MELBA Journal Machine Learning for Biomedical Imaging

Investigating sex bias in ECG classification for Atrial Fibrillation, Sinus Rhythm and Myocardial Infarction

Maria **Galanty** [1,2], Björn van der Ster [3], Alexander P. Vlaar [4], Clara I. Sánchez [1,2],

1 Informatics Institute, University of Amsterdam, The Netherlands
2 Department of Biomedical Engineering and Physics, Amsterdam UMC, University of Amsterdam, The Netherlands
3 Department of Anesthesiology, Amsterdam University Medical Center, University of Amsterdam, Amsterdam Cardiovascular Sciences, Amsterdam, the Netherlands
4 Department of Intensive Care, Amsterdam UMC, University of Amsterdam, The Netherlands

Abstract

Deep learning models are increasingly applied to electrocardiogram (ECG) analysis to optimise cardiovascular care. However, potential biases within these models may impact their reliability and clinical applicability. This study investigates potential sex bias in deep learning models for 12-lead ECG classification of Sinus Rhythm (SR), Atrial Fibrillation (AF), and Myocardial Infarction (MI). We evaluate three model architectures—Convolutional Neural Network, xResNet101, and a Residual Network with an attention mechanism—under varying sex ratios in the training data. Among these models, the attention-based Residual Network demonstrated the highest and most equitable performance, particularly in SR and AF classification. MI classification exhibited pronounced sex-based disparities, even with balanced training data. These findings underscore the importance of incorporating fairness considerations in the development of clinical deep learning systems to ensure reliable and unbiased performance across diverse patient populations. Moreover, optimising lead selection may further enhance both fairness and overall model performance.

Keywords

Deep learning, Electrocardiography, Classification, Bias

Article informations

https://doi.org/https://doi.org/10.59275/j.melba.2025-9fe7

Volume 3, Received: 2025-04-18, Published 2025-09-05

Corresponding author: m.galanty@uva.nl

Special issue: Fairness of AI in Medical Imaging (FAIMI) 2025
Guest editors: Veronika Cheplygina, Aasa Feragen, Andrew King, Ben Glocker, Enzo Ferrante, Eike Petersen, Esther Puyol-Antón, Melanie Ganz-Benjaminsen

1. Introduction

Deep learning in the medical domain—including electrocardiogram (ECG) analysis—is advancing rapidly, with numerous studies highlighting its potential for detecting and predicting abnormalities (Ribeiro et al., 2020; Singh et al., 2022). By automating complex pattern recognition, deep learning offers opportunities to improve diagnostic precision and streamline clinical workflows. However, despite its potential, concerns about algorithmic bias have gained increasing attention. Studies have shown that the performance of deep learning models can be compromised by various biases (Vokinger et al., 2021). Bias in medical AI has been documented across several domains, including ophthalmology (Khan et al., 2021), radiology (Kaushal et al., 2020; Petersen et al., 2022; Larrazabal et al., 2020), and cardiology (Lee et al., 2023, 2025).

Interpretation of ECGs can be significantly impacted by physiological factors, including age, sex, and race, each of which influences cardiac electrophysiology. These variations can lead to diagnostic inaccuracies or disparities in health care outcomes, highlighting the need for diversified ECG evaluation criteria (Zheng et al., 2025; Kittnar, 2023). Variability in ECG characteristics across racial groups presents important clinical challenges, influencing diagnostic accuracy and clinical decisions. Because traditional ECG norms are predominantly based on studies involving Caucasian

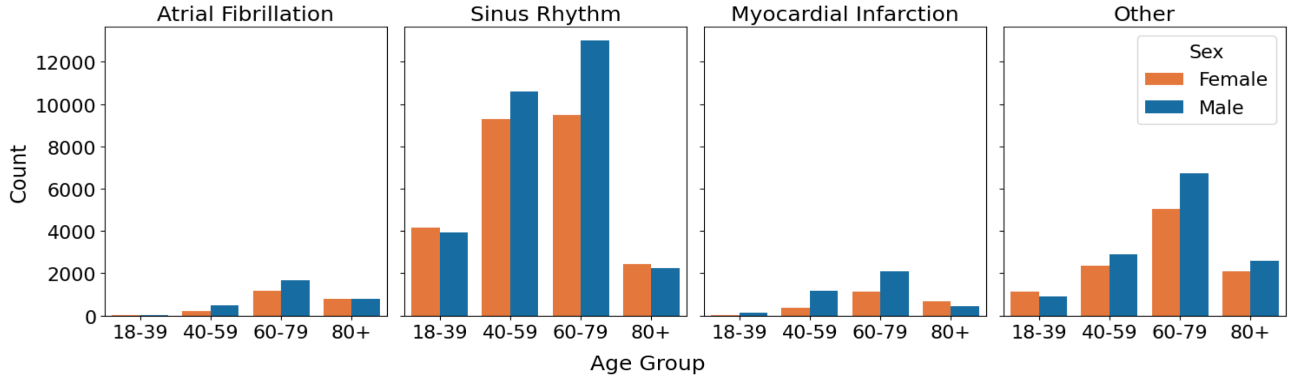


Figure 1: Distribution of labels across age groups (18-39, 40-59, 60-79, 80+) for males and females. The graph shows the prevalence of Atrial Fibrillation, Sinus Rhythm, and Myocardial Infarction in different age categories, separated by sex.

populations, applying these norms universally can result in diagnostic inaccuracies in other racial groups. A notable example includes the higher incidence of early depolarisation observed among African Americans, frequently mistaken for myocardial infarction (MI), leading to unwarranted invasive procedures and medical treatments (Zheng et al., 2025). Additionally, age-related hormonal changes can alter ECG patterns significantly, contributing to further diagnostic complexity across different age groups (Kittnar, 2023).

Sex differences in electrocardiography are also well documented. As reported by Kittnar (2023), men generally exhibit certain distinct ECG characteristics due to variations in heart size and muscle mass, whereas women tend to have a higher heart rate throughout adolescence and adulthood. These distinctions are influenced by sex hormones, particularly during puberty, when hormonal changes begin to shape the heart's electrical activity differently in males and females. Importantly, these differences are not static—they evolve with age and hormonal changes, such as those occurring during menopause, which can significantly increase cardiovascular risk in women. Zeitler et al. (2022) emphasises the clinical implications of these sex-specific patterns, especially for diagnosing and managing arrhythmias. For example, a prolonged QT interval—which reflects the time the heart's ventricles take to recover between beats—carries a higher risk of arrhythmias in women, highlighting the need for sex-specific thresholds in diagnosis and treatment (Zeitler et al., 2022).

Atrial fibrillation (AF) is the most common cardiac arrhythmia. While AF is more prevalent in men, women are more likely to present with persistent AF and experience atypical symptoms such as weakness and fatigue (Giammarino et al., 2025). Accumulating evidence highlights sex-specific differences in the physiological, electrical, and structural characteristics of the atria in AF (Giammarino et al., 2025; Odening et al., 2019). Variations in sex hormones influence ECG features such as P-wave morphology—males typically exhibit longer P-waves with lower variability—and higher age-related prevalence of AF. Additionally, males tend to have longer RR intervals than females (Odening et al., 2019; Laureanti et al., 2020).

Building on this, Sau et al. (2025) introduced an AI-ECG biomarker, known as the sex discordance score, to identify females with disproportionately elevated cardiovascular risk, enabling more targeted risk factor modification and enhanced clinical surveillance. Despite such advances, female representation in cardiovascular clinical trials remains low, and the generalisation of male-centric findings to all populations has introduced risks of misdiagnosis or undertreatment (Tobb et al., 2022). Neglecting these sex-specific differences in both clinical practice and AI development perpetuates systematic disadvantages for one sex. Given that cardiovascular disease remains the leading cause of mortality worldwide (Roth et al., 2020), it is essential that ECG interpretation—and the AI tools supporting it—account for physiological differences, including those between males and females, to ensure accurate diagnosis, equitable treatment, and improved outcomes for all patients.

Despite these known physiological distinctions, only a few deep learning studies have explicitly investigated sex-related biases in the development or evaluation of ECG-based models. For instance, Kaur et al. (2024) investigated disparities in race, sex, and age in the performance of convolutional deep learning models predicting heart failure within five years using 12-lead ECGs. Notably, model performance declined with age and was significantly worse in Black patients aged 0 to 40 compared to other racial groups within this cohort, with the most pronounced disparity observed in young Black women. Integrating race, ethnicity, sex, and age into the model architecture, training separate models for racial groups, and ensuring equal racial representation did not resolve these disparities. However, using individualised probability thresholds led to improved F1 scores.

Alday et al. (2022) examined biases related to sex, age,

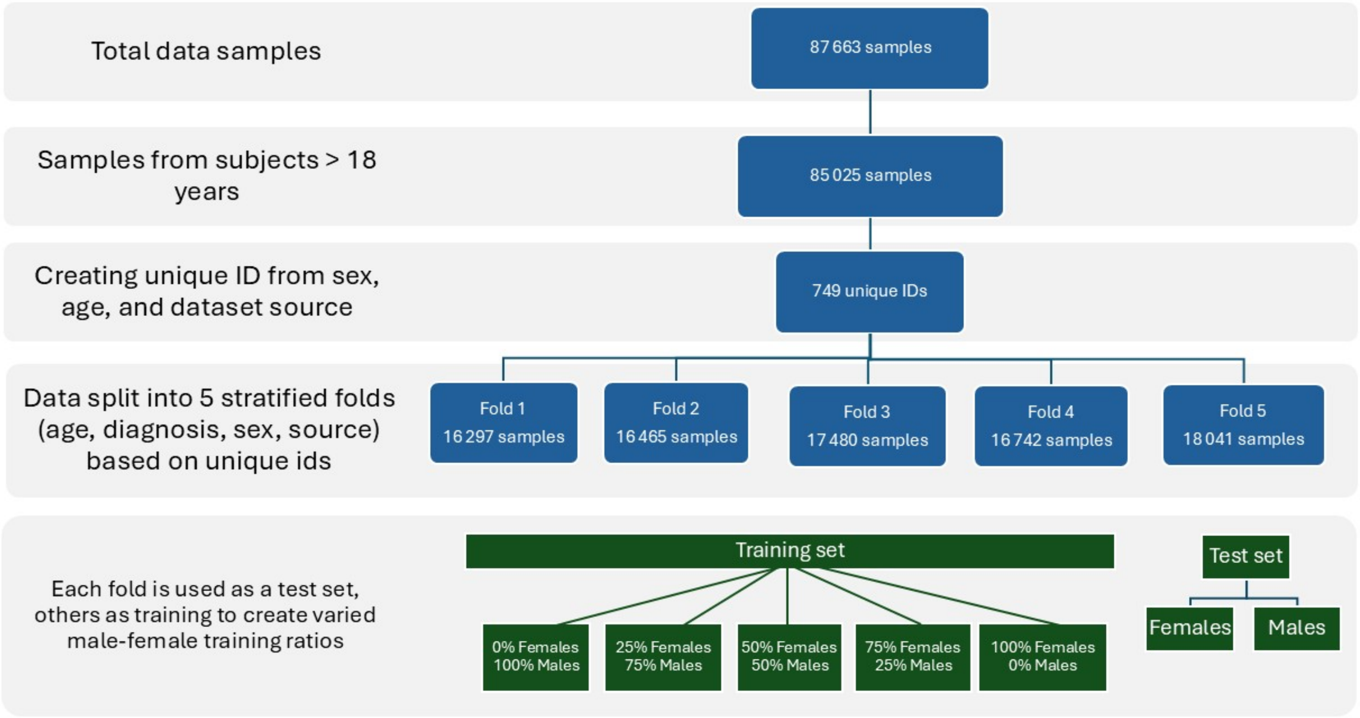


Figure 2: Graphical representation of the dataset creation process, illustrating the key steps involved in data preprocessing, stratification, and distribution across different training regimes

