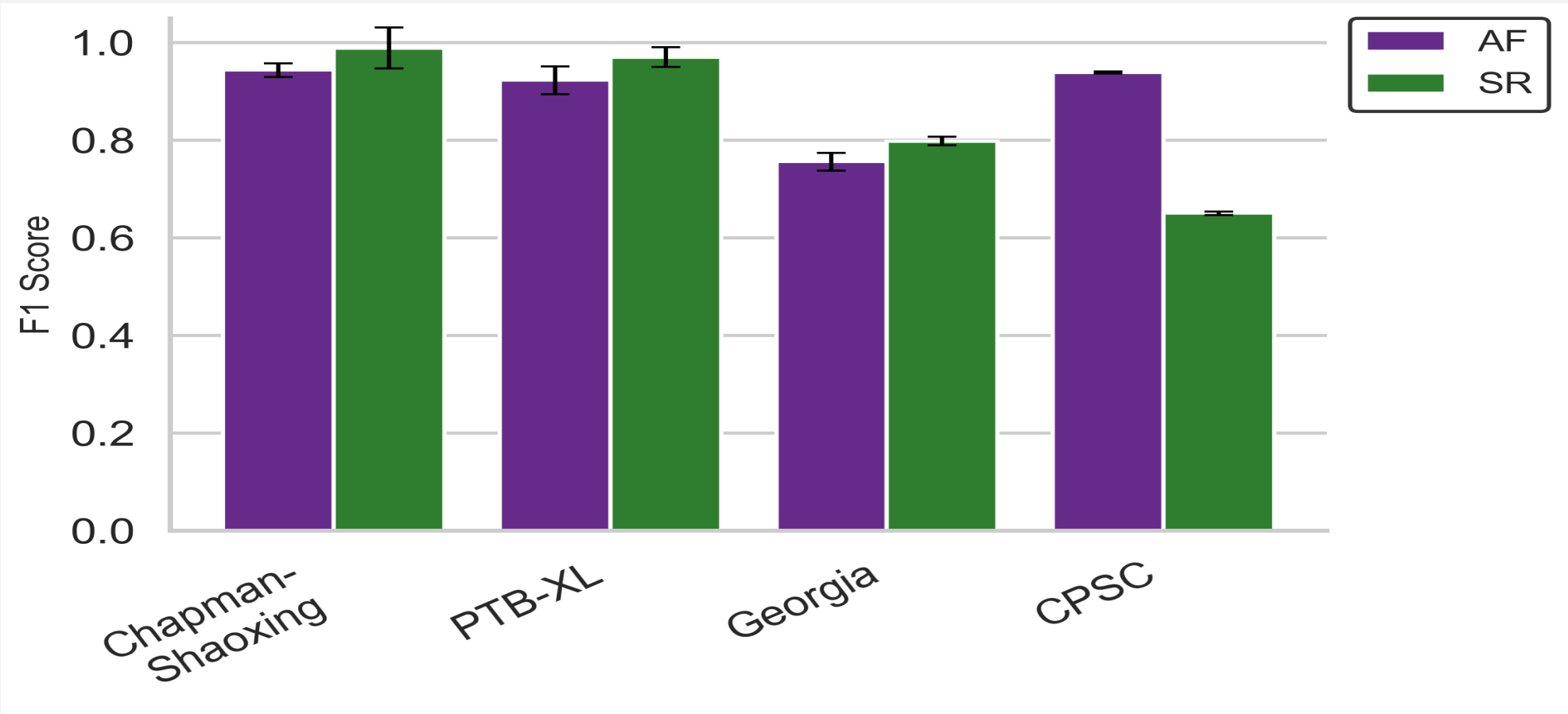

Figure 3: Cross-validation results showing mean F1 scores with 95% confidence intervals (CIs) across four public datasets

Experiment 2

Table 1: Classification performance of publicly available fine-tuned ECG-FM on the AUMC Expert for SR and AF. The left column indicates datasets used for fine-tunnings.

