Journal of Healthcare Informatics Research (2023) 7:225–253
https://doi.org/10.1007/s41666-023-00133-6

RESEARCH ARTICLE

Bias Analysis in Healthcare Time Series (BAHT) Decision Support Systems from Meta Data

Sagnik Dakshit[1] · Sristi Dakshit[1] · Ninad Khargonkar[1] · Balakrishnan Prabhakaran[1]

Received: 12 October 2022 / Revised: 19 April 2023 / Accepted: 12 May 2023 /
Published online: 19 June 2023

Abstract

One of the hindrances in the widespread acceptance of deep learning–based decision support systems in healthcare is bias. Bias in its many forms occurs in the datasets used to train and test deep learning models and is amplified when deployed in the real world, leading to challenges such as model drift. Recent advancements in the field of deep learning have led to the deployment of deployable automated healthcare diagnosis decision support systems at hospitals as well as tele-medicine through IoT devices. Research has been focused primarily on the development and improvement of these systems leaving a gap in the analysis of the fairness. The domain of FAccT ML (fairness, accountability, and transparency) accounts for the analysis of these deployable machine learning systems. In this work, we present a framework for bias analysis in healthcare time series (BAHT) signals such as electrocardiogram (ECG) and electroencephalogram (EEG). BAHT provides a graphical interpretive analysis of bias in the training, testing datasets in terms of protected variables, and analysis of bias amplification by the trained supervised learning model for time series healthcare decision support systems. We thoroughly investigate three prominent time series ECG and EEG healthcare datasets used for model training and research. We show the extensive presence of bias in the datasets leads to potentially biased or unfair machine-learning models. Our experiments also demonstrate the amplification of identified bias with an observed maximum of 66.66%. We investigate the effect of model drift due to unanalyzed bias in datasets and algorithms. Bias mitigation though prudent is a nascent area of research. We present experiments and analyze the most prevalently accepted bias mitigation strategies of under-sampling, oversampling, and the use of synthetic data for balancing the dataset through augmentation. It is important that healthcare models, datasets, and bias mitigation strategies should be properly analyzed for a fair unbiased delivery of service.

Keywords Fairness · Bias analysis · Bias mitigation · Synthetic data · Decision support systems

✉ Sagnik Dakshit
sdakshit@utdallas.edu

Extended author information available on the last page of the article

1 Introduction

Recent advances in machine learning and deep learning (DL) with multi-modal data have led to their deployment in consumer-centric applications. The superior performance of deep learning models has been shown to be competitive with experts [1]. Despite the success of DL systems in terms of model performance, they are not without shortcomings. The presence of undetected biases, data availability owing to privacy and cost, model drift, and lack of proper evaluation methods are some of the concerns hindering the widespread acceptance of DL-based decision support systems. The deployment of such automated systems in critical domains such as disease detection and fraud prevention aggravates the need to address these challenges leading to the advent of fair machine learning (FAIR ML). Fair ML is a nascent research domain with the goal of introducing accountability, transparency, and fairness in intelligent systems. It also deals with ensuring that the systems do not partake in any discrimination or show disparate impact towards end-users.

Fairness in ML as defined by Oneta et al. [2] is the field of study ensuring that model outputs are not dependent on sensitive attributes. Models showing dependence on protected attributes are considered unfair and have a disparate impact. Intelligent systems limited in this capacity perform poorly in the real world also termed model drift and consequently hinder their public acceptance. Some illustrative instances of model drift due to unchecked bias (leading to unfair systems) are as follows: (1) Google photos classifying African people as chimpanzees due to data bias, (2) localized higher interest rates based on zip codes as proxy variables, and (3) bias towards colored people [3]. These decision errors demonstrate the need for an in-depth analysis of the fairness of decision support systems in critical domains such as healthcare, finance, and laws. However, the traditional metrics for intelligent system algorithms such as accuracy, precision, recall, and F1 score do not evaluate bias or fairness. The above shortcomings in intelligent systems have led to an increased research focus on dealing with fairness, accountability, and transparency in datasets and models used to train decision support systems and the deep learning model for critical domains.