and race among 56 algorithms participating in the 2021 PhysioNet Challenge. The training and validation sets used in the challenge were balanced in terms of sex (male and female) and race (Asian, Black, White, Other). They found significant performance differences of algorithms in several evaluation metrics across sex, race and age groups. e.g, the Challenge Score was 12% lower for female subjects on the test set.

Noseworthy et al. (2020) analysed model performance by race and ethnicity for a convolutional neural network designed to identify patients with low left ventricular ejection fraction based on 12-lead ECG. This model was developed in a predominantly homogeneous population (96% non-Hispanic White), and their study demonstrated that although ECG characteristics vary by race, this did not affect the model's overall performance.

Previous studies (Kaur et al., 2024; Alday et al., 2022; Noseworthy et al., 2020) have made valuable contributions to the field by highlighting the presence of demographic disparities in ECG-based deep learning models. These works have played a substantial role in raising awareness of potential biases and promoting the development of fairer AI systems. However, this research has primarily focused on evaluating the fairness of already-developed algorithms (Kaur et al., 2024; Alday et al., 2022; Noseworthy et al., 2020), without explicitly investigating how sex imbalance in training data may contribute to model bias. While demographic bias has been explored across intersecting dimensions—such as race, age, and sex—sex itself has rarely been isolated as the central variable of analysis. Given the well-established physiological differences between males and females, there is a strong need for a dedicated investigation into sex-specific bias in ECG classification and the extent to which these differences may affect model performance. Such a study would also benefit from evaluating whether mitigation strategies—such as sex-balanced datasets or sex-specific model development—can enhance performance accuracy for both sexes. Addressing this gap is essential for understanding the role of sex-based physiological variation in algorithmic behaviour and for advancing fairness in clinical AI applications.

To address this need, the present study systematically evaluates sex bias in deep learning models for ECG classification. We explore three model architectures—Convolutional Neural Network (CNN) (LeCun et al., 1989), xResNet101 (He et al., 2016) and Residual Network with an Attention Mechanism (Nejedly et al., 2021). Using a dataset comprising multiple diagnostic categories—Sinus Rhythm (SR), Atrial Fibrillation (AF), and Myocardial Infarction (MI)—we assess the impact of sex imbalance in training data and quantify performance differences between male and female patients. Our goal is to explore whether sex-based physiological differences contribute to model bias. This work contributes to the growing field of fairness in medical AI by providing focused insights into sex-related bias in ECG-based cardiovascular diagnostics.

2. Methods

2.1 Dataset

In this study, we utilised the following publicly available datasets from the PhysioNet Computing in Cardiology Challenge 2021 (Reyna et al., 2021, 2022): the China 12-Lead ECG Challenge Database, the China 12-Lead ECG Challenge Database Extra, PTB-XL, the Georgia 12-Lead ECG dataset, the Chapman University dataset, the Shaoxing People's Hospital 12-Lead ECG Database, and the Ningbo First Hospital 12-Lead ECG Database. These datasets constitute a publicly accessible 12-lead ECG repository, encompassing multiple datasets with standardised labelling scores. The repository includes patient demographic information such as sex and age, but does not provide unique patient identifiers. Figure 1 illustrates the distribution of age, sex, and diagnostic categories across the merged dataset.

The following label mappings were applied to the investigated categories: **Sinus Rhythm** encompassed sinus rhythm, sinus bradycardia, and sinus tachycardia; **Atrial Fibrillation** corresponded to atrial fibrillation; and **Myocardial Infarction** included acute myocardial infarction, anterior myocardial infarction, and myocardial infarction.

The data was resampled to a unified frequency of 500 Hz, underwent pre-processing to remove low-frequency and high-frequency noise using a zero-phase filtering approach. Specifically, a third-order Butterworth bandpass filter (Butterworth, 1930) was applied with a frequency range of 1 Hz to 47 Hz to remove unwanted signal components while preserving relevant physiological information (Nejedly et al., 2021). The samples were randomly segmented to a fixed length of 4096 data points or zero-padded when necessary. Subsequently, z-score normalisation was applied to standardise the data across all samples.

2.2 Training and test datasets creation

To investigate the impact of varying sex ratios on model performance, a structured division of the dataset was essential. The accompanying graphical representations in Figure 2 visualise the training and test datasets creation process. Patients younger than 18 years, as well as those with missing values for sex or age, were excluded from the analysis. Due to the absence of unique patient identifiers, there was a heightened risk of data leakage, potentially resulting in overly optimistic performance estimates. Without unique identifiers, records from the same patient could inadvertently appear in both training and testing datasets, compromising the validity of model evaluation. To mitigate this, we implemented a strategy to minimise data leakage. Specifically, we generated pseudo-identifiers for patients based on their age, sex, and source dataset. Patients sharing identical values for these attributes were assigned exclusively to either the training or the testing set. Subsequently, patients were categorized into four distinct age groups: 18–40, 40–60, 60–80, and 80+ years. The dataset was then partitioned into five folds, with stratification carefully maintained across age groups, sex, data sources, and the diagnostic categories of interest, to ensure representative and unbiased splits.

From the fold designated as the test set, we created two separate test sets—one for male and one for female patients—by selecting an equal number of patients per diagnostic category while maintaining stratification for age group and data source. For the training sets, we constructed datasets with varying sex ratios: 0% female 100% male (F_0), 25% female 75% male (F_{25}), 50% female 50% male (F_{50}), 75% female 25% male (F_{75}), and 100% female 0% male (F_{100}). These proportions were maintained not only across the entire training set but also within each diagnostic category under investigation. 10% of the training set was utilised as a validation set, with the selection process ensuring the diagnostic distribution. Detailed distribution of training and test set sizes across categories stratified by age group can be found in the Appendix (Table 4) as well as test set distribution for each fold (Table 3). Furthermore, we ensured that the training set size remained consistent across different folds and experimental configurations.

2.3 Models

In this study, we evaluated the potential presence of sex-bias in deep learning-based ECG diagnosis using three distinct multi-class (3 classes) neural networks: a standard convolutional neural network (CNN) (LeCun et al., 1989), an xResNet101 (He et al., 2016), and the winning architecture from the PhysioNet Computing in Cardiology Challenge 2021 (Nejedly et al., 2021). The selected models span a range of architectural complexities, from relatively simple convolutional neural networks to deeper and more advanced residual networks with the attention mechanism. This diversity enables a comprehensive assessment of model performance and potential biases. Additionally, CNNs and residual networks are among the most widely used architectures for ECG classification tasks (Ansari et al., 2023).

The CNN model comprised four convolutional blocks, each including a convolutional layer with kernels sized [5, 5, 3, 3] and filter counts of [64, 64, 128, 128]. Each convolutional layer was immediately followed by a ReLU activation function and a max-pooling layer. The convolutional blocks were subsequently connected to two fully connected linear

Table 1: Integrated mean performance metrics—ROC AUC, pROC AUC, and PR AUC—for ResNet with Attention, xResNet101, and CNN on female (F) and male (M) test sets across varying sex ratios in the training data. Statistical significance of differences between female and male test sets is assessed using the Mann-Whitney U test (as in Larrazabal et al. (2020)) and indicated next to male values as follows: **** $(P \leq 0.0001)$, *** $(0.0001 < P \leq 0.001)$, ** $(0.001 < P \leq 0.01)$, * $(0.01 < P \leq 0.1)$; no marker indicates non-significant differences $(P > 0.1)$.