Fine-Tuning Datasets	Class	Accuracy	Precision	Recall	F1 Score
PhysioNet 2021 [2]	AF	0.914	0.718	0.903	0.800
	SR	0.561	0.939	0.299	0.454
MIMIC-IV-ECG [8]	AF	0.664	0.359	0.965	0.523
	SR	0.557	0.610	0.762	0.677

Experiment 3

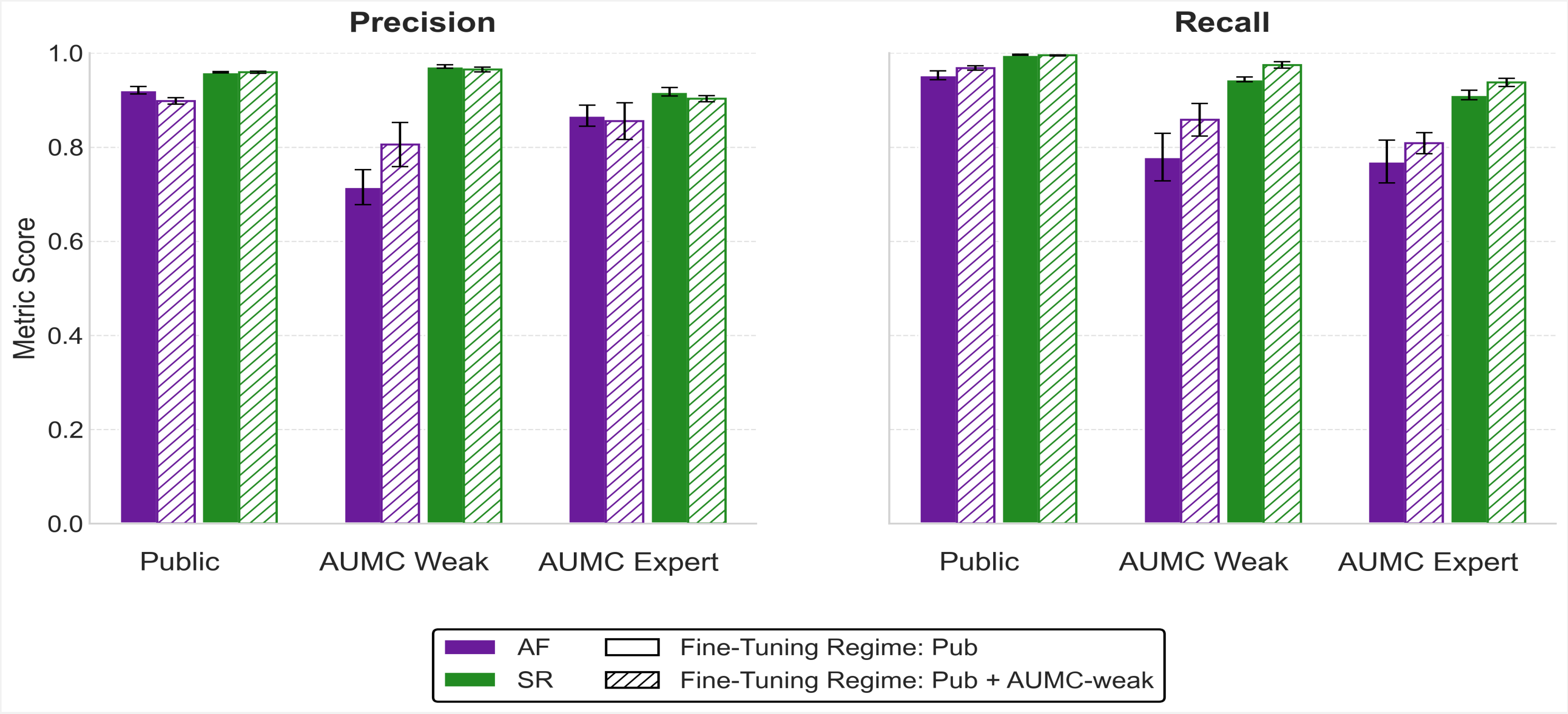

Figure 4: Precision and Recall across three evaluation datasets (Public test, AUMC Weak, AUMC Expert) for AF and SR. Solid lines show models trained without weak AUMC data, dashed lines with weak AUMC data. Error bars indicate 95% CIs.

