NPL
National Physical Laboratory

NPL Report Number MS 65

Investigation of the Impact of Demographic Imbalance in the Training Data on Fairness in Machine Learning

Philip J. Aston, Padmini Krishnadas, Tameem Adel, David Whittaker, Paul Duncan

March 2026

National Physical Laboratory

National Physical Laboratory (NPL)

Investigation of the Impact of Demographic Imbalance in the Training Data on Fairness in Machine Learning

Philip J. Aston, Padmini Krishnadas, Tameem Adel, David Whittaker, Paul Duncan

We are the UK's National Metrology Institute (NMI),
a world-leading centre of excellence that provides cutting-edge measurement science, engineering and technology that underpins prosperity and quality of life in the UK.

MCR-001, Issue 1, July 2025

ISSN: 1754-2960

https://doi.org/10.47120/npl.MS65

National Physical Laboratory

Hampton Road, Teddington, Middlesex, TW11 0LW

This work was funded by the UK Government's Department for Science, Innovation & Technology through the UK's National Measurement System programmes.

Contents

Introduction

Artificial intelligence (AI) technologies have rapidly expanded across healthcare, driven by advances in sensing, data availability and machine learning (ML) methods, alongside a maturing regulatory ecosystem that increasingly emphasises trustworthiness, transparency and risk management. Frameworks such as the NIST AI Risk Management Framework (RMF) [1] set out multidimensional characteristics for trustworthy AI, including safety, reliability, explainability, privacy and fairness, and are now widely referenced across health and public-sector domains. As AI is increasingly embedded in clinical workflows these characteristics are now central to its evaluation across the total product life cycle.

The AI Standards Hub (AISH) is an initiative established through the UK's National AI strategy involving three elements of the UK's National Quality Infrastructure which aims to bring together industry, government, regulators, consumers and civil society, and academia to promote sound, coherent and effective standards to inform and strengthen AI governance practices domestically and internationally.

Through the pre-normative research program of AISH, the National Physical Laboratory (NPL) has developed a Trustworthy and Safe AI Life Cycle (TSALC) [2] (see Figure 1) which provides a structured, risk-oriented framework that links AI risks directly to software engineering practices and measurable trustworthiness metrics across the full product life cycle. The TSALC stresses the importance of risk analysis, design and development controls, verification, validation and post-deployment monitoring, drawing on international guidance such as the NIST AI RMF [1] and principles of Good Machine Learning Practice (GMLP) [3]. This framework provides a practical bridge between high-level regulatory expectations and the concrete technical and organisational processes needed to develop AI systems that are safe, reliable and aligned with organisation's needs.

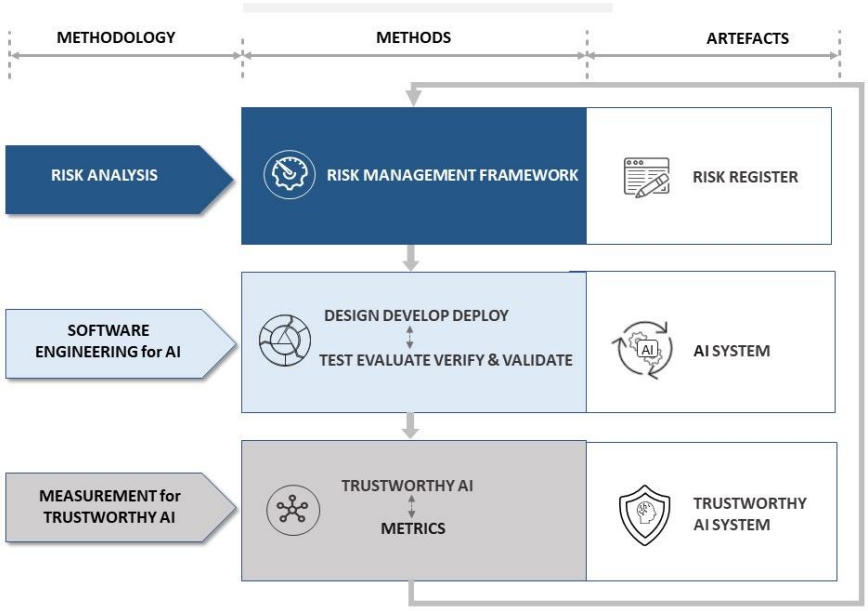

Figure 1 The Trustworthy and Safe AI Life Cycle [2].

NPL have developed a case study [4] to demonstrate how the TSALC can be applied to atrial fibrillation (AF) detection with wearable photoplethysmography (PPG) signal sensors, linking risk analysis to software engineering practice and measurable trustworthiness metrics in a realistic, clinician-initiated screening scenario. This case study established a structured approach for risk identification and mitigation, informed by expert input from academia, industry and the NHS, and highlighted the need for quantitative measures of trustworthiness in safety-critical settings.

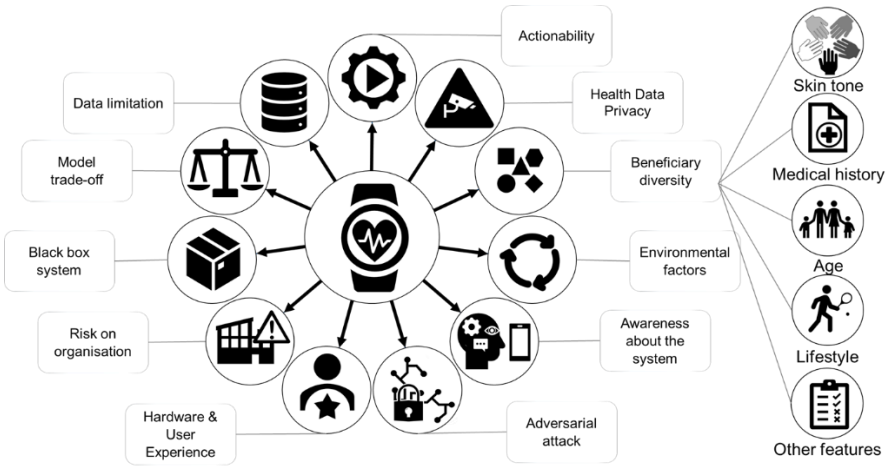

Figure 2 A schematic of the high-level risk categories for wearables AF detection [4].

These risks are shown in Figure 2 where a number of high-level risks surrounding AF detection with wearables are shown. Following the structure of the TSALC in Figure 1, these risks are then used to determine a set of trustworthiness principles the AI systems

must satisfy where “beneficiary diversity” or fairness is called out as a key element. This means that the AI system should produce predictions and decisions that do not adversely affect users based on characteristics such as gender, age, ethnicity and body mass index (BMI).

Concurrently, international frameworks have converged on fairness as a defining attribute of trustworthy AI. The NIST AI Risk Management Framework [1] specifies “Fair, with harmful bias managed” as one of seven trustworthiness characteristics and provides process guidance for governance, measurement and ongoing risk management. Regulatory guidance for medical AI has also matured: the US Food and Drug Administration (FDA), the UK Medicines and Healthcare products Regulatory Agency (MHRA) and Health Canada introduced Good Machine Learning Practice principles [3] in 2021, which were subsequently harmonised by the International Medical Device Regulators Forum (IMDRF) as a global GMLP reference in 2025 [5], reinforcing expectations for representative datasets, human-AI team performance and post-market monitoring. Within the UK, the MHRA’s Software and AI as a Medical Device Change Programme [6] and associated guidance emphasise life-cycle oversight, transparency and real-world performance for software and AI devices.

In this report we expand on the initial TSALC paper and subsequent case study by focusing on the beneficiary diversity requirements shown in Figure 2. We consider the simple binary classification problem of the detection or absence of AF in a short PPG signal. Our interest though, is to consider the effect on this prediction problem of using either balanced or biased datasets with respect to a number of factors relating to personal characteristics of the subjects. In particular, we consider gender, age, ethnicity and body mass index (BMI) individually.

Datasets

For this study, we created various datasets from the large *MIMIC-III-Ext-PPG* dataset [7] which contains over 4.9 million segments of 30 s PPG signals from 6,131 subjects derived from the matched waveform subset of *MIMIC-III* [8]. The annotations included in the dataset include AF and sinus rhythm (SR) and so we constructed datasets using these

labels. This data was recorded from critical care patients and so models trained on this data may not be suitable for detecting AF in healthy subjects.

The *MIMIC-III-Ext-PPG* dataset includes 12 folds which are stratified by heart rhythm events, age distributions, and cardiovascular ICD-10 discharge diagnoses. Each fold also contains unique patient identifiers. We will use these folds to create our own equal sized folds for each dataset, ensuring that each fold similarly has unique patient identifiers and contains a balanced set of labels. A signal quality index was used to classify all the signals in the dataset as either high or low quality or flagged as having issues. Our datasets were constructed using only high quality signals. For each category, we generated training datasets split into a number of folds and an independent test dataset which can be used to test all the different models so that the effect of varying the balance of the various personal characteristics can be assessed on a single independent dataset.

Gender

We created 5 training datasets, all consisting of 20,000 records which were equally split between the SR and AF labels and divided into 10 folds with each fold containing 1,000 SR and 1,000 AF records. These were derived from the first 10 stratified folds of the dataset. The 5 datasets had a different gender balance across both the SR and AF records as follows:

- 100 % female
- 75 % female, 25 % male
- 50 % female, 50 % male
- 25 % female, 75 % male
- 100 % male

In addition, we created a balanced, independent test set from the last two stratified folds with a total of 8,000 records, made up of 2,000 records in each of the categories female SR, female AF, male SR, male AF.

Age

The *MIMIC-III-Ext-PPG* dataset includes the age of all the subjects. We generated datasets based on age in the three intervals 46 - 60, 61 - 75, 76 - 90 years. We wanted to create balanced datasets with equal numbers of SR and AF records using the 12 stratified folds provided with the dataset. However, there were not 1,000 AF records in the lowest age range for every fold. Thus, we created 11 folds each with 1,000 SR and 1,000 AF records by combining stratified folds 4, 5 and 7, 8 and splitting stratified fold 2 into two separate folds by subject. In addition, we moved the 207 AF records (and the corresponding SR records) for subject 81475 from stratified fold 8 to fold 3. This gave 11 folds each with unique patient identifiers. From each fold, we randomly extracted 1,000 high quality records in each of the three age ranges and for both SR and AF. These extracted records were used to generate the following training datasets which each consisted of 20,000 records equally split between SR and AF and selected from the first 10 folds:

- *dataset_*all: Equal number of records in each age group.
- *dataset_young*: All of the 46 - 60 records.
- *dataset_mid*: All of the 61 - 75 records.
- *dataset_old*: All of the 76 - 90 records.