	ResNet with Attention						xResNet101						CNN					
Train	ROC AUC		pROC AUC		PR AUC		ROC AUC		pROC AUC		PR AUC		ROC AUC		pROC AUC		PR AUC	
Ratio	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M
Atrial Fibrillation																		
F_0	0.97	0.97	0.94	0.95	0.84	0.82	0.97	0.97	0.94	0.94	0.78	0.77	0.97	0.96**	0.92	0.92	0.74	0.73
F_{25}	0.97	0.97	0.95	0.95	0.84	0.82	0.97	0.97*	0.94	0.94	0.79	0.78	0.97	0.97**	0.93	0.93***	0.77	0.74*
F_{50}	0.99	0.98	0.97	0.96	0.88	0.86	0.97	0.97*	0.94	0.93**	0.79	0.76**	0.97	0.96**	0.93	0.92***	0.76	0.74*
F_{75}	0.98	0.98	0.96	0.96	0.86	0.86	0.98	0.97**	0.95	0.94***	0.81	0.78**	0.97	0.96***	0.94	0.92***	0.77	0.73
F_{100}	0.99	0.98	0.97	0.96	0.87	0.85	0.98	0.97***	0.95	0.93***	0.80	0.75***	0.97	0.96***	0.93	0.91***	0.75	0.71**
Sinus Rhythm																		
F_0	0.96	0.96	0.90	0.90	0.99	0.99	0.94	0.93	0.86	0.85	0.99	0.99	0.93	0.93	0.84	0.84	0.99	0.99
F_{25}	0.96	0.96	0.91	0.90	0.99	0.99*	0.94	0.93	0.86	0.85	0.99	0.99	0.94	0.92**	0.85	0.83	0.99	0.99**
F_{50}	0.97	0.96	0.92	0.91	0.99	0.99	0.94	0.93*	0.85	0.84	0.99	0.99	0.93	0.92*	0.85	0.83	0.99	0.99*
F_{75}	0.96	0.96	0.91	0.89	0.99	0.99	0.94	0.93**	0.87	0.85**	0.99	0.99**	0.93	0.92**	0.85	0.83*	0.99	0.99*
F_{100}	0.97	0.96	0.92	0.91	0.99	0.99	0.94	0.92***	0.86	0.82***	0.99	0.99***	0.93	0.91**	0.84	0.82*	0.99	0.98**
Myocardial Infarction																		
F_0	0.93	0.95*	0.86	0.89*	0.64	0.72**	0.94	0.95*	0.86	0.89**	0.64	0.72***	0.93	0.95**	0.85	0.88**	0.63	0.70***
F_{25}	0.94	0.94	0.87	0.88*	0.66	0.71	0.94	0.95*	0.87	0.89**	0.66	0.72***	0.94	0.95	0.86	0.88*	0.63	0.70***
F_{50}	0.95	0.95	0.88	0.89	0.69	0.73	0.94	0.95	0.87	0.89*	0.67	0.73***	0.94	0.95	0.87	0.88	0.64	0.70***
F_{75}	0.94	0.95	0.88	0.88	0.68	0.71	0.94	0.95	0.88	0.89	0.68	0.71	0.94	0.95	0.87	0.88	0.66	0.68
F_{100}	0.95	0.95	0.88	0.89	0.69	0.70	0.95	0.95	0.88	0.88	0.68	0.69	0.94	0.94	0.87	0.87	0.65	0.65

layers.

The xResNet101 architecture employed a deep residual learning approach, consisting of 101 layers and leveraging skip connections that facilitate the training of significantly deeper models without degradation in performance.

The PhysioNet Challenge-winning architecture was based on custom ResNet blocks optimised for ECG classification, using large convolutional kernels (15 in the first layer, 9 in residual layers with stride $2\times$). Outputs were processed through a multi-head attention mechanism and refined via adaptive max pooling. While the proposed architecture used an ensemble of models, we employed a single copy of the proposed model. Further in work we will be referring to this model as ResNet with Attention.

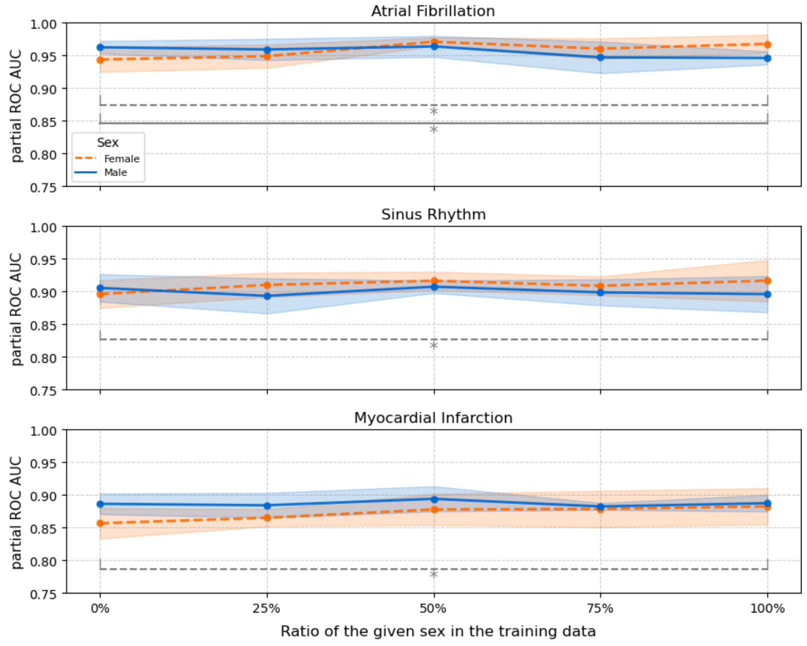
Given the significant class imbalance in our dataset, we used focal loss (Lin et al., 2017) for optimisation, with hyperparameters set at $\alpha = 0.75$ and $\gamma = 2$. The Adam optimiser (Kingma and Ba, 2014) was employed for model training. The initial learning rate was set at 0.001 and was decreased by a factor of 0.1 whenever the validation loss failed to improve for five consecutive epochs. Training was halted early if no improvement in loss was observed for five epochs after reaching the lowest permitted learning rate, with a maximum of two learning rate reductions allowed and an overall epoch limit set at 100. All experiments were conducted with a batch size of 128. Models were implemented in Python utilising the PyTorch library (Paszke et al., 2019).

2.4 Performance evaluation

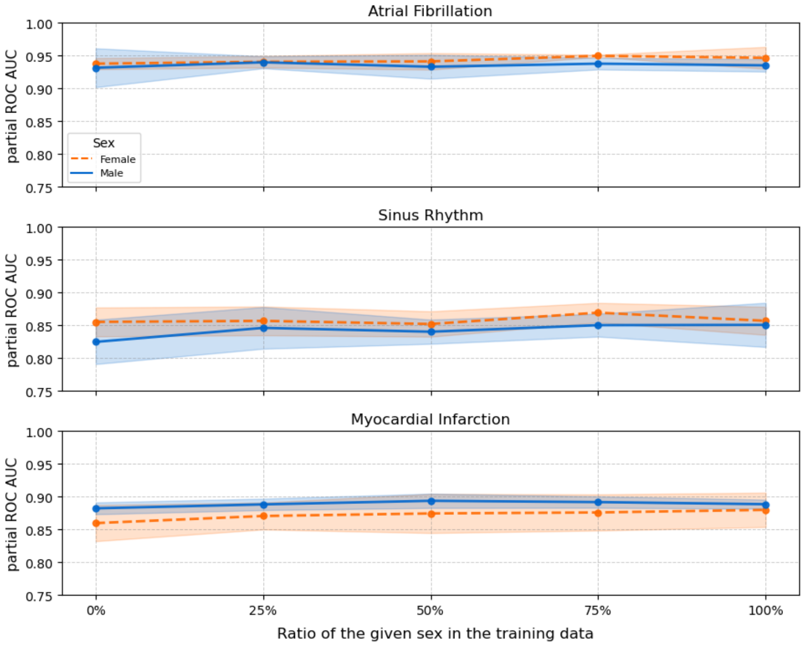
Our analysis focused on comparing model performance between male and female test sets under identical training conditions and investigating how variations in training set composition influenced results for each sex. Model performance was assessed using standard performance metrics: Receiver-operating characteristic curve area under the curve (ROC AUC) and precision-recall AUC (PR AUC). Due to high class imbalance, partial ROC AUC (pROC AUC) metric with a False Positive rate cut-off point at 0.20 has also been reported. All models were evaluated using 5-fold cross-validation to ensure robust and generalisable performance estimates. To investigate if there is a statistical significance difference between various training conditions among the two populations, we used Mann-Whitney U test and report significance following Larrazabal et al. (2020).

2.5 Additional experiments

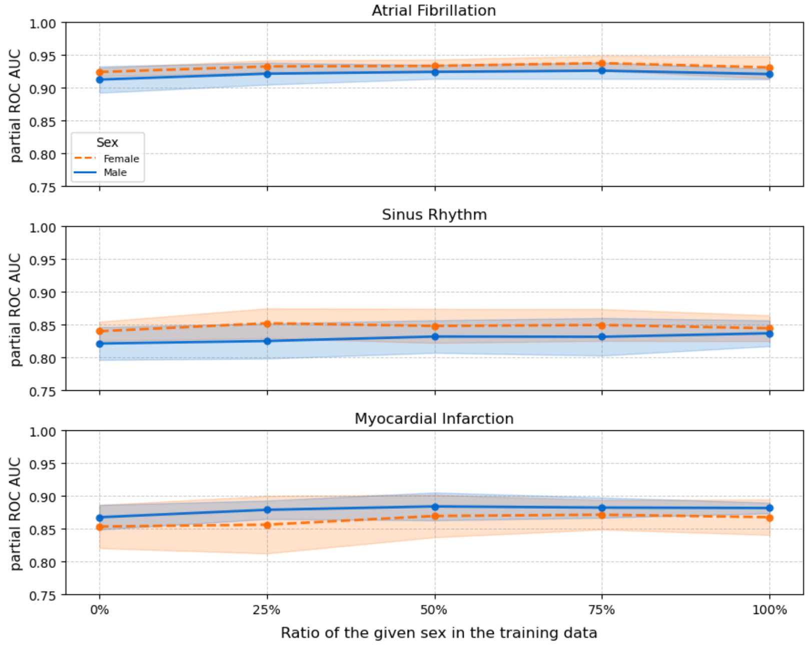
We conducted two additional analyses using the model that achieved the highest overall classification performance. First, we analysed performance stratified by patient age to investigate potential age-related variability in model accuracy. This evaluation aims to assess the model's consistency and reliability across different age groups.



(a) ResNet with Attention



(b) xResNet101



(c) CNN

Figure 3: Line plots showing partial pROC AUC performance for three models—ResNet with Attention (a), xResNet101 (b), and CNN (c)—on female (F) and male (M) test sets across varying sex ratios in the training data. The x-axis represents the percentage of each sex in the training set, and the y-axis shows the corresponding pROC AUC values. Shaded areas indicate 95% confidence intervals. Horizontal or grey lines highlight statistically significant differences between models trained with 0% and 100% sex representation based on the Mann-Whitney U test (see Table **??**). Absence of such lines denotes no significant difference.