1.1 Bias in Healthcare

Healthcare data finds computational applications prevalent in clinical and tele-medicine settings. The use of decision support systems in safety–critical applications to improve services to patient and clinicians bolsters the requirement for a systematic analysis of bias. In the domain of healthcare, 2D images, electronic healthcare records (EHR) and time series 1D signals are the three primary modalities of data. Bias and fairness have been well studied in data modalities of 2D images and EHR leaving a gap in the investigation of 1D time series signals [4–10]. Our work focuses on 1D time series physiological signals which are used to diagnose physiological conditions such as changes in heart and brain activity. To the best of our knowledge, this is the first work on analyzing the bias in time series

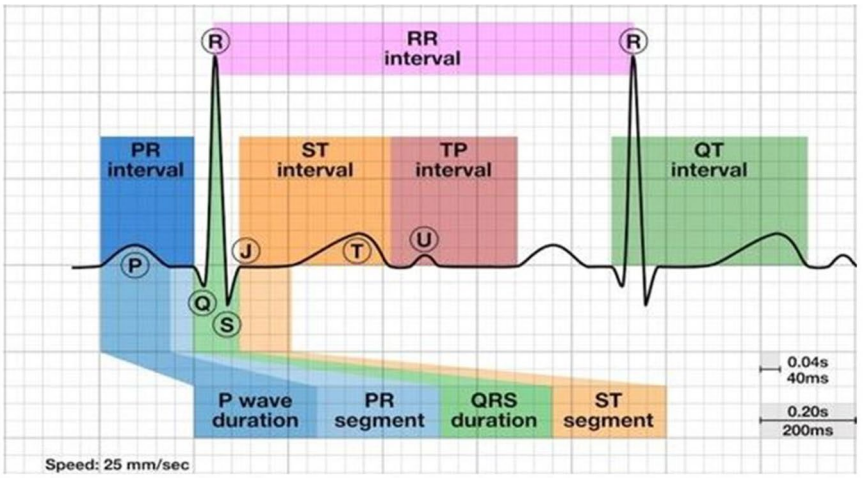
healthcare electrocardiogram (ECG) and electroencephalogram (EEG) datasets. The challenge of data availability owing to privacy laws in healthcare such as HIPAA leads to imbalanced and noisy datasets responsible for bias leading to disparate impact in the real world. In most domains, tasks such as semantic segmentation and object detection allow for the identification of imbalances inherently. However, the same does not extend completely to the healthcare domain, where 2D, 3D scans, and bio-signals do not have any information related to the bias. Once identified, the mitigation of bias is crucial at the source as privacy laws do not allow their unrestricted sharing without obfuscation and desensitization. Furthermore, the bias analysis methods on 2D and 3D data do not extend to 1D time series signals.

1.2 Contributions

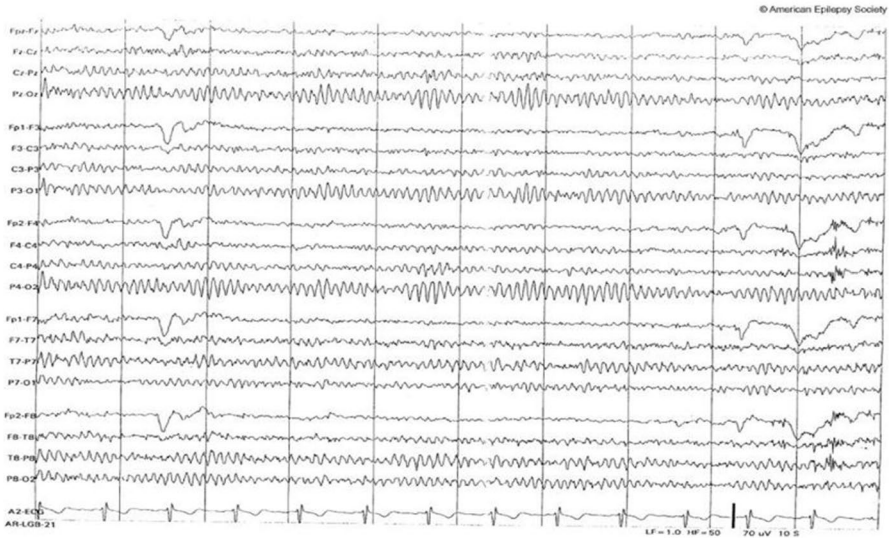
In this paper, we propose a framework for bias analysis in healthcare time series data (BAHT). Bias in raw time series signals such as electrocardiogram (ECG; Fig. 1a) and electroencephalogram (EEG; Fig. 1b), unlike categorical or image data, is difficult to identify owing to their desensitized nature. Towards this, our proposed BAHT framework extracts such information from the metadata at the source, which is structured reference data providing descriptions of the dataset itself. The unchecked bias often gets amplified in the learning process, leading to the problem of bias amplification. We compare the prevalent bias mitigation strategies and their effect on reducing bias amplification for arrhythmia detection using ECG signals for supervised algorithms such as deep learning.