Clearly, the first dataset is balanced with respect to the three age classes, while the other datasets consist of only one of the age classes. We also created a balanced *test_dataset* consisting of the 6,000 records from fold 11 which has 1,000 records for each age class and for both categories of SR and AF.

Ethnicity

The *MIMIC-III-Ext-PPG* dataset has a total of 33 different classes of ethnicity recorded. However, we restrict attention to the four major classes of White, Black, Hispanic or Latino (which we will refer to as simply Hispanic) and Asian, and we ignore the finer subdivision of these classes. A summary of the number of available records in these categories is shown in Table 1. Not surprisingly, there were many fewer AF records than SR, and the number of AF records varied significantly by ethnic group.

MCR-001, Issue 1, July 2025

Table 1 The number of SR and AF records with high quality signals for the four ethnic groups.

	SR	AF
White	1,911,764	186,907
Black	282.815	11,745
Hispanic	99,885	2,608
Asian	107,879	7,154
Total:	2,402,343	208,414

Because of the low number of Hispanic AF records, we generated 5 folds of data with each fold containing 500 SR and 500 AF records and with a unique set of subjects. Thus, the total number of records was 5,000 records for each ethnic group. We used a different approach for different ethnic groups, which we now describe.

White, Black

For the White and Black subjects, there were plenty of records available in both rhythm classes and so we simply combined the stratified folds together into 5 groups and then randomly selected 500 SR records and 500 AF records from each of these 5 groups to make up our 5 folds.

Hispanic

For the Hispanic subjects, there were only 2,608 AF records available (see Table 1), and we required 2,500 of these for our 5 folds. Thus, there were very few spare records. In this case, we found that it was not possible to combine the 12 stratified folds into 5 groups such that each group had at least 500 Hispanic AF records and so we combined subjects into 5 groups instead. We note that two groups contain records from a single subject which is not ideal. We then randomly selected 500 records from each group to give the AF records of our 5 folds.

As there are many Hispanic SR records available, we combined the stratified folds into 5 groups in the same way as for White and Black subjects. In order to have unique subjects

in each fold, we excluded records from each group associated with subjects with AF records that were not in that group and then randomly selected 500 records from the remaining records to give our 5 folds.

Asian

There are almost 3 times as many Asian AF records as we need (see Table 1), but these records were not evenly distributed across the 12 stratified folds. The combination of folds used for White and Black subjects did not give at least 500 AF records in each fold and so we combined the folds in a different way such that there were over 500 AF records in each of the 5 groups. We then randomly selected 500 SR and 500 AF records from each of these groups to make up our 5 folds.

Datasets

The extracted records described above were used to create the following training datasets from the first 4 folds of data, and each with 4,000 records equally split between SR and AF:

- *dataset_all*: 4,000 records in total consisting of 1,000 each for White, Black, Hispanic and Asian.
- *dataset_WHITE*: All 4,000 White records.
- *dataset_BLACK*: All 4,000 Black records.
- *dataset_HISPA*: All 4,000 Hispanic records.
- *dataset_ASIAN*: All 4,000 Asian records.

Again, the first dataset is balanced with respect to the four ethnic groups while the others consist of data for just one ethnic group. We also created a balanced *test_dataset* consisting of the 4,000 records from the fifth fold which has 500 records for each ethnic group and for both categories of SR and AF.

Body Mass Index

We considered BMI as another category which may influence ML results using PPG signals. BMI is calculated as weight (kg) / height $(m)^2$ and so has units of kg m^{-2}. We note that BMI is a simplistic but commonly used measure of obesity. A report by the WHO [9] notes that “Body mass index (BMI) provides the most useful, albeit crude, population-level

measure of obesity. It can be used to estimate the prevalence of obesity within a population and the risks associated with it. However, BMI does not account for the wide variation in body fat distribution, and may not correspond to the same degree of fatness or associated health risk in different individuals and populations."

The NHS categorises adults based on their BMI [10] as follows:

- < 18.5: Underweight
- 18.5 - 24.9: Healthy
- 25 - 29.9: Overweight
- 30 - 39.9: Obese
- ≥ 40: Severely obese

These ranges are slightly different for people with an Asian, Chinese, Middle Eastern, Black African or African-Caribbean family background but we do not take account of this.

The *MIMIC-III-Ext-PPG* dataset includes the weight and height for most subjects which enables the BMI to be calculated. A few of the weights and heights appear to have been entered incorrectly and so we restrict interest to subjects with height in the range 1.3 - 2.3 m and weight in the range 30 - 150 kg.

There are insufficient AF records for the Underweight category, and so we only consider the other four categories. We created 6 groups based on the stratified folds contained in the dataset. In addition, we transferred all the records for all subjects other than 63116 which are in one stratified fold to another that was short of AF records. From each of these groups, we randomly extracted 1,000 SR and 1,000 AF records in each of the four classes to give our 6 folds. These records were then used to generate the following training datasets which each consisted of 10,000 records equally split between SR and AF and selected from the first 5 folds:

- *dataset_all*: An equal number of records in each BMI class.
- *dataset_Healthy*: All of the Healthy records.

- *dataset_Overweight*: All of the Overweight records.
- *dataset_Obese*: All of the Obese records.
- *dataset_SeverelyObese*: All of the SeverelyObese records.

As for the previous cases, the first dataset is balanced with respect to the four BMI classes while the other datasets consist of data for a single BMI class. We also created a balanced *test_dataset* consisting of the 8,000 records from the sixth fold which contains 1,000 records for each BMI class and for both categories of SR and AF.

Machine learning methods

We used three different ML methods using the above datasets, namely ML using spectral features, ML using engineered features and deep learning using the raw PPG signals. We now describe each of these approaches in more detail.

Machine learning using spectral features

This approach makes use of a feedforward neural network [11] (FNN) trained and tested on spectral magnitudes (or amplitudes) of raw PPG signals following a Fourier transform (FT) [12]. Because the overarching aim of this work is not to tune algorithm parameters for best classification performance or compare that of different algorithms, no attempt to optimise it with respect to the number of hidden layers and neurons in each and the number of training epochs using validation sets is made. Instead, and following the work of Jimanez et al [13], a single hidden layer comprising 125 neurons (so shallow—not deep—learning according to the general definition of the terms) is used and the algorithm is trained for 200 epochs. This layer uses a rectified linear unit (ReLU) as the activation function. Although the work described in [13] analysed ECG rather than PPG signals, both follow cardiac rhythm (the former electrically and the latter optically) and so they should share structural similarities in the frequency domain.

The Adam optimiser [14] is used by the FNN and the Sparse Categorical Cross Entropy loss function is chosen for binary classification. The optimiser learning rate is set to 0.001,

MCR-001, Issue 1, July 2025

and dropout [15] with a dropout rate of 0.2 is used throughout training (this is not used during inference, so use is not made of what is termed Monte Carlo dropout).

Spectral magnitudes are equal to the square root of the sum of the squares of the real and imaginary output of the FT. A sampling rate, F_S, of 125 Hz was used to log the PPG time domain signals, and the code selects a constant, contiguous range of frequencies discarding the DC (0 Hz) component and those close to the Niquist limit of $F_S/2$.

Preprocessing of these frequency-resolved datasets is minimal: magnitudes are smoothed by averaging over batches of ten consecutive frequencies and all are normalised by the total area under the spectrum concerned. (It would be an interesting exercise, though one not directly within the scope of this work, to compare results with and without such normalisation, to determine the extent to which the potential predictive power of input data is affected by absolute power spectral density.) Dimensional reduction, for instance using principal component analysis [16], is not attempted, but the number of training samples is much (over an order of magnitude) greater than the number of resulting dimensions in the form of magnitudes fed as input variables or features to the FNN.

Machine learning using engineered features

Feature engineering is the process of selecting features containing crucial attributes from a dataset relevant to the task at hand. This helps focus the ML models attention on important characteristics of the data while eliminating noise to improve the models predictive performance [17] [18] [19]. The proposed features extracted from the PPG signals provide interpretable representations of cardiac activity by capturing aspects of cardiac timing, morphology, spectral characteristics and more that are directly influenced by underlying rhythm differences. When used with ML models that can provide rankings of feature importances, they contribute transparency and reliability to the model's predictions.

The raw PPG data in the form of .dat and .hea files were loaded and pre-processed using the WFDB package [20]. A zero phase Butterworth bandpass filer of order 3 with default passbands of 0.5 - 0.8 Hz was applied to suppress baseline drift and high frequency noise. Signals were then partitioned into overlapping windows of 60 s each with 50 % overlap. A final trailing window is added if any remainder of the signal exceeded 25 % of the window length. Windows shorter than 0.5 s were excluded. The SciPy package's 'find_peaks' function was used to find systolic peaks in the PPG signal for each window, with a

minimum inter peak distance of 0.3 s and an adaptive prominence threshold of 2 % of the window's standard deviation. From this, interbeat intervals (IBI) were computed as the time differences between consecutive peaks. Beat centred segments were extracted as well to capture morphological irregularities using a fixed window around each detected peak. From the IBI information, we compute the follow time domain based features of variability:

1. RMSSD: Root mean square of successive IBI differences
2. SDNN: Standard deviation of IBIs
3. CV of IBI: Computed as the ratio of the standard deviation of the IBIs to the mean IBI.
4. Sample entropy of the IBI sequence provided there were a minimum of 10 IBIs present for a window

The second set of features captures beat morphology stability and are detailed below:

5. Amplitude variability: Variability of the beat prominence from local peak prominence. This can be expressed as the standard deviation of amplitude divided by the median amplitude
6. Width variability: Variability of beat widths measured at half prominence, expressed as the standard deviation of width divided by the median width
7. Consecutive bean correlation: Defined as the median Pearson correlation of z-scored, resampled bean waveforms (each beat resampled to a fixed length, given there are more than three valid beats)
8. Peak timing instability: Standard deviation of residuals after a linear fit to cumulative peak times
9. Dicrotic notch stability: For each detected systolic peak, the algorithm searched 60 - 300 ms afterward for a local minimum representing a candidate for a dicrotic notch. The per beat delay in seconds from the systolic peak to this minimum is computed. The standard deviation of these delays represents the window level dicrotic notch stability. Larger values of this feature indicate unstable vascular movement

The third set of features captures spectral characteristics of the signal by computing Welch's periodogram for each window to derive the following:

10. Spectral entropy: Shannon entropy of the normalised spectrum

11. Spectral flatness (Weiner entropy): This feature quantifies the degree to which the PPG power spectrum resembles a broadband noise like distribution. Sinus rhythm produces sharp heart rate peaks yielding low flatness values while irregular rhythms distribute energy across many frequencies resulting in higher flatness. Flatness is computed as the ratio of the geometric mean to the arithmetic mean of the spectrum.
12. Dominant peak power reduction: This feature measures show strongly the spectrum is dominated by its largest frequency component. Lower values indicate a clear dominant heart rate frequency while higher values reflect reduced spectral concentration and increased spread is characteristic of irregular rhythms such as atrial fibrillation.
13. Broadband power fraction: Fraction of power outside a ± 25 Hz band around the dominant heart rate peak within the range 0.7 - 3.0 Hz.