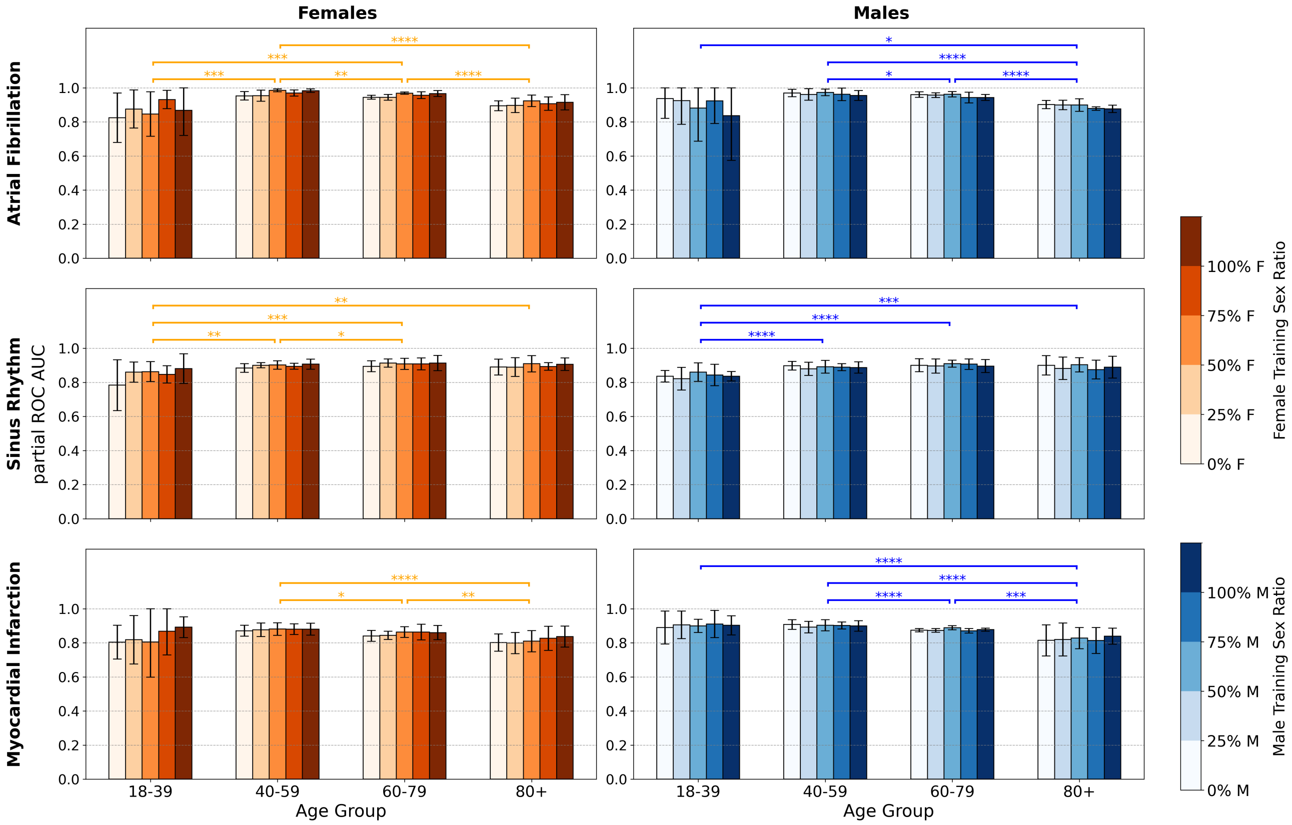


Figure 4: Bar plots showing partial ROC AUC performance across four age groups (18–39, 40–59, 60–79, and 80+) for three different classes (Atrial Fibrillation, Sinus Rhythm, and Myocardial Infarction), stratified by sex ratio in training. The left column shows results for the female test set (in orange), while the right column shows results for the male test set (in blue). Within each column, bar colour intensity increases with the proportion of the corresponding sex in the training data (e.g., 100% Female or 100% Male). Error bars represent 95% confidence intervals, and brackets with stars indicate statistical significance between specified age groups. Horizontal lines indicate statistically significant differences between age groups, based on the Mann-Whitney U test (see Table 1). Statistical significance notations are consistent with those used in Table 1, where asterisks ($*$) denote levels of significance; absence of a line indicates no significant difference.

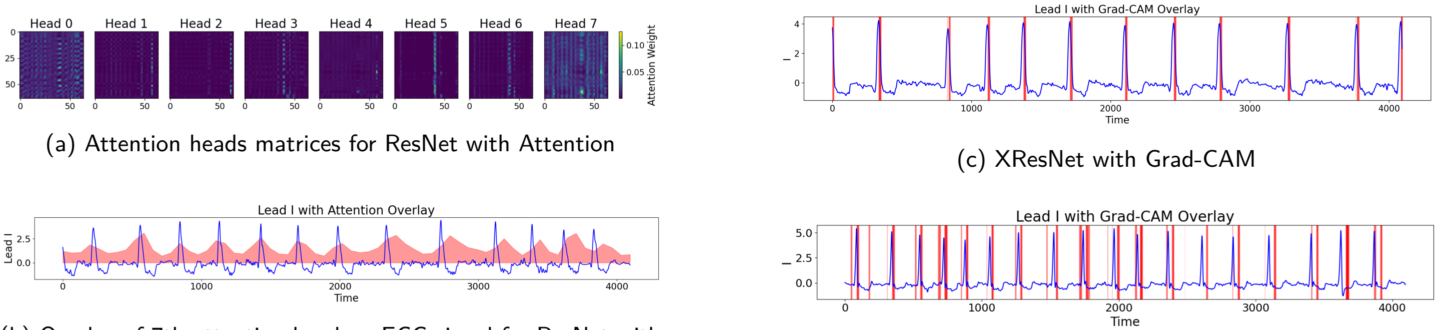
Second, motivated by prior evidence from Grzelak et al. that ECG leads III and V6 provide stronger representation of left atrial activity—particularly relevant for atrial fibrillation detection—we investigated whether sex-related disparities in classification accuracy vary with different combinations of ECG leads. Specifically, we evaluated model performance using only leads III and V6 to assess whether this lead configuration influences classification accuracy differently between sexes. This experiment was motivated by prior findings indicating that these leads are particularly informative for left atrial activity, and aimed to investigate whether their use introduces or mitigates sex-related disparities in model performance.

2.6 Interpretability analysis

To gain insight into the models' decision-making processes and better understand sources of performance disparities, we incorporated model interpretability techniques. For the ResNet with Attention architecture, we visualised the learned attention maps to highlight which regions of the ECG signal the model focuses on during prediction. For models without attention mechanisms, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) to generate gradient-based saliency visualisations. Additionally, we used Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) to project the latent feature representations from the penultimate layer of each trained model into a lower-dimensional space, with test samples annotated by sex. This approach enables a qualitative assessment of the structure of the learned embeddings and their relationship to demographic subgroups.

3. Results

The performance results for the three evaluated models: ResNet with Attention, xResNet101, and CNN—are sum-



(a) Attention heads matrices for ResNet with Attention

(b) Overlay of 7th attention head on ECG signal for ResNet with Attention

(c) XResNet with Grad-CAM

(d) CNN with Grad-CAM

Figure 5: Interpretability visualisations of different models for atrial fibrillation samples

Table 2: Mean performance metrics—ROC AUC, partial ROC AUC (pROC AUC), and precision-recall AUC (PR AUC)—for the ResNet with Attention model using only leads III and V6. Abbreviations and statistical significance notations are consistent with those used in Table 1.

ResNet with Attention (leads III and V6)						
	ROC AUC		pROC AUC		PR AUC	
	F	M	F	M	F	M
Atrial Fibrillation						
F_0	0.99	0.99	0.97	0.97	0.88	0.88
F_{25}	0.99	0.99	0.97	0.97	0.89	0.88
F_{50}	0.99	0.99	0.97	0.97	0.88	0.88
F_{75}	0.99	0.98	0.97	0.97	0.89	0.87
F_{100}	0.99	0.98	0.97	0.97	0.88	0.87
Sinus Rhythm						
F_0	0.96	0.96	0.90	0.90	0.99	0.99
F_{25}	0.96	0.96	0.90	0.90	0.99	0.99
F_{50}	0.96	0.96	0.91	0.90	0.99	0.99
F_{75}	0.96	0.95	0.90	0.89	0.99	0.99*
F_{100}	0.96	0.96	0.90	0.89	0.99	0.99
Myocardial Infarction						
F_0	0.90	0.93*	0.80	0.85**	0.51	0.64**
F_{25}	0.91	0.93*	0.81	0.85*	0.54	0.63*
F_{50}	0.90	0.92*	0.81	0.84	0.53	0.61*
F_{75}	0.90	0.92	0.80	0.83	0.52	0.57*
F_{100}	0.91	0.92	0.82	0.84	0.54	0.59*

marized in Table 1. This table reports mean AUC values across varying sex-based training data ratios, along with corresponding significance levels. To illustrate how model performance changes as the proportion of a given sex in the training data increases, Figure 3 depicts trends in mean partial ROC AUC. For each model, results on male test sets are shown in solid blue, while results on female test sets are represented by dashed orange lines. Performance is reported across three classes: AF, SR, and MI.

Further analysis of age-specific performance for the highest-performing model, ResNet with Attention, is presented in Figure 4. This figure displays bar plots of partial ROC AUC values stratified by four age groups (18–39, 40–59, 60–79, and 80+ years). Results are further stratified by the sex of the test data: outcomes for female test sets are shown in orange (left column), and those for male test sets in blue (right column). Within each group, darker bar shades represent a higher proportion of the corresponding sex in the training data (e.g., 100% Female or 100% Male).

Figure 5 shows interpretability visualisations for samples labelled as atrial fibrillation, focusing on Lead I. In Figure 5a, we display query–key attention maps across all heads for a given sample. Figure 5b overlays attention scores from the 7th head directly onto the normalised ECG waveform, highlighting relevant regions. Figure 5c presents Grad-CAM visualisations from the XResNet-101 model using the first layer of the third block. In Figure 5d, Grad-CAM is applied to the CNN model from its final convolutional layer, using a threshold of 0.2 to improve clarity.

In addition, we conducted a targeted experiment using only ECG leads III and V6, informed by prior work suggesting their enhanced sensitivity to left atrial activity relevant for atrial fibrillation detection (Grzelak et al.). The results, summarised in Table 2, present mean values for ROC AUC, pROC AUC, and PR AUC for the ResNet with Attention model, grouped by sex.

4. Discussion

This study demonstrates that sex imbalance in training data can impact the performance and generalisation of deep learning models in 12-lead ECG classification, with effects varying by task, age group, and model architecture. Among the evaluated models, the ResNet Attention Mechanism emerged as the most robust and equitable across sexes, particularly in the classification of AF and SR. Its performance remained relatively unbiased across varying sex ratios, suggesting that certain architectural features—such as attention mechanisms—may confer greater resilience to demographic bias. Our findings also suggest that overrepresentation of one sex in training does not guarantee improved performance for that subgroup. Instead, balanced or female-inclusive training datasets tend to yield more generalisable models, even enhancing outcomes for male patients. This raises critical questions about the utility of developing entirely separate models for different demographic groups and underscores the importance of fairness-aware model design.