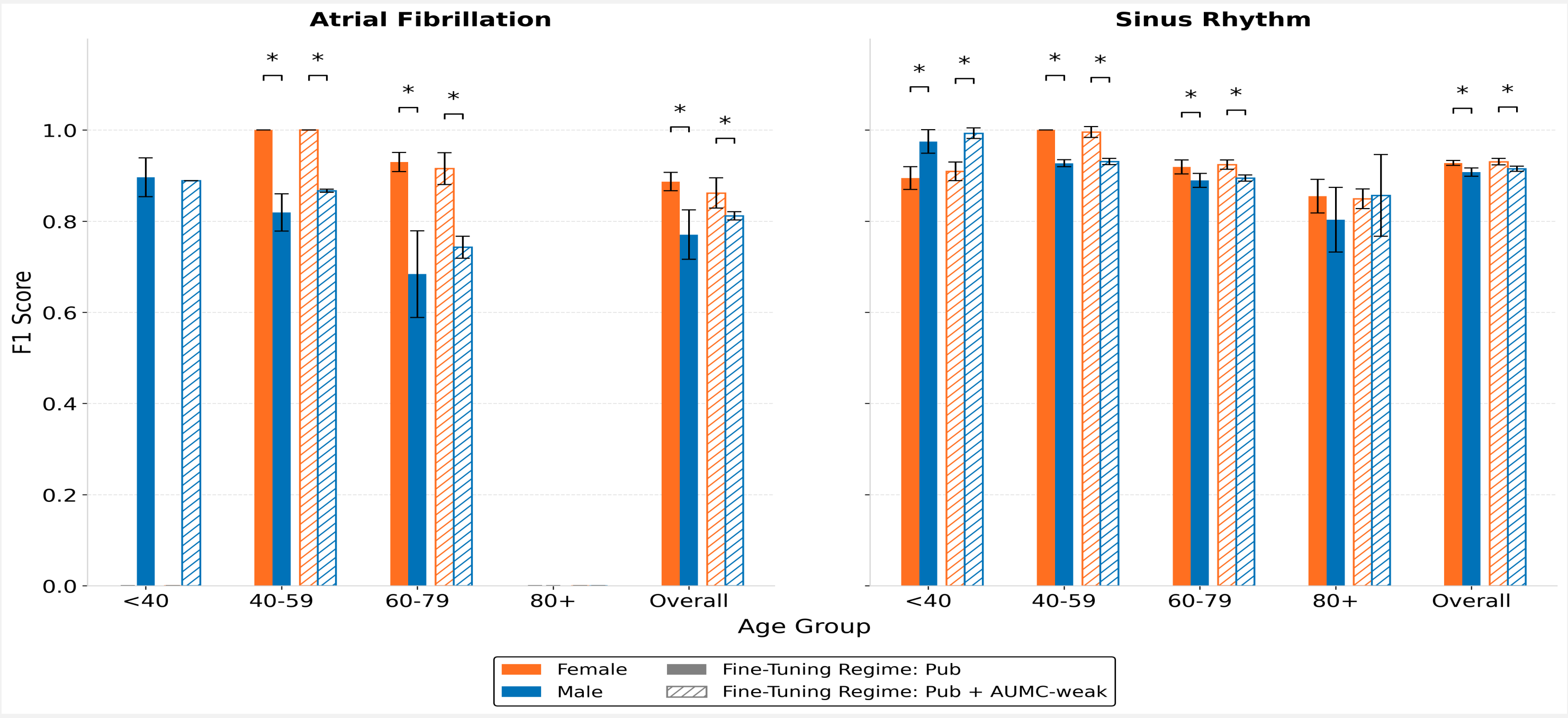

Figure 5: F1 scores for AF and SR stratified by sex and age. Results are shown for AUMC Expert dataset; models are fine-tuned without (solid) and with (hatched) weakly labelled AMC data. Bars indicate means over five runs with 95% CIs.

Conclusions

ECG-FM provide a scalable, clinically relevant method for detecting arrhythmias in the ICU. Fine-tuning on public datasets, with or without weakly labelled in-house data, **achieved strong performance, generalised well to external ICU data, and was stable across different lead configurations**.

Challenges & Future Directions

This study was limited by using a single foundation model and a small, single-centre AUMC dataset that only captured bedside monitor data. Future work should expand to multi-centre datasets and investigate the performance gap between female and male patients to determine whether it arises from data-specific factors or model bias.

RPA AI for Health Decision-making

qurAI

Universiteit van Amsterdam
Instituut voor Informatica

University of Amsterdam
Faculty of Humanities

Amsterdam UMC