In addition to these metrics, the count of detected beats per window, window duration, and an estimate of beats per minute is also reported.

These features were inputted to a Random Forest model implemented with 10-fold cross validation for different versions of the personal characteristic datasets.

The Expected Calibration Error (ECE) is a metric of confidence calibration which measures how well a model's estimated probabilities match observed probabilities by taking a weighted average over the absolute difference between average accuracy and the average confidence [21] [22]. The Uncertainty Calibration Error (UCE) metric extends the calibration assessment by measuring the alignment of a model's predicted uncertainty and true predictive performance. UCE was computed for every trained model following the general formulation of UCE for multiclass classification [22] but approximated for binary classification, as the weighted difference between mean uncertainty and empirical error across uncertainty binned groups.

Deep learning using the raw PPG signals

The method implemented herein is a one-dimensional convolutional neural network (1D-CNN). A 1D-CNN [23] is a specialised type of neural network designed to process sequential data, unlike its more famous 2D counterpart which is primarily used for image processing. In a 1D-CNN, the convolutional kernels (filters) move across the data in one dimension,

making it particularly effective for tasks involving time series, signals (like ECG or raw PPG signals), or text where the order and local patterns within a sequence are critical. This architecture excels at automatically learning hierarchical features from raw input, reducing the need for extensive manual feature engineering.

The core component of a 1D-CNN is the Conv1D layer. This layer applies a set of learnable filters, each acting as a pattern detector, across the input sequence. Each filter is a small window of weights that slides over the input, performing a dot product with the local data segment that it covers. This operation, known as convolution, generates a feature map that highlights the presence of specific patterns in different parts of the sequence. The number of filters mainly determines the depth of the output feature map, and the kernel size defines the width of the sliding window. Activation functions, such as ReLUs, are typically applied after the convolution to introduce nonlinearity.

Following the convolutional layers, pooling layers like MaxPooling1D are commonly used. These layers downsample the feature maps, reducing their dimensionality and computational complexity while retaining the most salient features. MaxPooling1D achieves this by taking the maximum value within a sliding window, effectively creating a more abstract and robust representation that is less sensitive to minor shifts in the input. After several convolutional and pooling layers, the high-dimensional feature maps are typically flattened into a single vector. This flattened vector then feeds into one or more dense (fully connected) layers, which perform the final classification or regression task.

In essence, a 1D-CNN learns to identify relevant local patterns within a sequence and then combines these patterns to make higher-level predictions [24]. Its ability to automatically extract meaningful features and its robustness to temporal variations make it a powerful tool for analysing various forms of sequential data. For binary classification tasks, like distinguishing between different heart rhythms, the final dense layer usually has a single output unit with a sigmoid activation function, outputting a probability score between 0 and 1.

The primary advantages of 1D-CNNs lie in their ability to automatically learn local features and patterns directly from raw sequential data, reducing the need for extensive domain-specific feature engineering. The architecture of 1D-CNNs allows for translational invariance [25], meaning that a learnt pattern can be detected regardless of its position in the sequence.

This makes 1D-CNNs more robust to variations in timing or phase, which is particularly beneficial in fields like biomedical signal processing (e.g., detecting anomalies in ECG signals), natural language processing (e.g., sentiment analysis on text sequences), and industrial monitoring (e.g., fault detection in sensor data). Furthermore, by stacking multiple convolutional and pooling layers, 1D-CNNs can build hierarchical representations, capturing both fine-grained local details and more abstract, global patterns within the sequence, leading to powerful and efficient models for various sequential data analysis tasks.

Machine learning experiments

For each of the ML methods described and datasets generated, we performed the following experiments:

- Cross validation on each training dataset using the available folds. This allows a comparison of performance in detecting AF using either a balanced dataset or a single group dataset.
- Each of the models generated in the cross validation was applied to the test dataset in order to compare the effects of training data with a different balance of the particular characteristic of interest on a common dataset.

Results

Effect of gender imbalances

When evaluating the performance of the three methods after varying the split of genders in the dataset, we observe in Figure 3 when validating on male only and female only sets, that models trained on raw PPG signals show strong sensitivity to gender imbalance in the training data, achieving highest accuracy when the gender proportions in the training set match the validation set. Models trained on interpretable features maintain stable accuracy regardless of the gender composition of the validation set, indicating robustness and poor sensitivity to gender distribution. Models trained on spectral features show moderate dependence on gender distribution with reduced accuracy when trained exclusively on the

opposite gender but less severe in variation as compared to models trained on raw signals.

When considering the mixed gender validation set, the model trained on raw PPG signals once again favours conditions where the proportion of female records is the highest, while accuracy gradually decreases as the number of female samples in the training set reduce. The model trained on interpretable features performs best when the validation set is gender balanced but shows no meaningful accuracy changes when the proportion of male or female samples is increased. The spectral features model shows the opposite pattern, reporting lowest accuracy on the gender balanced validation set but improved accuracy when either male or female samples dominate.

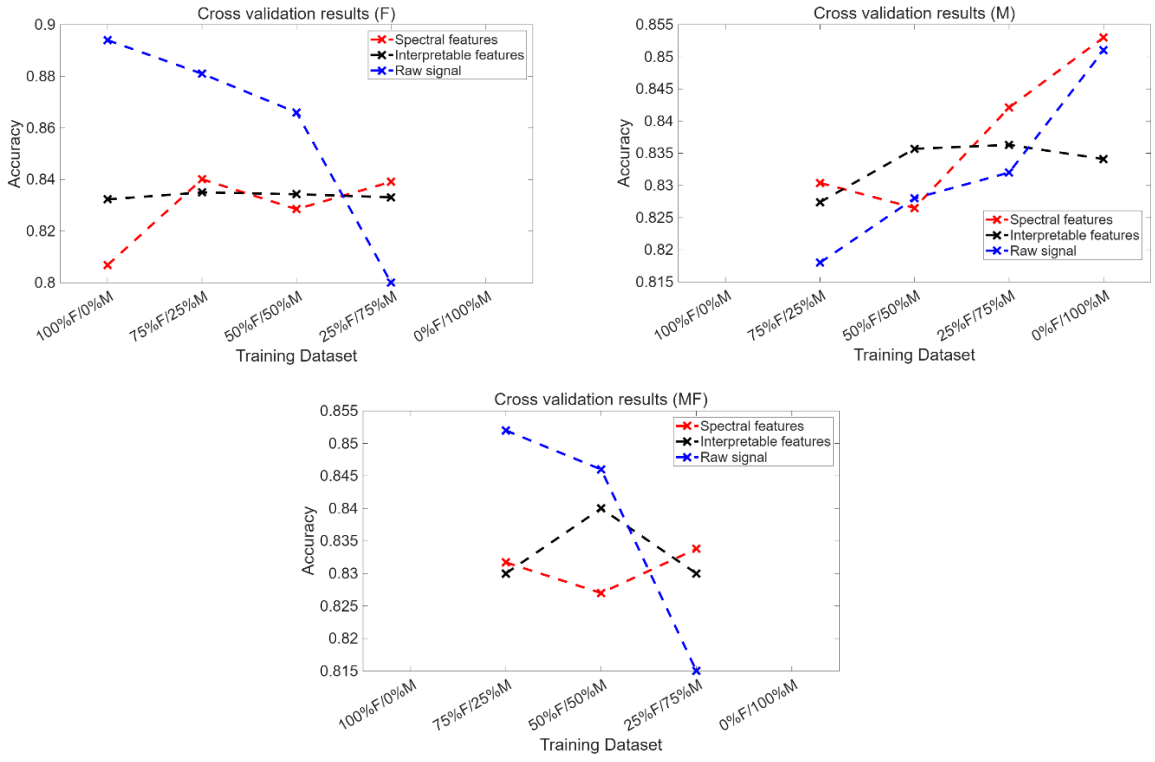

Figure 3 Plots depicting performance metrics to predict AF in (left) only female (right) only males (bottom) males and females, when using the three defined methods of ML classification models.

Each of the cross validation models for all three methods were evaluated on the hold out test set and results considered for the groups female only, male only and mixed gender, with the mean and standard deviation of the accuracies being reported (see Figure 4). On the gender-specific test sets, the models trained on raw PPG signals once again showed strong sensitivity to training set gender composition, achieving highest accuracy when the training and test sets matched in gender distribution, showing preference for female

dominated training sets. It also displayed the largest variability when trained on a balanced dataset when tested on female only data. The interpretable feature model remained stable across all training splits, reporting only a slight drop in performance when trained exclusively on male data. The spectral features model showed low variability as well with the highest accuracy when trained on male only data and tested on female only records indicating mild gender sensitivity but substantially less when compared to the raw signal model.

On the mixed-gender test set, the raw signal model again performed best when trained with the highest proportion of female data, though with considerable standard deviation, and its performance declined as female representation in the training set decreased. The interpretable features model continued to show minimal dependence on gender composition and achieved its best (though only marginally higher) accuracy when trained on a balanced dataset. The spectral features model recorded its lowest accuracy when trained on female-only data, but showed similar accuracy across all other splits, further indicating only mild sensitivity to training gender composition.