Age-stratified analysis further revealed that model performance peaks in middle-aged groups (40–79 years), while younger (18–39) and oldest (80+) cohorts showed greater variability, likely due to sample scarcity and higher comorbidity rates, respectively. These age effects intersect with sex-based disparities and reinforce the need for demographically representative datasets during training.

Finally, our results highlight a persistent performance gap in MI classification between male and female patients, even under-balanced training. This may stem from physiological sex differences in ECG presentation, diagnostic interpretation challenges, or disparities in the quality of care provided—particularly for females experiencing MI.

General models performance

Among the models, ResNet with Attention consistently achieved the highest or comparable performance across all metrics and classification tasks, regardless of sex or training ratio. For instance, in the MI task with balanced training data (F_{50}), the ResNet with Attention reached a ROC AUC of 0.95 for both females and males, a pROC AUC of 0.88/0.89, and PR AUC of 0.69/0.73, respectively. xResNet101 followed closely, while CNN trailed in most tasks, particularly in PR AUC—suggesting it struggles more with correctly identifying positive cases under class imbalance.

Sex-specific performance across varying training set composition

All three evaluated models—ResNet with Attention, xResNet101, and CNN—exhibited varying levels of disparity between sexes when classifying AF and SR across different sex ratios in the training data. Performance variation was more pronounced in MI classification, where all models' sensitivity to sex-based imbalances became more apparent. While ROC AUC scores remained consistently high—partly due to class imbalance—pROC AUC and PR AUC provided deeper insight into performance trade-offs and bias, particularly under skewed data conditions.

ResNet with Attention demonstrated the smallest sex-based disparities. For both AF and SR classification, the best results were achieved when the training data had either a balanced sex ratio or consisted entirely of female samples—a pattern observed for both male and female test sets. In the AF task under fully balanced training (F_{50}), the PR AUC was 0.88 for females and 0.86 for males—a gap of 0.02. Similar trends were observed across other ratios and diagnostic categories. These performance differences between male and female test sets were not statistically significant across the various sex ratios, suggesting that the model maintained equitable performance regardless of training composition. In contrast, xResNet101 showed increasing performance disparity between sexes as the proportion of female samples in the training data grew, particularly for AF and SR. In the AF task under F_{100}, female PR AUC reached 0.80, while male PR AUC dropped to 0.75-a statistically significant difference. Likewise, in SR classification under F_{100}, male pROC AUC fell to 0.82, while female performance reached 0.86. CNN exhibited the most pronounced sex-based discrepancies. In the AF task at F_{75}, female PR AUC was 0.77 compared to 0.73 for males.

Across all models MI classification exhibited the largest sex-based performance gaps. Even under fully balanced training (F_{50}), the PR AUC for MI in females remained consistently lower than in males (ResNet with Attention: 0.69 vs. 0.73; xResNet: 0.67 vs. 0.73; CNN: 0.64 vs. 0.70). Notably, the PR AUC gap of 0.09 in the CNN model (F_0) was the widest, highlighting how female predictions suffer in precision. Even with balanced training, female PR AUC scores for MI remain lower, suggesting that performance disparities cannot be fully attributed to dataset imbalance. These persistent gaps might stem from underlying physiological factors, labelling inaccuracies, as well as possible historical disparities in diagnosis or treatment.

As illustrated in Figure 3a, which plots pROC AUC for the highest-performing model (ResNet with Attention) across varying sex ratios in the training data, a noteworthy trend emerges for AF and SR: performance on male test sets decreases as the proportion of male samples in the training set increases. In the AF classification task, models trained exclusively on female data (F_{100}) achieved higher pROC AUC scores (0.96) on male test sets than models trained solely on male data (F_0: 0.95), with the difference being statistically significant. On the other hand, female test performance increased consistently with greater female representation in training, across all diagnostic categories. This suggests that training on female ECGs may enhance generalisation even for male patients. A possible explanation lies in physiological differences, such as females' higher average heart rate Kittnar (2023), which results in more cardiac cycles—and potentially more informative patterns—within fixed-length ECG segments. These findings highlight the asymmetric impact of training data composition on model performance and underscore the potential benefits of incorporating female-rich data for improved generalisability.

Age-specific performance for ResNet with Attention

Figure 4 illustrates the partial ROC AUC performance of ResNet with Attention across age groups (18–39, 40–59, 60–79, and 80+) with results stratified by sex and training set composition. A general trend emerges, showing more stable performance in the middle-aged population (40–79 years).

Female test performance in AF peaked in the 40–59 and 60–79 age groups, especially under female-heavy training conditions. The 18–39 group consistently lagged, but this is likely due to the sparse representation of this age group for AF and MI classes (on average 3 males and 6 females per fold), limiting statistical power. In male test sets, AF performance also favoured the middle-aged groups. Notably, models trained with lower male representation performed better on male test sets, again indicating that female data may provide features with broader generalisability.

SR classification results remained similar across female age groups, but the lowest and most unstable performance was observed in the 18–39 group when training used only male data, emphasising the detrimental effect of sex-mismatched training for the female population. Male SR classification showed minimal variation across age, with a slight dip in the 40–79 age range as male representation increased in training. This further supports the idea that over-representation of one sex does not equate to optimal performance for that sex.

MI classification showed the most pronounced age-related trends among the diagnostic tasks. Female pROC AUC scores peak in the 40–59 and 60–79 age groups. The most substantial performance gains were observed when comparing various training settings in the youngest group (18–39); however, due to the under-representation of the young population in the dataset for MI, results for this group should be interpreted with caution. For male test sets, the highest MI classification performance was observed in the 40–59 age group. Balanced training data consistently yields strong results across all male age groups above 39. Interestingly, for the 80+ age group in both sexes, the highest classification performance was achieved when models were trained on data exclusively from the same sex. However, this group also exhibited wide confidence intervals, indicating greater performance variability and lower reliability.

Interpretability analysis

Figure 5 offers interpretability visualisations for samples labelled as AF, revealing how different model architectures process and prioritise input features. In Figure 5a, the attention matrices across all eight heads show structured patterns, with many heads consistently attending to periodically aligned regions. These patterns reflect the model's ability to capture and cyclically attend to recurring temporal features in the ECG waveform—such as QRS complexes or R-R intervals—critical for detecting AF. This temporal self-alignment might allow for distributed and context-aware representation learning. As illustrated in Figure 5b, the 7th attention head overlays a smooth and physiologically coherent importance pattern over the ECG trace, reinforcing the model's capacity to attend to features spanning longer time horizons, not just directly neighbouring ones, suggesting that the model is attending not only to local events but also to temporally extended dependencies.

The Grad-CAM visualisations from XResNet (Figure 5c) and CNN (Figure 5d) highlight prominent features like R-peaks, demonstrating both models' ability to detect key ECG patterns. A 0.2 threshold was applied to enhance clarity, with visualisations shown from a single network layer. However, due to their underlying architectures, these models lack explicit mechanisms for capturing long-range temporal and cyclic dependencies. This limitation stems from their architectural design: while xResNet incorporates residual connections that mitigate vanishing gradients and enable deeper feature integration, allowing it to capture longer-term dependencies than standard CNNs, it still lacks the explicit attention mechanism required for capturing dependencies over extended and periodic intervals. Classical CNNs without residual connections remain constrained to local spatial features.

This architectural constraint stands in contrast to the attention-based ResNet, which is specifically designed to integrate information across long-range temporal dependencies. As a result, it produces a more distributed and temporally nuanced representation of the ECG signal. These differences in model design may help explain observed performance disparities, particularly across demographic subgroups. For instance, sex-related physiological characteristics—such as the typically faster heart rates and distinct P-wave morphology seen in females with AF—can lead to subtle variations in ECG patterns. Attention-based models, by capturing the global context of the signal, are better positioned to detect these nuanced and dispersed features. This capability likely contributes to their superior overall performance and greater fairness across different demographic groups.

Additionally, we applied UMAP to visualise (Figure 6 in the Appendix) the latent space representations. The resulting UMAP plots did not show a clear separation between male and female samples, regardless of the model's overall predictive performance. This indicates that, in the learned latent space, sex is not a primary axis of variation for the models. Thus, the models do not appear to encode sex as a dominant feature in their final representations, suggesting that observed performance differences are not due to explicit clustering by sex but may arise from more subtle interactions between ECG features and sex-related physiology.

ResNet with Attention using Leads III and V6

This experiment assessed whether using only leads III and V6—known to better capture left atrial activity (Grzelak et al.)—affects sex-related disparities in ECG classification performance. For AF, results using only leads III and V6 matched or slightly outperformed those from the 12-lead configuration. Moreover, compared to 12-lead settings, the results are less influenced by the training set ratio, with no significant differences observed between male and female test sets. This supports the clinical relevance of these leads for atrial signal detection and suggests that lead selection can reduce sex-based disparities in AF classification. For SR, performance remained high and comparable across both configurations, with no substantial sex differences observed. In contrast, for MI, reducing to two leads led to a noticeable drop in performance, especially for female patients. Significant gaps in PR AUC and pROC AUC between sexes were observed across all training ratios, indicating that leads III and V6 are insufficient for capturing MI-related features, which are more spatially distributed. Overall, these findings suggest that targeted lead selection may improve fairness and efficiency for some conditions where left atrial activity and P-wave morphology are central(e.g., AF), but can introduce or amplify disparities in others (e.g., MI), depending on the underlying cardiac pathology and signal distribution.

4.1 Limitations and Future Work

Despite the insights offered by this study, several limitations should be acknowledged.

Firstly, although we investigated three deep learning architectures, the findings may not generalise to other model types or clinical tasks beyond ECG-based classification of AF, SR, and MI. Further analysis should focus on the emerging class of foundation models for ECG analysis, such as ECG-FM (McKeen et al., 2024). However, these models are typically trained on large collections of datasets—including the ones used in this study—making it infeasible to fairly evaluate them under the current experimental design. Future work could assess how these large-scale models address or amplify bias, and whether their generalisation comes at the cost of subgroup performance consistency.

Given the strong performance and fairness characteristics of the attention-based ResNet model observed in our study, future research could explore combining this architecture with complementary fairness-enhancing strategies—such as data augmentation to address subgroup imbalance, debiasing techniques during training, or post-hoc calibration—may further improve model equity and generalisability.

Secondly, the dataset exhibited inherent demographic imbalances, particularly in the younger (18–39) age group for AF and MI. These subgroups were under-represented, limiting the robustness of subgroup analyses. Additionally, the absence of unique patient identifiers means that, despite our use of a pseudo-identifier strategy to minimise overlap between training and test sets, we cannot fully exclude the possibility of residual data leakage.