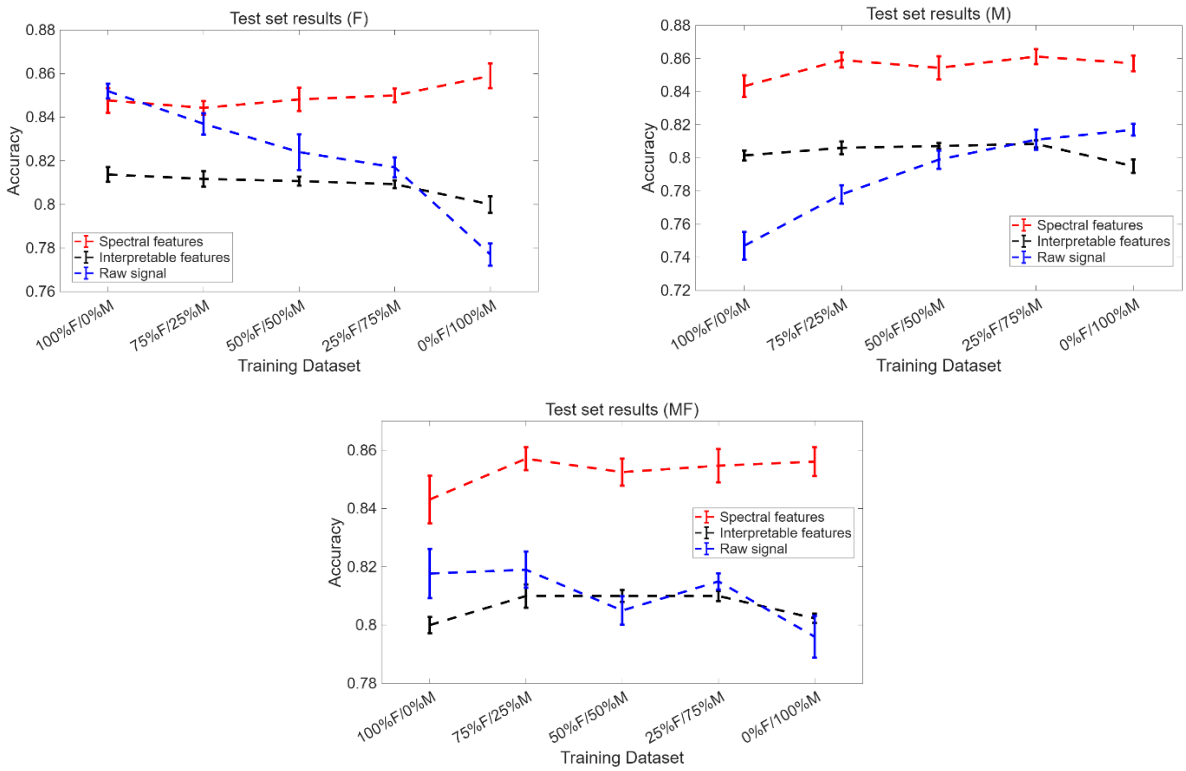

Figure 4 Mean and standard deviation of all the cross validation models applied to each of the test datasets. The plots show the results restricted to female, male and both records.

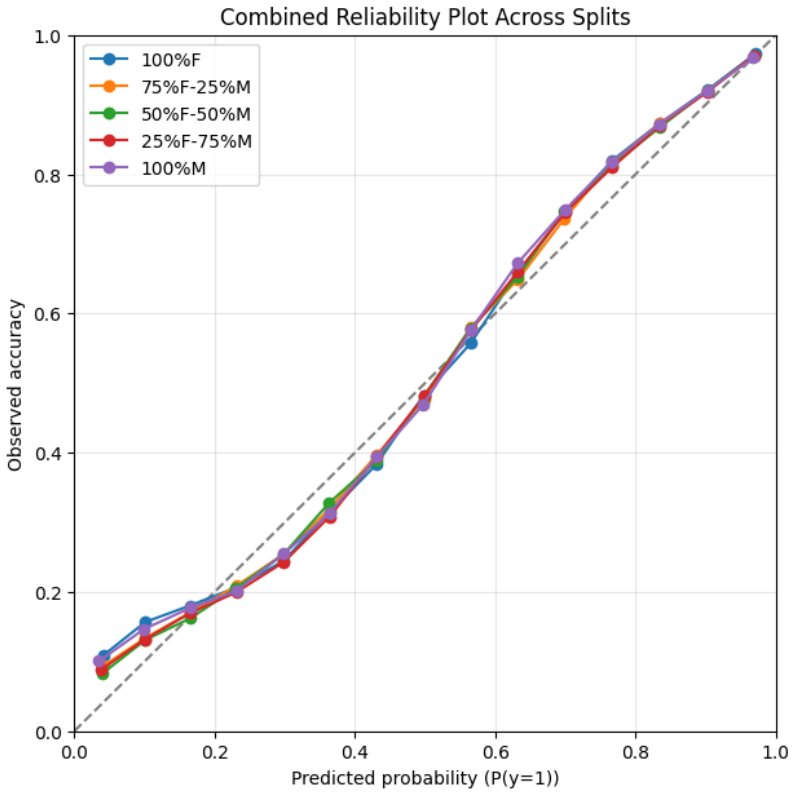

Figure 5 Reliability plot depicting the calibration of models trained on interpretable features when classifying AF.

Table 2 Calibration metrics for models trained on interpretable features. The lowest errors are shown in bold.

Training data version	UCE	ECE
All female	0.0247	0.0326
75 % female 25 % male	0.0199	0.0248
50 % female 50 % male	**0.0198**	**0.0236**
25 % female 25 % male	0.0221	0.0271
All male	0.0280	0.0340

As a metric of interpretability, the reliability plot together with UCE and ECE metrics are used to assess the model calibration and confidence. These were used in conjunction with the models trained on interpretable features to build interpretability and reliability to model predictions. Figure 5 represents the reliability plot for models trained on interpretable features. The plot depicts how the models are nearly alike irrespective of shifts in the

proportion of male or female records during training, corroborating the model's robustness to gender composition in the training set. Table 2 represents the calibration metrics for the models trained on interpretable features. Models report low values of UCE and ECE with minimal variability between different training sets.

Effect of age imbalances

When investigating the effect of varying the composition of training data in terms of age groups present, model performances were compared when trained on datasets restricted to specific age groups (young, mid and old) as well as versus all ages combined, evaluated using cross validation.

When considering cross validation results (depicted in Figure 6 by dotted lines), models trained on raw signals and interpretable features using all age groups report stable accuracies when validated across different age groups. When considering models trained on the spectral features, models validated on the Mid age group reported the lowest accuracy.

When the models were trained on different age groups, trends and patterns in performance metrics emerge. When considering models trained on raw signals, models trained on the Mid age group report a sharp decrease in accuracy as compared to when trained on other groups. Highest accuracy was recorded when trained on the dataset with all age groups. Considering models trained on interpretable features, there is a sharp and distinct decrease in accuracy when training on the Young dataset. When comparing models trained on spectral features, there is a sharp increase in accuracy when training on the Mid age group as compared to other age groups.

Each of the cross validations models was then tested on a hold out test set, recording the mean and standard deviation of the accuracies.

Models trained on raw PPG signals report large standard deviations in accuracy when tested on a dataset with all categories of age groups except for when the training set matched the test set and consisted of similar composition of age data. When the models were tested on the Young test set, the models once again report large deviations in accuracies irrespective of the training set composition. The model recorded the lowest

accuracy when trained on the Mid age group. These results are consistent even when the models were tested on the Mid age group and Old age groups.

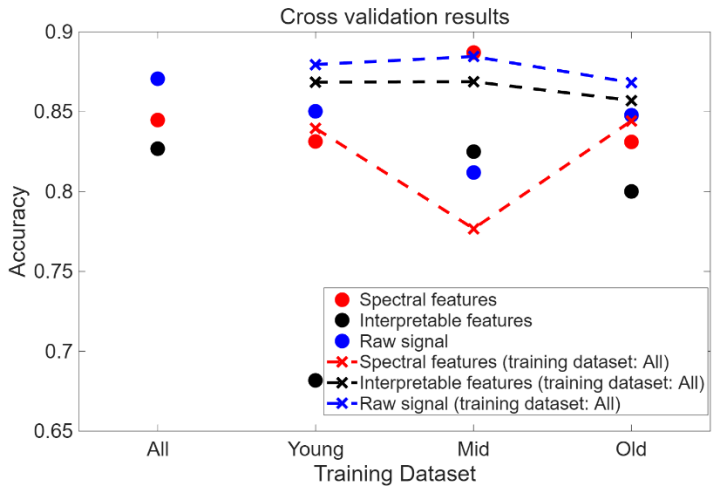

Figure 6 Cross validation results for the different training datasets. The crosses linked by dashed lines show the breakdown when trained on 'All' for the different age groups.

Models trained on interpretable features report the lowest accuracies as compared to other methods when varying age compositions in the hold out test set. When comparing model performances when tested on the dataset with all age groups, the All and Young models report the lowest accuracies (~50 %), a sharp decrease when compared to other methods. The accuracy picks up for the Mid and Old groups. This trend is also followed when testing on the Young and Mid dataset, however the accuracies reported by the Mid and Old models are higher in this case, with lower standard deviations. When testing on the Old dataset, the model trained on All, Mid and Old data report near ~70 % accuracy but very low accuracy for models trained on Young data. Additionally, models trained on the Old and All datasets report large standard deviations. These results suggest that models trained on interpretable features are highly sensitive to the age composition of the training and test sets and its performance and reliability is dependent on the balance of the dataset.

Models trained on spectral features of different age groups report nearly similar accuracies when tested with different age groups. This suggests that models trained on spectral features show low sensitivity to the composition of age in the training records when classifying AF.