Third, our evaluation was limited to binary biological sex as a proxy for physiological differences. This binary approach may overlook critical nuances in ECG patterns influenced by hormone levels, comorbidities, and social determinants of health, potentially failing to capture the full spectrum of cardiovascular risk within each sex (Sau et al., 2025). Future research direction could involve integrating more comprehensive clinical and demographic metadata, such as hormonal profiles, BMI, and medication use.

Our study revealed a persistent performance gap in MI classification between males and females, even under balanced training conditions. While all models showed higher precision-recall performance on male test sets, performance for females consistently lagged behind. This gap may stem from sex-based differences in ECG manifestations of MI, under-representation of females in historical diagnostic datasets, or differences in clinical care pathways. Future research should investigate whether these gaps are a result of physiological signal differences, data labelling inaccuracies, or broader clinical biases, and whether fairness-aware training strategies could help mitigate performance disparities.

5. Conclusion

In conclusion, this study emphasises that fairness investigations are essential in clinical machine learning applications, particularly when working with heterogeneous populations. Physiological differences between sexes do not inherently degrade model performance, but if unaddressed, training biases can lead to systematic disparities. While ensuring representative training data may improve model fairness and generalisation, its impact should be carefully assessed and validated. At the same time, model architecture appears to play a critical role in mitigating bias and enhancing equity in AI-driven healthcare systems. In particular, architectures that incorporate mechanisms for capturing long-range temporal and cyclical dependencies—such as attention—demonstrate not only superior overall performance but also greater robustness to demographic variability, suggesting that architectural design choices can directly influence a model's fairness. Finally, targeted lead selection may help improve fairness in some conditions (e.g., AF) but may worsen it in others, underscoring the need for condition-specific evaluation.

Acknowledgments

This work was supported by the University of Amsterdam Research Priority Area Artificial Intelligence for Health

Decision-making.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding the treatment of animals or human subjects.

Conflicts of Interest

APJV consulting fees and grants from Edwards Lifesciences/BD paid to the institution.

Data availability

The datasets used in this study are publicly available from the official PhysioNet *Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021* repository (Reyna et al., 2021, 2022). The repository is accessible at https://physionet.org/content/challenge-2021/1.0.3/. The code developed for this project is available at https://github.com/MariaGalanty/Investigating-sex-bias-in-ECG-classification.git.

References

Erick A Perez Alday, Ali B Rad, Matthew A Reyna, Nadi Sadr, Annie Gu, Qiao Li, Mircea Dumitru, Joel Xue, Dave Albert, Reza Sameni, et al. Age, sex and race bias in automated arrhythmia detectors. *Journal of electrocardiology*, 74:5–9, 2022. DOI: 10.1016/j.jelectrocard.2022.07.007.

Yaqoob Ansari, Omar Mourad, Khalid Qaraqe, and Erchin Serpedin. Deep learning for ecg arrhythmia detection and classification: an overview of progress for period 2017–2023. *Frontiers in Physiology*, 14:1246746, 2023. DOI: 10.3389/fphys.2023.1246746.

Stephen Butterworth. On the theory of filter amplifiers. *Wireless Engineer*, 7(6):536–541, 1930.

Lucilla Giammarino, Lluis Matas, Nicolò Alerni, András Horváth, Varjany Vashanthakumar, Saranda Nimani, Miriam Barbieri, Sahej Bains, Ruben Lopez, Julien Louradour, et al. Sex and sex hormonal regulation of the atrial inward rectifier potassium current (ik1): insights into potential pro-arrhythmic mechanisms. *Cardiovascular Research*, page cvaf074, 2025. DOI: 10.1093/cvr/cvaf074.

Jakub Grzelak, Shaheim Ogbomo-Harmitt, Andrew P King, Fu Siong Ng, and Oleg Aslanidi. In-silico investigation of the right and left atrial contributions to the p-wave morphology in ecg of healthy and atrial fibrillation patients. URL `https://cinc.org/archives/2024/pdf/CinC2024-122.pdf`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Dhamanpreet Kaur, J Weston Hughes, Albert J Rogers, Guson Kang, Sanjiv M Narayan, Euan A Ashley, and Marco V Perez. Race, sex, and age disparities in the performance of ecg deep learning models predicting heart failure. *Circulation: Heart Failure*, 17(1):e010879, 2024. DOI: 10.1161/CIRCHEARTFAILURE.123.010879.

Amit Kaushal, Russ Altman, and Curt Langlotz. Geographic distribution of us cohorts used to train deep learning algorithms. *Jama*, 324(12):1212–1213, 2020. DOI: 10.1001/jama.2020.12067.

Saad M Khan, Xiaoxuan Liu, Siddharth Nath, Edward Korot, Livia Faes, Siegfried K Wagner, Pearse A Keane, Neil J Sebire, Matthew J Burton, and Alastair K Denniston. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health*, 3(1): e51–e66, 2021. DOI: 10.1016/S2589-7500(20)30240-5.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. DOI: 10.48550/arXiv.1412.6980.

Otomar Kittnar. Sex related differences in electrocardiography. *Physiological Research*, 72(Suppl 2):S127, 2023. DOI: 10.33549/physiolres.934952.

Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020. DOI: 10.1073/pnas.1919012117.

Rita Laureanti, Giulio Conte, Valentina DA Corino, Stefan Osswald, David Conen, Laurent Roten, Nicolas Rodondi, Peter Ammann, Christine S Meyer-Zuern, Leo Bonati, et al. Sex-related electrocardiographic differences in patients with different types of atrial fibrillation: Results from the swiss-af study. *International journal of cardiology*, 307:63–70, 2020. DOI: 10.1016/j.ijcard.2019.12.053.

Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

Tiarna Lee, Esther Puyol-Antón, Bram Ruijsink, Keana Aitcheson, Miaojing Shi, and Andrew P King. An investigation into the impact of deep learning model choice on sex and race bias in cardiac mr segmentation. In *Workshop on Clinical Image-Based Procedures*, pages 215–224. Springer, 2023. DOI: 10.1007/978-3-031-45249-9$_2$1.

Tiarna Lee, Esther Puyol-Antón, Bram Ruijsink, Sebastien Roujol, Theodore Barfoot, Shaheim Ogbomo-Harmitt, Miaojing Shi, and Andrew King. An investigation into the causes of race bias in artificial intelligence–based cine cardiac magnetic resonance segmentation. *European Heart Journal-Digital Health*, 6(3):350–358, 2025. DOI: 10.1093/ehjdh/ztaf008.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. DOI: 10.1109/TPAMI.2018.2858826.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. DOI: 10.48550/arXiv.1802.03426.

Kaden McKeen, Laura Oliva, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model. *arXiv preprint arXiv:2408.05178*, 2024. DOI: 10.48550/arXiv.2408.05178.

Petr Nejedly, Adam Ivora, Radovan Smisek, Ivo Viscor, Zuzana Koscova, Pavel Jurak, and Filip Plesinger. Classification of ecg using ensemble of residual cnns with attention mechanism. In *2021 computing in cardiology (CinC)*, volume 48, pages 1–4. IEEE, 2021. DOI: 10.1088/1361-6579/ac647c.

Peter A Noseworthy, Zachi I Attia, LaPrincess C Brewer, Sharonne N Hayes, Xiaoxi Yao, Suraj Kapa, Paul A Friedman, and Francisco Lopez-Jimenez. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ecg analysis. *Circulation: Arrhythmia and Electrophysiology*, 13(3): e007988, 2020. DOI: 10.1161/CIRCEP.119.007988.

Katja E Odening, Sebastian Deiß, Dagmara Dilling-Boer, Maxim Didenko, Urs Eriksson, Sotirios Nedios, Fu Siong Ng, Ivo Roca Luque, Pepa Sanchez Borque, Kevin Vernooy, et al. Mechanisms of sex differences in atrial fibrillation: role of hormones and differences in electrophysiology, structure, function, and remodelling. *EP Europace*, 21(3):366–376, 2019. DOI: 10.1093/europace/euy215.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Eike Petersen, Aasa Feragen, Maria Luise da Costa Zemsch, Anders Henriksen, Oskar Eiler Wiese Christensen, Melanie Ganz, and Alzheimer's Disease Neuroimaging Initiative. Feature robustness and sex differences in medical imaging: a case study in mri-based alzheimer's disease detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2022. DOI: 10.1007/978-3-031-16431-6$_9$.

Matthew Reyna, Nadi Sadr, Annie Gu, Erick Andres Perez Alday, Chengyu Liu, Salman Seyedi, Amit Shah, and Gari Clifford. Will two do? varying dimensions in electrocardiography: The physionet/computing in cardiology challenge 2021, 2022. URL `https://doi.org/10.13026/34va-7q14`.

Matthew A Reyna, Nadi Sadr, Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, et al. Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021. In *2021 computing in cardiology (CinC)*, volume 48, pages 1–4. IEEE, 2021. DOI: 10.13026/gt86-a263.

Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1760, 2020. DOI: 10.1038/s41467-020-15432-4.

Gregory A Roth, George A Mensah, Catherine O Johnson, Giovanni Addolorato, Enrico Ammirati, Larry M Baddour, Noël C Barengo, Andrea Z Beaton, Emelia J Benjamin, Catherine P Benziger, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the gbd 2019 study. *Journal of the American college of cardiology*, 76(25):2982–3021, 2020. DOI: 10.1016/j.jacc.2020.11.010.

Arunashis Sau, Ewa Sieliwonczyk, Konstantinos Patlatzoglou, Libor Pastika, Kathryn A McGurk, Antônio H Ribeiro, Antonio Luiz P Ribeiro, Jennifer E Ho, Nicholas S Peters, James S Ware, et al. Artificial intelligence-enhanced electrocardiography for the identification of a sex-related cardiovascular risk continuum: a retrospective cohort study. *The Lancet Digital Health*, 7(3):e184–e194, 2025. DOI: 10.1016/j.landig.2024.12.003.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Jagmeet P Singh, Julien Fontanarava, Grégoire de Massé, Tanner Carbonati, Jia Li, Christine Henry, and Laurent Fiorina. Short-term prediction of atrial fibrillation from ambulatory monitoring ecg using a deep neural network. *European Heart Journal-Digital Health*, 3(2):208–217, 2022. DOI: 10.1093/ehjdh/ztac014.