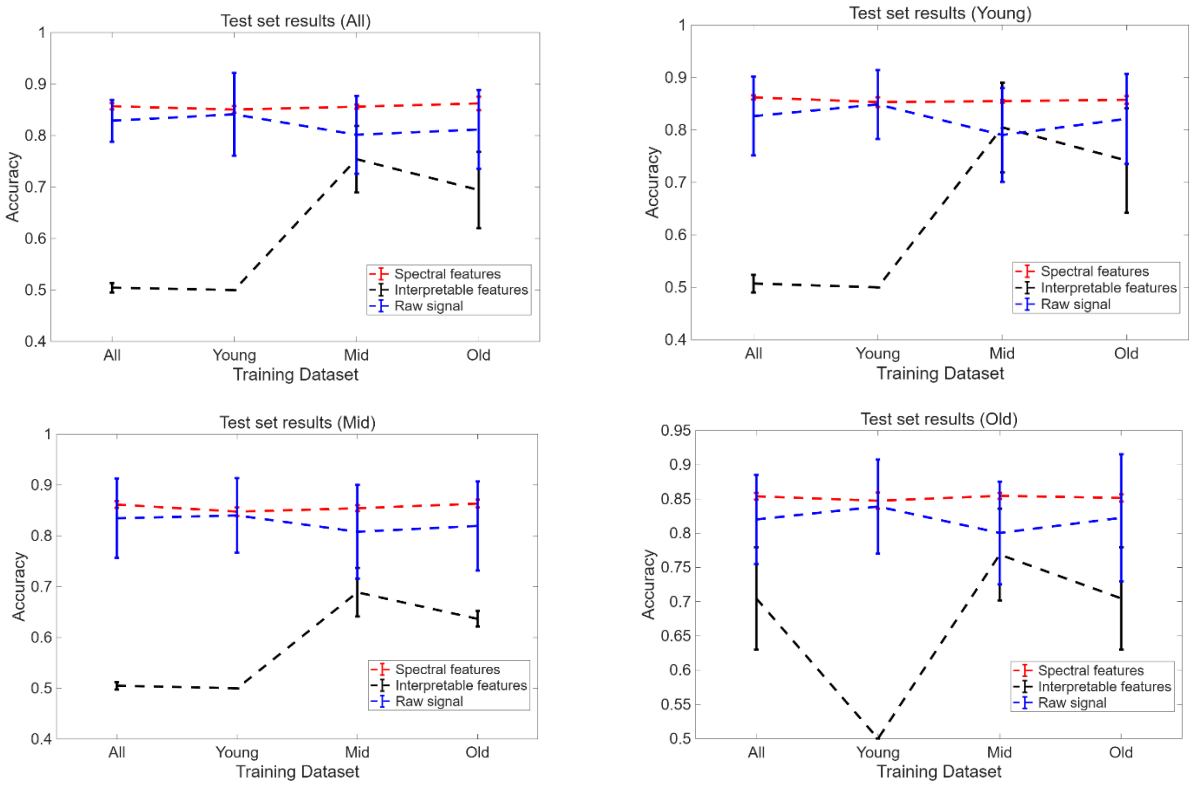

Figure 7 Mean and standard deviation of all the cross validation models applied to each of the test datasets. The plots show the results for all age groups and for each individual age group.

As a metric of interpretability, the reliability plot together with the UCE and ECE metrics are used to assess the model calibration and confidence. These were used in conjunction with the models trained on interpretable features to build interpretability and reliability to model predictions. Figure 8 represents the reliability plot for models trained on interpretable features. The reliability curve for the models trained on the Young dataset deviate furthest from calibration. Models trained on the Old and All datasets report least deviation from perfect calibration. It is also observed that in the higher confidence bins (0.55 - 0.85), all models' curves lie above the perfect calibration line indicating overconfidence. i.e., models are less correct than they think they are, and this varies with the age group compositions.

Table 3 reports UCE and ECE metrics for models trained on interpretable features when trained on datasets of different age composition. Models trained on Young data report very high UCE and ECE metrics in comparison to other models. The models trained on the Mid age group also report relatively high ECE scores. Models trained on the Old age group report the lowest ECE and UCE. These results highlight the impact of age and its composition in training data when classifying cardiac events.

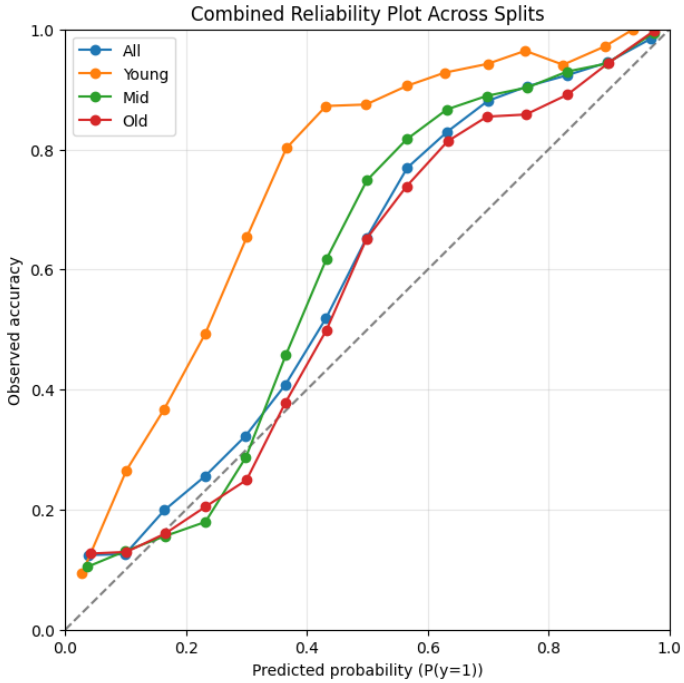

Figure 8 Reliability curves for models trained on various age compositions.

Table 3 Calibration metrics for models trained on interpretable features for different age compositions. The lowest errors are shown in bold.

Training data version	UCE	ECE
All	0.0797	**0.0482**
Young	0.1055	0.2268
Mid	0.0544	0.0926
Old	**0.0517**	0.0699

Effect of ethnicity imbalances

When investigating the effect of ethnicity, the datasets considered were All (containing a mixed balance of all ethnicities), White, Black, Hispanic or Asian. The cross validation models trained on the All dataset were additionally validated on the ethnicity specific datasets as well and are shown in Figure 9.

When considering models trained on raw PPG signals on the All dataset, we observe that when validated on the specific ethnic groups they report a declining trend in accuracy.

Models trained on All and validated on the White dataset report the highest accuracy (near ~82.5 %) while those validated on the Hispanic dataset report a noticeable decrease in accuracy (~72.5 %). When considering models trained and validated on matching ethnic specific groups we see a clear trend. Models trained and validated on White data report high accuracy (near ~82.5 %). This performance decreases drastically for other ethnic groups. Models trained on Black datasets report only ~72.5 % accuracy, Hispanic datasets report ~55 % accuracy and Asian datasets report ~65 % accuracy, suggesting a clear dependence of model reliability and performance on composition of ethnicity in the training data.

When considering models trained on interpretable features on the All dataset and validated on individual ethnicities, the White dataset once again reports the highest accuracy of ~82.5 %. The Hispanic dataset reports the lowest accuracy again at 75 %. When comparing the performance of models trained on individual ethnicity datasets, the White dataset once again reports the highest accuracy, which declines strongly for other datasets. Models trained on Black datasets report ~72.5 % accuracy, Hispanic at ~55 % and Asian ~65 %. These results once again suggest dependence of model performance and reliability on the composition of ethnicity in the training data.

When considering models trained on spectral features on the All dataset and validated on the individual ethnicity datasets, there is a slight deviation in trends as compared to other methods. The models report highest accuracy when validated on the Asian dataset ~90 % and lowest for models trained on the Black dataset (~75 %). For models trained and validated on matching datasets, the White datasets report highest performance accuracies (marginally ~85 %), showing close agreement with those trained on the All dataset. The models trained on the Black dataset report the lowest accuracy (~70 %).

Models trained on each of the ethnicity specific datasets were then tested on the hold out test set.

Models trained on the raw PPG signals tested on the individual ethnicities do not show much change in accuracy when trained on various ethnicity datasets. Most models report ~72 - 80 % accuracies, with models trained on the Black dataset often recording lowest accuracy, and in some cases, Hispanic datasets reporting slightly higher standard deviations (10 %).

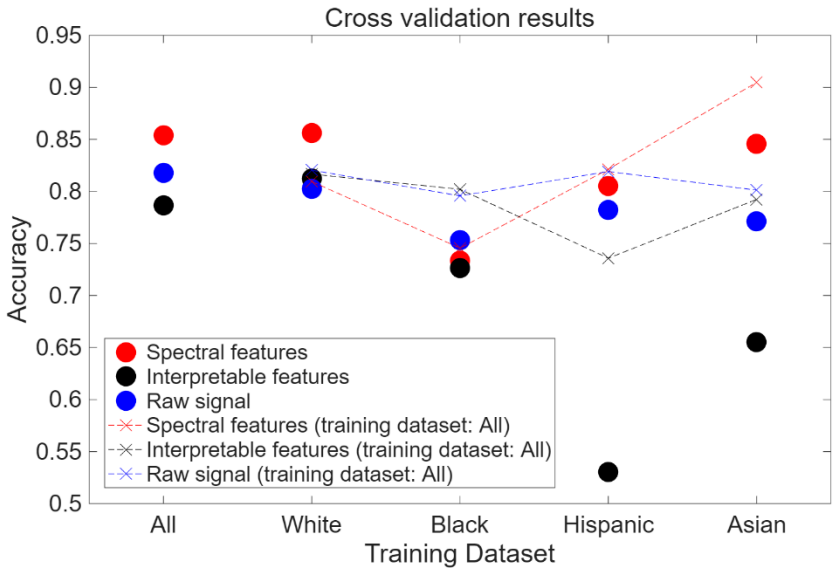

Figure 9 Cross validation results for the different training datasets. The crosses linked by dashed lines show the breakdown when trained on 'All' for the different ethnicity classes.

Models trained on interpretable features show noticeable differences in the accuracy when changing the training or testing datasets. When models trained on different ethnicities were tested on the All test set, models trained on white data report the highest accuracies (~75 %) while models trained on Hispanic data reported the lowest accuracy (50%). Models trained on Black data had high standard deviations (15 %). When the models were tested on a White test set, models trained on White data report the highest accuracy as expected, however, the models trained on Black and Hispanic data again show a stark decrease in accuracy and report large standard deviations (15 %). When testing models on the Black hold out test set, oddly enough, once again models trained on White data reported the highest accuracy, and models trained on Black data the least accuracy. Models trained on the Black and Hispanic data again report higher standard deviations. These trends are corroborated when testing on Hispanic data as well, although with lower standard deviations for models trained on Black and Hispanic data. When testing the models on Asian data, the highest accuracy was recorded for models trained on the All dataset, while the models trained on the Hispanic dataset recorded poorest accuracy (~50 %). Models trained on the Black dataset once again report very high standard deviations (~30 %).

When considering models trained on spectral features, accuracy and standard deviation trends are unique. When testing on the All dataset, models trained on the White dataset report the highest accuracy with minimal standard deviation. Models trained on the

Hispanic dataset report lowest accuracy with larger standard deviation. These trends are corroborated when testing on the White dataset, although standard deviations reported are considerably lower. When testing on the Black dataset, accuracy is nearly uniform for all models irrespective of the training set composition (~75 %). When testing on the Hispanic dataset, models trained on the White dataset report the highest accuracy while models trained on the Hispanic dataset report the lowest accuracy with higher standard deviation (~72.5 ± 20 %). Models tested on the Asian dataset also report the highest accuracy when trained on the White dataset. Models trained on the Hispanic dataset once again report very poor accuracy with the largest recorded standard deviation (~60 ± 40 %).

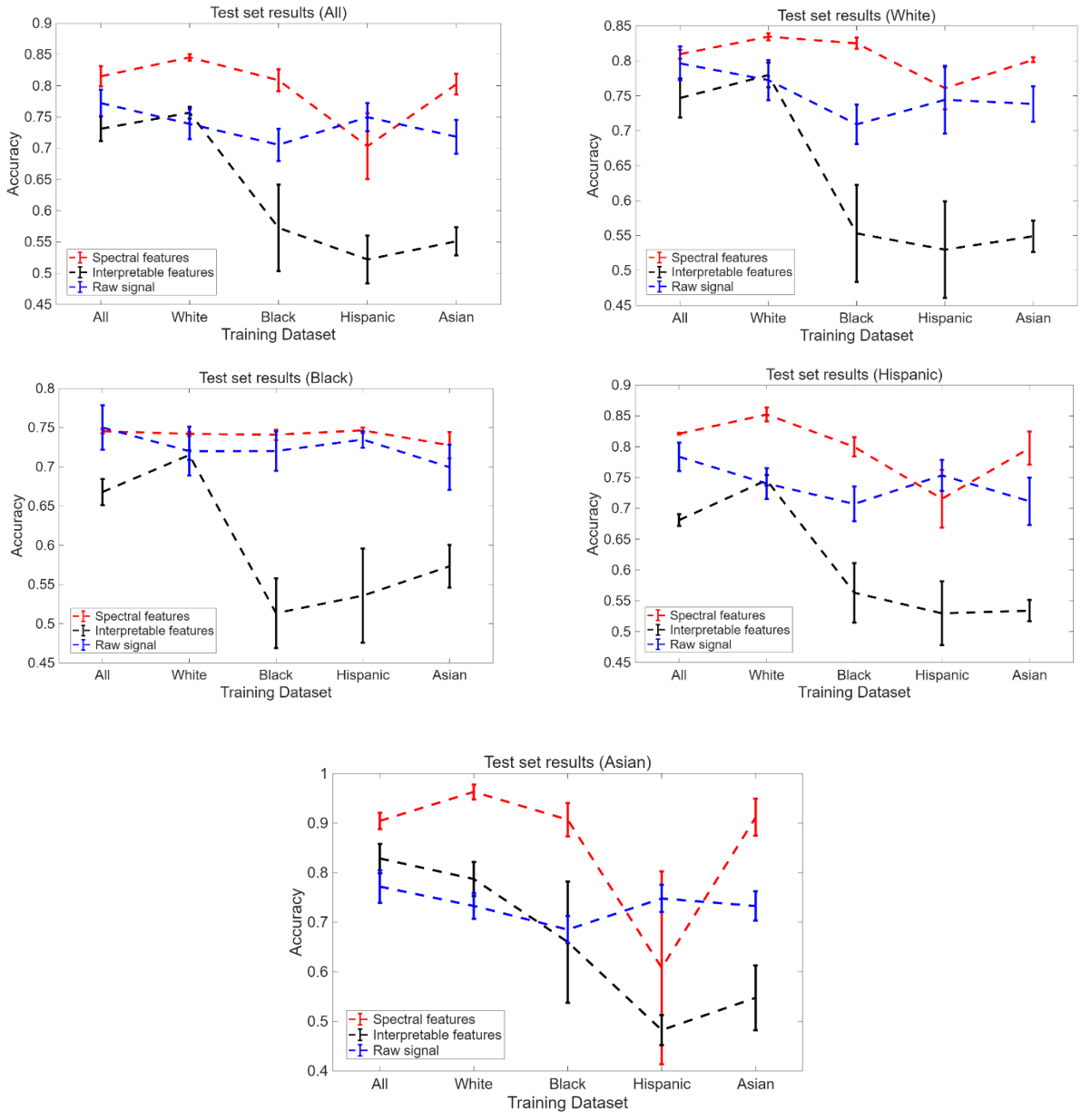

Figure 10 Mean and standard deviation of all the cross validation models applied to each of the test datasets. The plots show the results for all ethnicity groups and for each individual ethnic group.

As a metric of interpretability, the reliability plot together with the UCE and ECE metrics are used to assess the model calibration and confidence. These were used in conjunction with the models trained on interpretable features to build interpretability and reliability to model predictions. Figure 11 depicts the calibration curves for models trained on the various ethnicity datasets. It was observed that the models trained on the White dataset are very close to perfect calibration while the models trained on the Hispanic dataset are furthest from perfect calibration. These models not only deviate from perfect calibration but also show strong fluctuations in calibration, demonstrating strong overconfidence in lower confidence bins and under confidence in higher confidence bins. The models trained on the Black dataset deviate from perfect calibration noticeably as well. The findings of this plot indicate that the models trained on the All and White data are near perfect calibration and are reliable in their predictions whereas models trained on the other ethnicity datasets are not reliable for providing accurate classifications of cardiac abnormality.

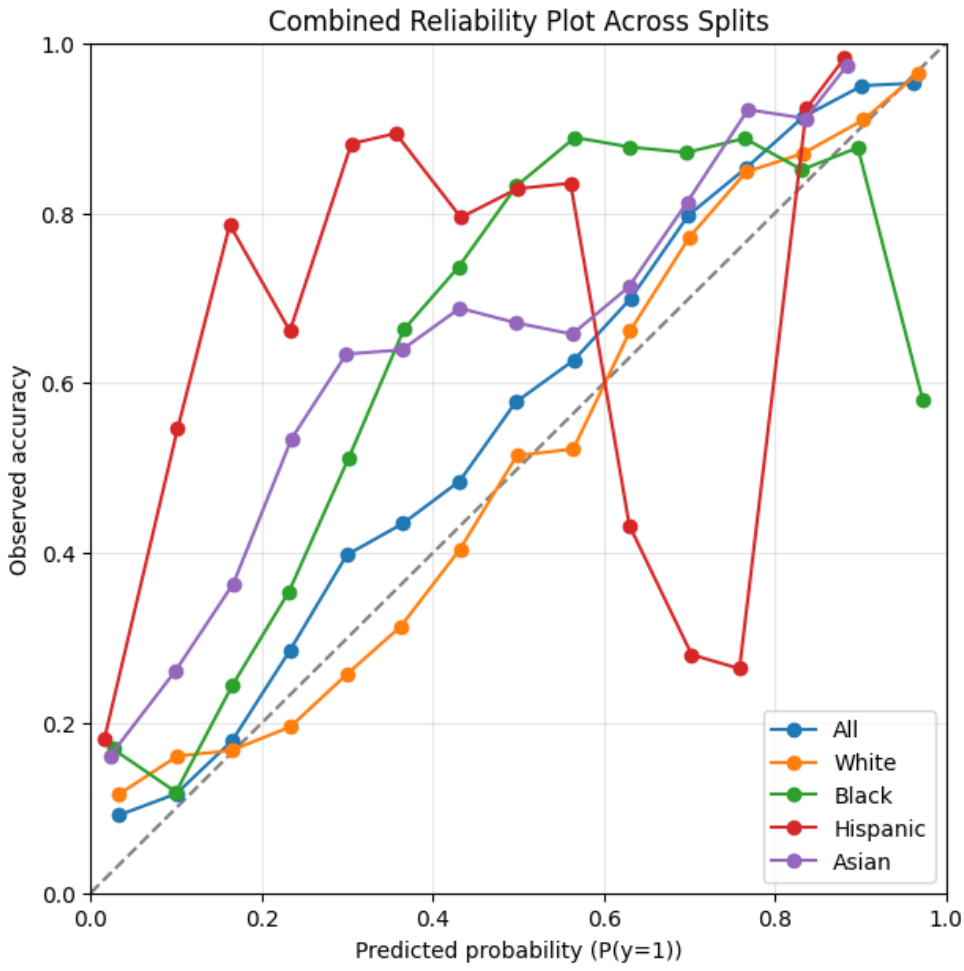

Figure 11 Reliability curves for models trained on datasets with varying composition of ethnicities.

Table 4 reports the UCE and ECE metrics for the models trained on datasets with varying ethnicities. Results from the table corroborate cross validation and hold out test set results reporting high UCE and ECE for models trained on the Hispanic dataset. Models trained on the Black dataset report relatively higher ECE, while models trained on the Asian dataset report relatively higher UCE and ECE metrics as well. These models have large errors and their predictions can be considered unreliable for high stakes decision making.

Table 4 Calibration metrics for models trained on datasets with varying ethnicities. The lowest errors are shown in bold.

Training data version	UCE	ECE
All	**0.0199**	0.0549
White	0.0352	**0.0400**
Black	0.0606	0.1758
Hispanic	0.3144	0.3420
Asian	0.1358	0.1858

Effect of BMI imbalances

The effect of composition of BMI in the dataset was assessed by training and validating the model on BMI specific groups as well as training on one dataset with all groups (All dataset) and validating on BMI specific groups.

When considering models trained on raw PPG features using the All dataset and validated on other BMI specific groups, models validated on the Severely Obese group reported the highest accuracy (~87.5 %) while other validation sets reported near similar accuracies (~85 %). When considering models trained and validated on BMI specific groups, models trained on the All dataset report the highest accuracy whereas models trained on the Healthy dataset report the lowest accuracy.

When considering the models trained on interpretable features using the All dataset and validated on the BMI specific groups, the highest accuracy was recorded when validated on the Obese category (~85 %) while the Overweight and Severely Obese categories record the lowest accuracy (~75 %). When considering models trained and validated on

BMI specific groups, the All, Healthy and Obese categories record similar accuracies (~80 %) while the Severely Obese category marks a noticeable drop in accuracy (~65 %).

When considering models trained on the spectral features using the All dataset and validated on BMI specific groups, reported accuracies were nearly similar irrespective of the training or test set composition. When considering models trained and validated on BMI specific groups, the Obese category reports the highest accuracy (~87.5 %) while the Overweight category reports the lowest accuracy (~80 %).