Kardie Tobb, Madison Kocher, and Renée P Bullock-Palmer. Underrepresentation of women in cardiovascular trials-it is time to shatter this glass ceiling. *American Heart Journal Plus: Cardiology Research and Practice*, 13:100109, 2022. DOI: 10.1016/j.ahjo.2022.100109.

Kerstin N Vokinger, Stefan Feuerriegel, and Aaron S Kesselheim. Mitigating bias in machine learning for medicine. *Communications medicine*, 1(1):25, 2021. DOI: 10.1038/s43856-021-00028-w.

Emily P Zeitler, Jeanne E Poole, Christine M Albert, Sana M Al-Khatib, Fatima Ali-Ahmed, Ulrika Birgersdotter-Green, Yong-Mei Cha, Mina K Chung, Anne B Curtis, Jodie L Hurwitz, et al. Arrhythmias in female patients: incidence, presentation and management. *Circulation research*, 130(4): 474–495, 2022. DOI: 10.1161/CIRCRESAHA.121.319893.

Jianwei Zheng, Chizobam Ani, Islam Abudayyeh, Yunfan Zheng, Cyril Rakovski, Ehsan Yaghmaei, and Omolola Ogunyemi. A review of racial differences and disparities in ecg. *International Journal of Environmental Research and Public Health*, 22(3):337, 2025. DOI: 10.3390/ijerph22030337.

Appendix A. Tables and Figures

Table 3: Number of samples per diagnostic category, stratified by age group and sex, for the test set in each fold. The abbreviations used are Atrial Fibrillation (AF), Sinus Rhythm (SR), Myocardial Infarction (MI), Females (F) and Males (M).

	AF	**SR**	**MI**	**Other**	**AF**	**SR**	**MI**	**Other**
Age	F				M			
Fold 0								
18-39	3	817	9	21	1	663	18	17
40-59	32	2104	89	35	73	1572	61	45
60-79	310	1300	228	70	307	1993	244	112
80+	163	610	115	49	127	603	118	34
Total	508	4831	441	175	508	4831	441	208
Fold 1								
18-39	7	1117	6	15	4	604	32	10
40-59	32	1244	48	65	83	2204	176	97
60-79	281	1980	220	94	316	1654	283	94
80+	184	490	257	22	101	369	40	30
Total	504	4831	531	196	504	4831	531	231
Fold 2								
18-39	1	433	4	28	3	595	8	15
40-59	57	1572	104	32	55	1788	200	15
60-79	125	2225	337	101	108	1817	221	111
80+	72	235	18	27	89	265	34	11
Total	255	4465	463	188	255	4465	463	152
Fold 3								
18-39	6	659	2	21	5	500	4	13
40-59	54	2063	78	61	83	1042	100	80
60-79	237	1694	141	137	295	2960	182	100
80+	197	508	120	63	111	422	55	28
Total	494	4924	341	282	494	4924	341	221
Fold 4								
18-39	4	1059	11	3	2	1026	19	26
40-59	34	2095	51	56	65	2061	96	78
60-79	214	2090	203	90	279	2506	265	57
80+	185	567	149	36	91	218	34	15
Total	437	5811	414	185	437	5811	414	176

Table 4: The mean and standard deviation of the number of samples for each diagnostic category, grouped by age and presented separately for females (F) and males (M), are reported for both test sets and the training sets under each training regime.

	AF		SR		MI		Other	
Age	F	M	F	M	F	M	F	M
Test Set								
18-39	4±2	3±2	817±283	678±203	6±4	16±11	18±9	16±6
40-59	42±13	72±12	1816±390	1733±457	74±24	127±59	50±15	63±33
60-79	233±71	261±87	1858±368	2186±538	226±71	239±39	99±24	95±22
80+	160±51	103±16	482±146	376±151	132±86	56±36	39±17	24±10
Total	440±107	440±107	4972±501	4972±501	438±69	438±69	205±44	198±33
0% Females 100% Males								
18-39	0	14±2	0	2757±184	0	59±11	0	577±62
40-59	0	291±12	0	7239±377	0	541±63	0	1876±135
60-79	0	984±102	0	8820±603	0	957±45	0	4330±270
80+	0	453±28	0	1526±144	0	201±25	0	1670±148
Total	0	1742±89	0	20341±575	0	1758±70	0	8453±369
25% Females 75% Males								
18-39	7±1	45±8	833±70	2067±137	4±1	11±1	232±9	435±46
40-59	74±6	406±47	1866±110	5427±290	42±3	220±8	465±29	1407±97
60-79	227±18	717±34	1897±88	6618±456	233±20	738±74	1006±73	3246±200
80+	132±21	151±19	497±36	1146±107	159±12	340±21	415±35	1251±114
Total	440±18	1318±52	5094±145	15257±434	438±29	1309±60	2118±106	6339±273
50% Females 50% Males								
18-39	13±2	30±5	1668±141	1378±89	8±1	7±0	461±17	290±32
40-59	149±12	270±31	3729±217	3615±191	84±7	149±7	934±57	938±64
60-79	453±35	478±23	3790±182	4413±298	465±33	493±49	2018±137	2161±135
80+	265±43	101±12	986±75	764±69	318±23	228±14	836±67	833±78
Total	880±35	879±35	10173±290	10170±289	875±48	876±38	4249±202	4222±185
75% Females 25% Males								
18-39	19±3	15±3	2501±211	688±44	13±2	3±0	694±26	145±17
40-59	223±18	135±16	5585±325	1806±96	126±11	75±5	1410±86	467±36
60-79	680±53	239±11	5687±275	2208±148	697±54	245±26	3021±210	1078±70
80+	397±64	50±6	1479±114	379±33	480±37	114±9	1252±99	418±39
Total	1319±52	439±18	15252±423	5082±137	1316±75	437±16	6377±309	2109±99
100% Females 0% Males								
18-39	26±4	0	3334±280	0	17±2	0	926±34	0
40-59	298±25	0	7443±433	0	167±13	0	1884±116	0
60-79	906±71	0	7579±361	0	934±71	0	4028±281	0
80+	529±86	0	1964±150	0	642±51	0	1674±134	0
Total	1758±70	0	20320±563	0	1760±107	0	8511±410	0

Table 5: Results for ResNet with Attention including means, standard deviation and p-values (p-val) for ROC AUC, partial ROC AUC (pROC AUC), and precision-recall AUC (PR AUC)

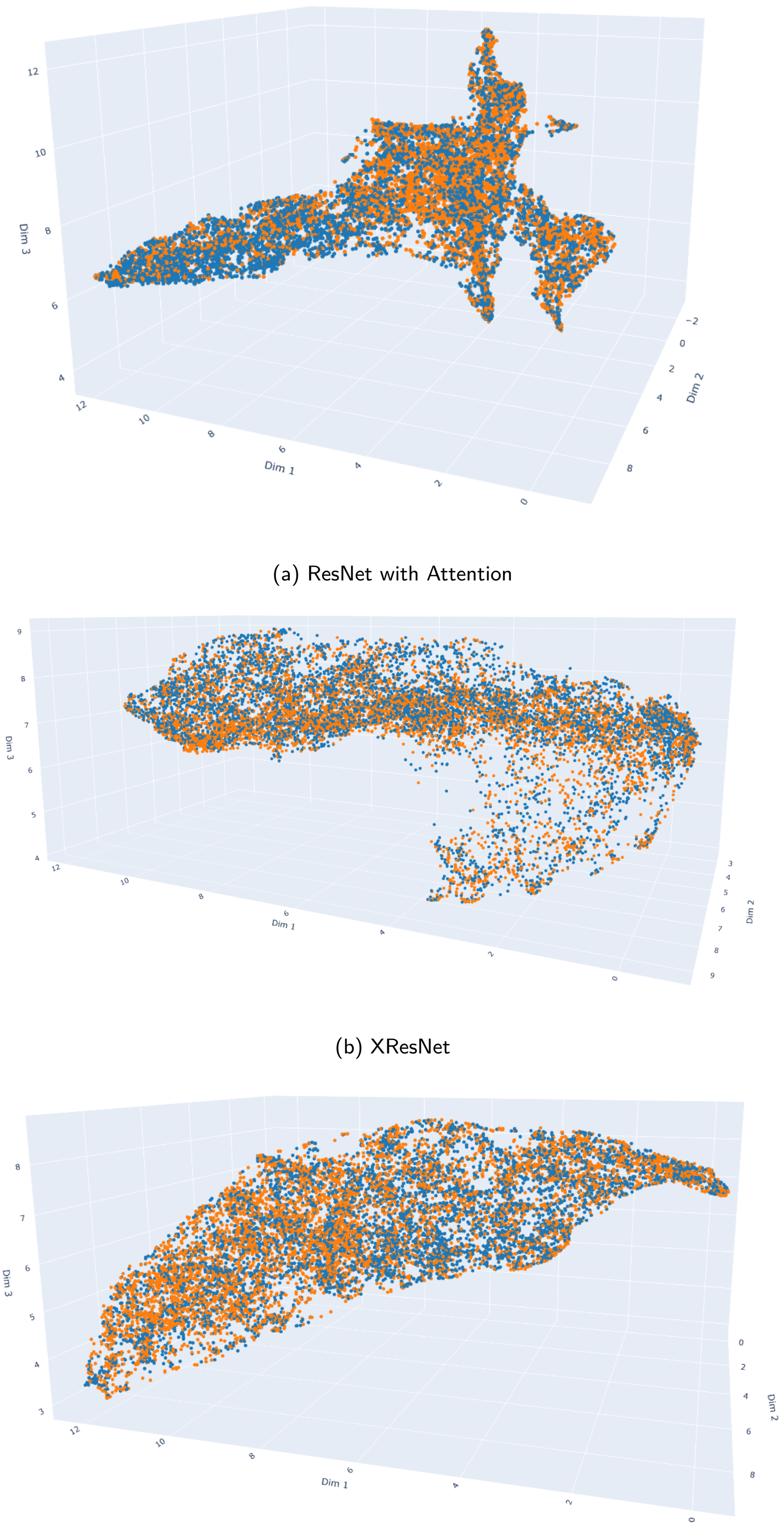
	ROC AUC			pROC AUC			PR AUC		
	F	M	p-val	F	M	p-val	F	M	p-val
Atrial Fibrillation									
F_0	0.97 ± 0.0	0.97 ± 0.01	0.4206	0.94 ± 0.02	0.95 ± 0.01	0.8413	0.84 ± 0.04	0.82 ± 0.01	0.1508
F_{25}	0.97 ± 0.01	0.97 ± 0.01	0.8413	0.95 ± 0.01	0.95 ± 0.02	1.0	0.84 ± 0.02	0.82 ± 0.04	0.4206
F_{50}	0.99 ± 0.0	0.98 ± 0.01	0.3095	0.97 ± 0.01	0.96 ± 0.01	0.3095	0.88 ± 0.02	0.86 ± 0.02	0.2222
F_{75}	0.98 ± 0.01	0.98 ± 0.01	0.6905	0.96 ± 0.01	0.96 ± 0.01	1.0	0.86 ± 0.04	0.86 ± 0.03	1.0
F_{100}	0.99 ± 0.01	0.98 ± 0.0	0.2222	0.97 ± 0.01	0.96 ± 0.01	0.3095	0.87 ± 0.04	0.85 ± 0.04	0.3095
Sinus Rhythm									
F_0	0.96 ± 0.01	0.96 ± 0.01	1.0	0.9 ± 0.02	0.9 ± 0.02	0.8413	0.99 ± 0.0	0.99 ± 0.0	0.8413
F_{25}	0.96 ± 0.01	0.96 ± 0.01	0.3095	0.91 ± 0.02	0.9 ± 0.02	0.3095	0.99 ± 0.0	0.99 ± 0.0	0.0952
F_{50}	0.97 ± 0.0	0.96 ± 0.0	0.3095	0.92 ± 0.01	0.91 ± 0.01	0.1508	0.99 ± 0.0	0.99 ± 0.0	0.5476
F_{75}	0.96 ± 0.01	0.96 ± 0.01	0.3095	0.91 ± 0.01	0.89 ± 0.02	0.2222	0.99 ± 0.0	0.99 ± 0.0	0.2222
F_{100}	0.97 ± 0.01	0.96 ± 0.01	0.2222	0.92 ± 0.03	0.91 ± 0.02	0.2222	0.99 ± 0.0	0.99 ± 0.0	0.1508
Myocardial Infarction									
F_0	0.93 ± 0.01	0.95 ± 0.01	0.0159	0.86 ± 0.02	0.89 ± 0.01	0.0317	0.64 ± 0.02	0.72 ± 0.04	0.0079
F_{25}	0.94 ± 0.01	0.94 ± 0.01	0.2222	0.87 ± 0.01	0.88 ± 0.0	0.0952	0.66 ± 0.03	0.71 ± 0.04	0.1508
F_{50}	0.95 ± 0.01	0.95 ± 0.01	0.5476	0.88 ± 0.02	0.89 ± 0.02	0.2222	0.69 ± 0.02	0.73 ± 0.05	0.4206
F_{75}	0.94 ± 0.01	0.95 ± 0.01	1.0	0.88 ± 0.02	0.88 ± 0.02	0.5476	0.68 ± 0.03	0.71 ± 0.05	0.5476
F_{100}	0.95 ± 0.01	0.95 ± 0.01	1.0	0.88 ± 0.02	0.89 ± 0.01	1.0	0.69 ± 0.05	0.7 ± 0.05	1.0