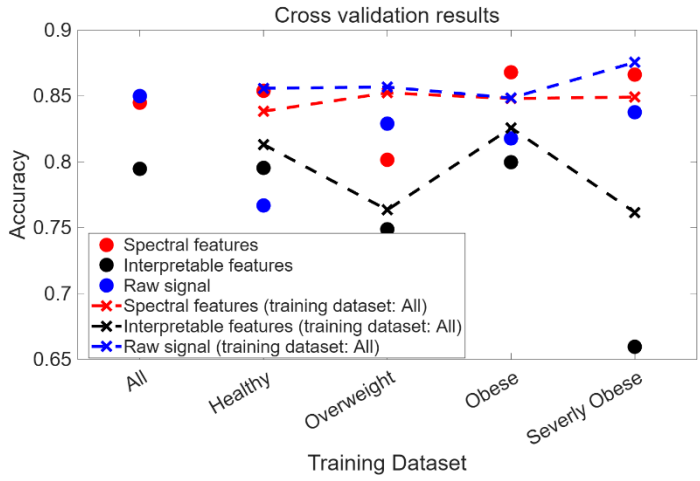

Figure 12 Cross validation results for the different training datasets. The crosses linked by dashed lines show the breakdown when trained on 'All' for the different BMI classes.

These models were then tested on the hold out test set to assess their performance and generalisability.

When considering models trained on raw PPG features tested on the All dataset, the All, Overweight and Severely Obese categories report ~80 % accuracy while the Healthy and Obese categories report ~78.5 % accuracy. All models report near 10 % standard deviations between cross validation models. These trends remain nearly similar for all test groups.

When considering models trained on interpretable features and tested on the All dataset, models, irrespective of the training set, report ~50 % accuracy with the Overweight and Obese models recording high standard deviations (~15 %). When testing on the Healthy test set, models once again record ~50 % accuracy except for the Obese dataset which recorded ~45 % with 15 % standard deviation. When testing on the Overweight and Obese

test sets, models report similar trends in performance recording ~50 % accuracy except for the Overweight, Obese and Severely Obese sets which report 55 % accuracy with 15 % standard deviations.

When considering models trained on spectral features on the All, Healthy, Overweight and Obese test sets, there are no noticeable differences in accuracy irrespective of changes in the training set composition. Accuracy only varies with the test set composition, recording ~80 %, ~87 %, ~82 %, ~87 % accuracies for All, Healthy, Overweight and Obese test sets, respectively. When testing on Severely Obese, nearly all models again report similar accuracies (~70 %) while the models trained on Healthy dataset report 75 % accuracy with 15 % standard deviation.

As a metric of interpretability, the reliability plot together with the UCE and ECE metrics are used to assess the model calibration and confidence. These were used in conjunction with the models trained on interpretable features to build interpretability and reliability to model predictions. Figure 14 shows the reliability curves for models trained on the different BMI groups. Models trained on the Severely Obese training data deviated furthest from calibration while the models trained on the Obese and Overweight datasets remain closer to perfect calibration, although still showing a moderate amount of deviation at higher confidence bins. All models are recorded above the line of perfect calibration indicating overconfidence. i.e. models think they are right more often than they actually are, and this shifts with the BMI composition of the training set.

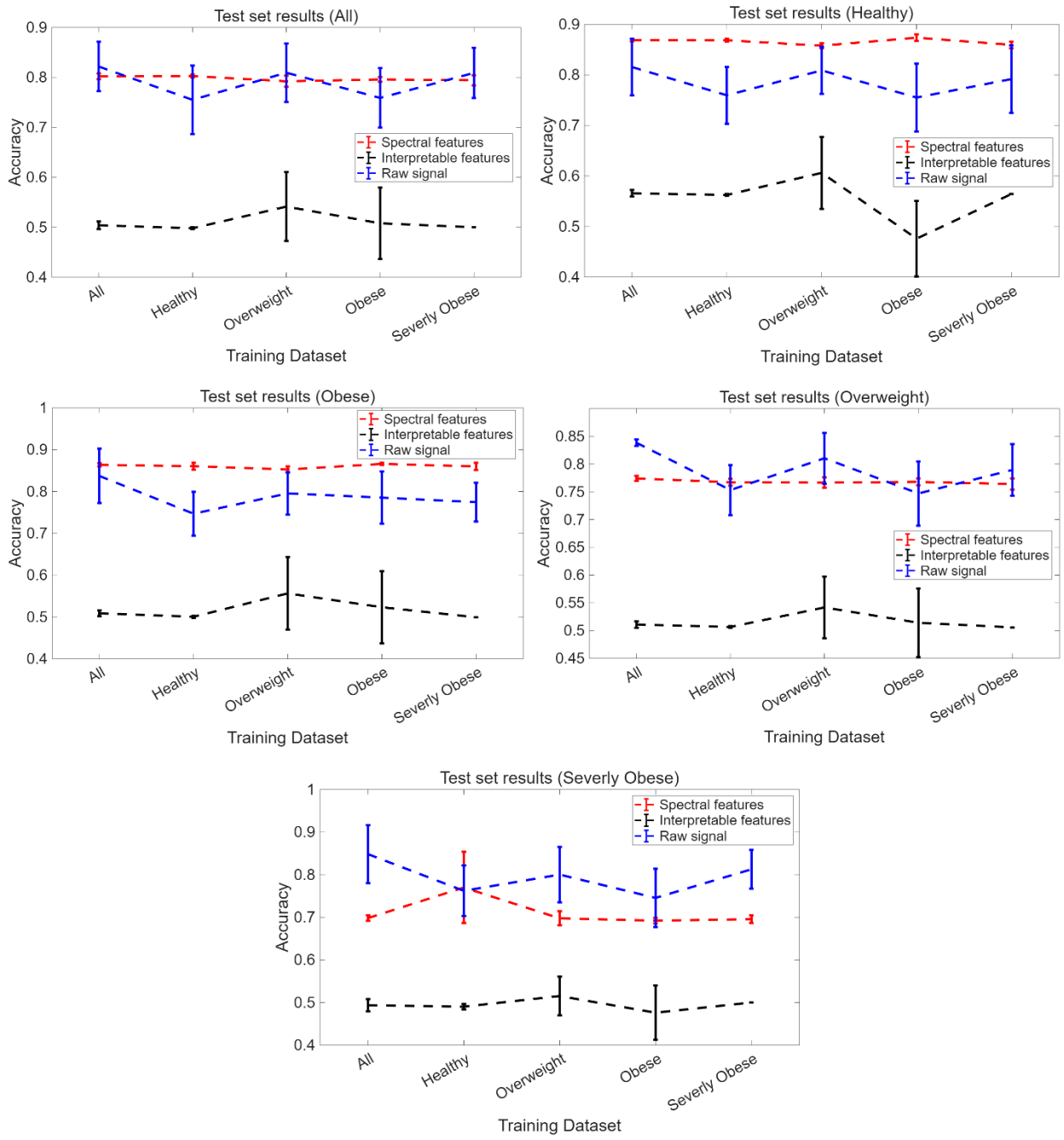

Figure 13 Mean and standard deviation of all the cross validation models applied to each of the test datasets. The plots show the results for all BMI classes and for each individual BMI class.

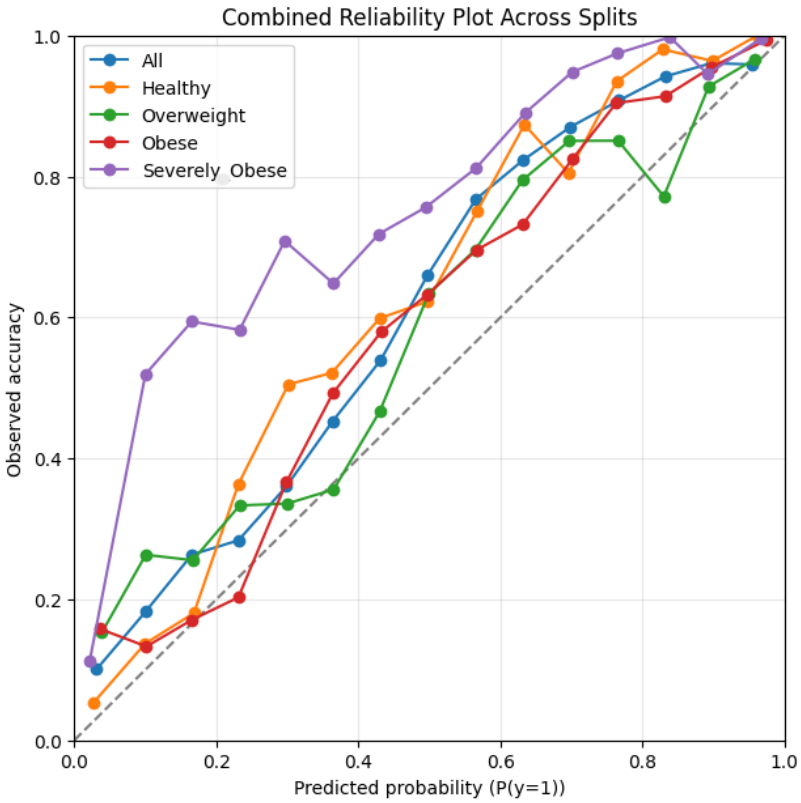

Figure 14 Reliability curves for models trained on datasets with varying composition of BMI groups.

Table 5 reports the UCE and ECE metrics for models trained on interpretable features with the varying BMI groups. The results from the table corroborate those with cross validation and hold out test set results reporting high ECE and UCE values for models trained on Severely Obese data.

Table 5 Calibration metrics for models trained on datasets with varying BMI groups. The lowest errors are shown in bold.

Training data version	**UCE**	**ECE**
All	0.0382	0.1002
Healthy	**0.0347**	0.1171
Overweight	0.0633	0.0956
Obese	0.0384	**0.0853**
Severely Obese	0.2444	0.2830

Discussions and conclusions

This report extends the Trustworthy and Safe AI Life Cycle case study [4] to a quantitative examination of how demographic imbalances in the training dataset can impact the detection of atrial fibrillation by influencing AI/ML model behaviour and reliability when trained on PPG data.

When investigating the impact of varying gender composition in the training set, models trained on raw PPG signals were highly sensitive to gender imbalance, depicting a clear preference for female dominated training sets while exhibiting large variability when the training sets were gender imbalanced. The models showed poor generalisability across gender shifts and are least robust to gender imbalance. Models trained on interpretable features were consistent, stable and robust to gender composition in the training set. The calibration metrics recorded were low and consistent confirming well calibrated confidence levels even when trained on single gender datasets, indicating high generalisability. Models trained on spectral features showed mild sensitivity. These results suggest that models trained on raw signals are picking up some gender differences in the signals, while the interpretable features are concentrating on features of the signal which are gender independent.

When investigating the impact of age in the training set, models trained on raw PPG signals showed mild dependence on age with the best performance for the Young cohort. In contrast, models trained on interpretable features showed very poor performance for the Young cohort and best performance for the Mid cohort, regardless of the composition of the training data. Correspondingly, calibration for the Young group was poorest, but similar for the other groups. Models trained on spectral features showed altogether different behaviour, with consistent performance across all age groups, thus showing insensitivity to this characteristic.

When investigating the impact of ethnicity composition in the training set, models trained on the raw PPG signals showed strong ethnicity dependence and large performance disparities when changing the composition of training or test sets. Models trained on the White dataset reported high accuracies which degraded sharply for models trained on other ethnicities, Hispanic being the least. Raw signal models thus showed strong

MCR-001, Issue 1, July 2025

sensitivity to ethnicity imbalance and performed significantly worse on minority ethnic groups. These results were corroborated by models trained on interpretable features which reported strong ethnicity sensitivity and very poor calibration for models trained on the Black, Hispanic and Asian ethnicities. The models exhibited poor generalisability and predictive performance and deviated strongly from ideal calibration under specific ethnicity imbalanced training circumstances. Spectral feature models in contrast showed higher robustness in general but still reveal some deviation for models trained on Hispanic data. It is notable that using only a single ethnic group for training did not give the best results for that ethnic group in the test data as might be expected, with the exception of White training data. Trends in ethnicity specific model performance can also be attributed to the dependence of PPG waveform quality on skin optical properties. Melanin concentration affects light absorption and scatter that can alter signal morphology and noise characteristics [26]. Signal to noise ratio and amplitude have also been shown to vary with skin tone [26], [27]. This drawback of PPG signals can justify why models trained on White datasets generalise poorly to other ethnicities with different skin pigmentation as observed for Black, Hispanic and Asian datasets. Since spectral feature models show higher robustness, it can be concluded that frequency domain representations can help mitigate ethnicity related signal variability but cannot eliminate it completely.

When investigating the impact of BMI composition in the training set on model performance, models trained on raw signals showed mild sensitivity to BMI composition, reporting nearly stable performance across all models. Models trained on an evenly distributed training set consistently performed better than other models while the Healthy models recorded the worst performance. When considering models trained on interpretable features, there was significant degradation in model performance as BMI composition shifts in the training set. Models reported mediocre performance with high variability, as well as consistent deviation from ideal calibration. Models trained on interpretable features were thus highly sensitive to BMI composition in the training set and were poorly calibrated with performance degrading on most hold out test sets. Spectral feature models were most robust to shifts in BMI composition in the training set and showed reasonable generalisability across all test sets. These models reported stable accuracy patterns and limited sensitivity to training set composition. Physiologically, individuals with higher BMI or obesity are at higher risk of cardiac abnormalities including

AF [28] and therefore, AF related physiological patterns may appear more predominantly in higher BMI cohorts, allowing models to learn these patterns more effectively.

The findings from this study reinforces the need for fairness, balance and demographic aware evaluations when integrating AI models in high stakes decision making such as in clinical settings. The case study in consideration – classification of atrial fibrillation – assessed across gender, age, ethnicity and BMI reveals that models consistently learned patterns rooted in demographic characteristics, not just pathology, that resulted in unequal predictive performance and unreliable model calibration when the training data was imbalanced or with mismatched testing conditions. Moreover, the categories assessed highlights the urgent need for fairness during model development since AF disproportionately affects older populations and those with higher BMI, while PPG signal quality is known to vary with skin tone. Overlooking such critical factors can compound existing biases and compromise the reliability of model predictions. Results from this report highlight exactly this; model reliability and calibration degrade the most when the models are used outside of the context they are trained and can exacerbate the risks associated with deployment in high stakes decision making. Additionally, the choice of input data (in our case, raw PPG signals, spectral or extracted features) along with the choice of ML model has a heavy impact on sensitivity to demographic bias and should be chosen appropriately. Ultimately these findings emphasise the need for fair model development with diverse and balanced training data accompanied by transparent reliability assessments to provide meaningful support to decision making processes.

References

[1] E. Tabassi, “Artificial Intelligence Risk Management Framework (AI RMF 1.0).” 2023. Accessed: Mar. 05, 2025. [Online]. Available: https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10

[2] M. Levene *et al.*, “A life cycle for trustworthy and safe artificial intelligence systems.” Accessed: Mar. 26, 2026. [Online]. Available: https://doi.org/10.47120/npl.MS57

[3] “Good machine learning practice for medical device development: Guiding principles,” GOV.UK. Accessed: Mar. 26, 2026. [Online]. Available: https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles/good-machine-learning-practice-for-medical-device-development-guiding-principles

[4] M. Alsuleman, P. Duncan, and A. Thompson, “Screening of atrial fibrillation using wearable PPG devices – a trustworthy and safe AI life cycle case study.” Accessed: Mar. 26, 2026. [Online]. Available: https://doi.org/10.47120/npl.MS61
[5] Artificial Intelligence/Machine Learning-enabled Working and IMDRF, “Good machine learning practice for medical device development: Guiding principles.” 2025. [Online]. Available: https://www.imdrf.org/sites/default/files/2025-02/IMDRF_AIML%20WG_GMLP_N88%20Final.pdf
[6] “Software and AI as a medical device change programme,” GOV.UK. Accessed: Mar. 26, 2026. [Online]. Available: https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme
[7] M. Moulaeifard, P. Charlton, and N. Strodthoff, “MIMIC-III-Ext-PPG: A PPG Benchmark Dataset for Cardiorespiratory Analysis.” [Online]. Available: https://physionet.org/content/mimic-iii-ext-ppg/1.0.0/
[8] B. Moody, G. Moody, M. Villarroel, G. D. Clifford, and I. Silva, “MIMIC-III Waveform Database Matched Subset.” [Online]. Available: https://physionet.org/content/mimic3wdb-matched/1.0/
[9] “WHO Consultation on Obesity (1999: Geneva, Switzerland) & World Health Organization (2000). Obesity : preventing and managing the global epidemic : report of a WHO consultation. World Health Organization.” [Online]. Available: https://iris.who.int/handle/10665/42330
[10] “Obesity.” [Online]. Available: https://www.nhs.uk/conditions/obesity/
[11] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 1st ed. USA: Prentice Hall PTR, 1994.
[12] R. W. Ramirez, *The FFT, Fundamentals and Concepts*. Prentice Hall, 1984.
[13] “Atrial fibrillation detection using feedforward neural networks and automatically extracted signal features | IEEE Conference Publication | IEEE Xplore.” Accessed: Mar. 24, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8331735
[14] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Jan. 30, 2017, *arXiv*: arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980.
[15] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in *Proceedings of the 33rd International Conference on Machine Learning*, PMLR, Jun. 2016, pp. 1050–1059. Accessed: Mar. 24, 2026. [Online]. Available: https://proceedings.mlr.press/v48/gal16.html
[16] “LIII. On lines and planes of closest fit to systems of points in space : The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics: Vol 2 , No 11 - Get Access.” Accessed: Mar. 24, 2026. [Online]. Available: https://www.tandfonline.com/doi/pdf/10.1080/14786440109462720
[17] R. Jahangir, M. N. Islam, Md. S. Islam, and Md. M. Islam, “ECG-based heart arrhythmia classification using feature engineering and a hybrid stacked machine learning,” *BMC Cardiovasc Disord*, vol. 25, no. 1, p. 260, Apr. 2025, doi: 10.1186/s12872-025-04678-9.
[18] C. Reid Turner, A. Fuggetta, L. Lavazza, and A. L. Wolf, “A conceptual basis for feature engineering,” *Journal of Systems and Software*, vol. 49, no. 1, pp. 3–15, Dec. 1999, doi: 10.1016/S0164-1212(99)00062-X.
[19] S. Saha, U. Saha, S. Saha, and S. A. Fattah, “PPG-Based Feature Extraction for Type II Diabetes Prediction: A Machine Learning Approach,” in *2024 2nd International Conference on Computer, Communication and Control (IC4)*, Feb. 2024, pp. 1–6. doi: 10.1109/IC457434.2024.10486372.

[20] "WFDB Software for Python: A toolkit for physiological signals | IEEE Conference Publication | IEEE Xplore." Accessed: Feb. 22, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10176714
[21] M. Pavlovic, "Understanding model calibration -- A gentle introduction and visual exploration of calibration and the expected calibration error (ECE)," Sep. 12, 2025, *arXiv*: arXiv:2501.19047. doi: 10.48550/arXiv.2501.19047.
[22] M.-H. Laves, S. Ihler, K.-P. Kortmann, and T. Ortmaier, "Uncertainty Calibration Error: A new metric for multi-class classification," Oct. 2020, Accessed: Mar. 24, 2026. [Online]. Available: https://openreview.net/forum?id=XOuAOv_-5Fx
[23] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. Inman, "1D convolutional neural networks and applications: A survey," *Mech. Sys. Sig. Proc.*, vol. 151, p. 107398, 2021.
[24] S. Guessoum *et al.*, "The short-term prediction of length of day using 1D convolutional neural networks (1D CNN)," *Sensors*, vol. 22, p. 9517, 2022.
[25] X. Xu, Y. Zhang, Y. Rao, Y. Xu, Y. Gao, and H. Tu, "A deep learning-guided ensemble empirical mode decomposition method for single-channel fetal electrocardiogram extraction," *Sensors*, vol. 26, no. 7, p. 2037, 2026, doi: 10.3390/s26072037.
[26] L. S. O. Sánchez, J. M. May, and P. Kyriacou, "Evaluation of skin pigmentation effect on photoplethysmography signals using a vascular finger phantom with tunable optical and mechanical properties," *J. Biomed. Opt.*, vol. 30, no. 11, p. 117002, Nov. 2025, doi: 10.1117/1.JBO.30.11.117002.
[27] "22HLT01 QUMPHY D3 deliverable report," 2026.
[28] T. M. Powell-Wiley *et al.*, "Obesity and cardiovascular disease: A scientific statement from the American Heart Association," *Circulation*, vol. 143, no. 21, pp. e984–e1010, May 2021, doi: 10.1161/CIR.0000000000000973.

National Physical Laboratory

Investigation of the Impact of Demographic Imbalance in the Training Data on Fairness in Machine Learning

Providing the measurement capability
that underpins the UK's prosperity
and quality of life

To find out more about NPL:

npl.co.uk

To get in touch:

npl.co.uk/contact

020 8977 3222

Hampton Road
Teddington
Middlesex
TW11 0LW

Please consider the environment
before printing this document