Table 6: Results for xResNet101 including means, standard deviation and p-values (p-val) for ROC AUC, partial ROC AUC (pROC AUC), and precision-recall AUC (PR AUC)

	ROC AUC			pROC AUC			PR AUC		
	F	M	p-val	F	M	p-val	F	M	p-val
Atrial Fibrillation									
F_0	0.97 ± 0.0	0.97 ± 0.01	0.4473	0.94 ± 0.01	0.94 ± 0.01	0.5639	0.78 ± 0.04	0.77 ± 0.04	0.5791
F_{25}	0.97 ± 0.0	0.97 ± 0.01	0.1394	0.94 ± 0.01	0.94 ± 0.01	0.2448	0.79 ± 0.03	0.78 ± 0.05	0.3597
F_{50}	0.97 ± 0.0	0.97 ± 0.01	0.1411	0.94 ± 0.01	0.93 ± 0.01	0.0995	0.79 ± 0.03	0.76 ± 0.04	0.0957
F_{75}	0.98 ± 0.0	0.97 ± 0.01	0.08	0.95 ± 0.0	0.94 ± 0.01	0.0294	0.81 ± 0.03	0.78 ± 0.04	0.0892
F_{100}	0.98 ± 0.0	0.97 ± 0.01	0.0393	0.95 ± 0.01	0.93 ± 0.02	0.0382	0.8 ± 0.04	0.75 ± 0.07	0.0334
Sinus Rhythm									
F_0	0.94 ± 0.01	0.93 ± 0.01	0.5196	0.86 ± 0.02	0.85 ± 0.03	0.7524	0.99 ± 0.0	0.99 ± 0.0	0.6817
F_{25}	0.94 ± 0.01	0.93 ± 0.01	0.2096	0.86 ± 0.02	0.85 ± 0.01	0.4803	0.99 ± 0.0	0.99 ± 0.0	0.3818
F_{50}	0.94 ± 0.01	0.93 ± 0.01	0.105	0.85 ± 0.02	0.84 ± 0.01	0.2267	0.99 ± 0.0	0.99 ± 0.0	0.1603
F_{75}	0.94 ± 0.01	0.93 ± 0.01	0.0667	0.87 ± 0.01	0.85 ± 0.03	0.0925	0.99 ± 0.0	0.99 ± 0.0	0.0531
F_{100}	0.94 ± 0.01	0.92 ± 0.02	0.0267	0.86 ± 0.02	0.82 ± 0.03	0.037	0.99 ± 0.0	0.99 ± 0.0	0.0167
Myocardial Infarction									
F_0	0.94 ± 0.01	0.95 ± 0.0	0.1189	0.86 ± 0.02	0.89 ± 0.01	0.0727	0.64 ± 0.04	0.72 ± 0.04	0.0167
F_{25}	0.94 ± 0.01	0.95 ± 0.0	0.1498	0.87 ± 0.02	0.89 ± 0.01	0.0686	0.66 ± 0.03	0.72 ± 0.04	0.0366
F_{50}	0.94 ± 0.01	0.95 ± 0.0	0.2281	0.87 ± 0.02	0.89 ± 0.01	0.1	0.67 ± 0.04	0.73 ± 0.03	0.0224
F_{75}	0.94 ± 0.01	0.95 ± 0.0	0.2458	0.88 ± 0.02	0.89 ± 0.01	0.2292	0.68 ± 0.02	0.71 ± 0.03	0.1563
F_{100}	0.95 ± 0.01	0.95 ± 0.0	0.7946	0.88 ± 0.02	0.88 ± 0.01	0.7806	0.68 ± 0.03	0.69 ± 0.04	0.362

Table 7: Results for CNN including means, standard deviation and p-values (p-val) for ROC AUC, partial ROC AUC (pROC AUC), and precision-recall AUC (PR AUC)

	ROC AUC			pROC AUC			PR AUC		
	F	M	p-val	F	M	p-val	F	M	p-val
				Atrial Fibrillation					
F_0	0.97 ± 0.0	0.96 ± 0.0	0.0997	0.92 ± 0.01	0.92 ± 0.01	0.3799	0.74 ± 0.05	0.73 ± 0.03	0.5072
F_{25}	0.97 ± 0.0	0.97 ± 0.01	0.0949	0.93 ± 0.01	0.93 ± 0.01	0.0275	0.77 ± 0.04	0.74 ± 0.04	0.1137
F_{50}	0.97 ± 0.0	0.96 ± 0.01	0.0754	0.93 ± 0.01	0.92 ± 0.01	0.0247	0.76 ± 0.03	0.74 ± 0.04	0.1198
F_{75}	0.97 ± 0.0	0.96 ± 0.01	0.0292	0.94 ± 0.01	0.92 ± 0.01	0.0229	0.77 ± 0.03	0.73 ± 0.04	0.1689
F_{100}	0.97 ± 0.0	0.96 ± 0.01	0.0225	0.93 ± 0.01	0.91 ± 0.02	0.0053	0.75 ± 0.03	0.71 ± 0.05	0.0589
				Sinus Rhythm					
F_0	0.93 ± 0.01	0.93 ± 0.01	0.3994	0.84 ± 0.01	0.84 ± 0.02	0.6468	0.99 ± 0.0	0.99 ± 0.0	0.4846
F_{25}	0.94 ± 0.01	0.92 ± 0.01	0.0709	0.85 ± 0.02	0.83 ± 0.02	0.1703	0.99 ± 0.0	0.99 ± 0.0	0.08
F_{50}	0.93 ± 0.01	0.92 ± 0.01	0.122	0.85 ± 0.02	0.83 ± 0.02	0.2034	0.99 ± 0.0	0.99 ± 0.0	0.1078
F_{75}	0.93 ± 0.01	0.92 ± 0.01	0.0681	0.85 ± 0.02	0.83 ± 0.02	0.1382	0.99 ± 0.0	0.99 ± 0.0	0.1053
F_{100}	0.93 ± 0.01	0.91 ± 0.01	0.0539	0.84 ± 0.02	0.82 ± 0.02	0.1037	0.99 ± 0.0	0.98 ± 0.0	0.0683
				Myocardial Infarction					
F_0	0.93 ± 0.01	0.95 ± 0.0	0.0913	0.85 ± 0.03	0.88 ± 0.01	0.0657	0.63 ± 0.03	0.7 ± 0.03	0.0209
F_{25}	0.94 ± 0.02	0.95 ± 0.01	0.165	0.86 ± 0.04	0.88 ± 0.01	0.1295	0.63 ± 0.04	0.7 ± 0.03	0.0352
F_{50}	0.94 ± 0.01	0.95 ± 0.01	0.3388	0.87 ± 0.03	0.88 ± 0.02	0.2295	0.64 ± 0.03	0.7 ± 0.04	0.0219
F_{75}	0.94 ± 0.01	0.95 ± 0.01	0.7422	0.87 ± 0.02	0.88 ± 0.01	0.3585	0.66 ± 0.03	0.68 ± 0.04	0.3109
F_{100}	0.94 ± 0.01	0.94 ± 0.01	0.7138	0.87 ± 0.02	0.87 ± 0.02	0.9954	0.65 ± 0.03	0.65 ± 0.05	0.9582



(a) ResNet with Attention

(b) XResNet

(c) CNN

Figure 6: Three-dimensional UMAP projection of latent representations extracted from the penultimate layer of the trained models: (a) ResNet with Attention, (b) XResNet, (c) CNN. Each point corresponds to a test sample, with colour indicating sex (blue: male, orange: female).