Our primary contributions can be summarized as follows:

- We propose a framework BAHT (Fig. 2) for a systematic evaluation of bias in time series healthcare decision support systems. To the best of the authors' knowledge, this is the first work in the domain of healthcare with ECG and EEG time series data.
- BAHT aids developers to analyze and evaluate the bias of decision support systems using ECG and EEG time series data through graphical interpretation in terms of protected variables from metadata.
- We introduce and analyze the concept of combination protected group bias (CPGB) in healthcare ECG and EEG time series datasets. CPGB helps identify the bias that exists in datasets when two or more protected variables are combined as an entity as illustrated in Section 3.
- BAHT's bias amplification module can serve as a tool for supervised model bias amplification measurement and provides a bias amplification analysis for time series ECG and EEG deep learning-based systems.
- We show a focused experiment for model drift on the task of core-set selection for incoming streams of ECG signals, in the real world due to unanalyzed bias in the dataset and learning algorithm.
- We present a comparison of prevalent bias mitigation strategies on the effect of bias amplification.



a ECG Time-Series Signal



b EEG Time-Series Signal

Fig. 1 **a** ECG time series signal. **b** EEG time series signal

2 Related Works

In our extensive search, we did not find any fairness studies and bias mitigation strategies focused on biomedical 1D heart and brain signals, though a large body of work discusses bias in AI, pitfalls, and benefits theoretically [11]. Bias mitigation is a nascent field and research in the same has been primarily focused on 2D images and 3D scan modalities of data in healthcare. Most of the existing work can

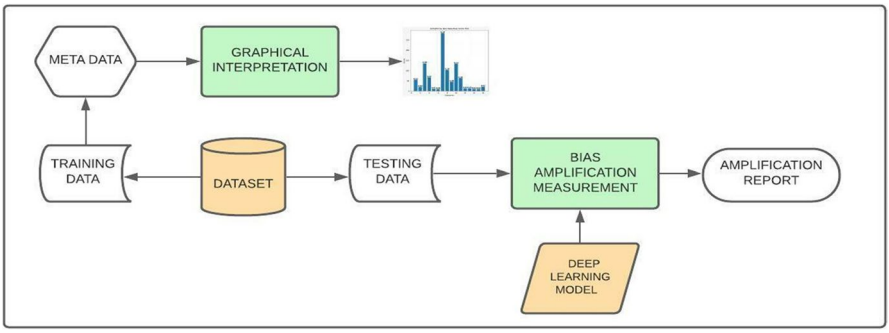


Fig. 2 Bias analysis in healthcare time series (BAHT) framework; yellow: input dataset to be analyzed and the deep learning model for bias amplification analysis. Green: graphical interpretation tool and bias amplification measurement tool. The bar plot from the GI module is to represent the type of generated output and is illustrated in detail in Fig. 3

be grouped as investigation studies and pre-processing, in-processing, and post-processing strategies. Torralba et al. [12] demonstrated the presence of bias in popular research 2D image datasets highlighting the problem and need for mitigation. Bias analysis and mitigation have also been discussed for critical tasks such as human trafficking detection classifiers by Hundman et al. [13]. They propose to mitigate bias by removing any information that relates to a biased feature and by maintaining the distribution of the biased feature for both positive and negative samples. Similarly, in [14], authors propose a two-stage framework to identify bias based on epistemological principles for the task of hiring decisions. Dixon, Lucas et al. [15] focus on the measuring and mitigation of bias for the task of text classification. The authors highlight the difference between unfair systems and the presence of unintended bias. The latter is attributed to class imbalance, and they propose mitigation by adding more samples. The authors also show bias quantification in the domain of textual data by calculating the likelihood of frequency of words and in terms of the model by area under the curve (AUC) and pinned AUC measurement. Their proposed bias identification and quantification approach are limited to the domain of natural language, and the measurement of AUC does not show the effect of how bias is amplified in model learning. Gurupur et al. [16] discusses inherent bias in healthcare decision support systems. They propose using information gain to measure the reliability of such systems without any experimental justification. The authors also acknowledge the requirement for higher levels of measurement to ascertain inherent bias. Puyol-Anton et al. [17] investigate fairness in cardiac magnetic resonance imaging (MRI) images to show the effect of an imbalanced dataset on racially biased models. They investigated strategies such as fair meta-learning, stratified batch sampling, and protected group models to reduce the bias. The results show that the best way to reduce this type of bias is using protected group models where models are developed separately for each protected group. Authors Duprez et al. [18] and Kishi et al. [19] discuss bias in cardiology with a focus on biases from race, ethnicity, and age on specific conditions of atherosclerosis and coronary artery disease. Unlike images and tabular data, signal datasets such as ECG are not widely available and

have fewer samples. This restricts the use of protected group models to train multiple data-hungry deep learning models. Puyol-Anton et al. used cardiac MRI data from the UK Biobank having samples of 5903 subjects in comparison to 47 subjects for the MIT-BIH arrhythmia dataset from PhysioNet [20] hindering the development of capable deployable classification models. The use of synthetic data has also been proposed as a strategy to mitigate bias due to imbalanced classes [21].

We can observe that bias analysis techniques and mitigation strategies are specific to the task, algorithm used, and data modality. To the best of the authors' knowledge, this is the first body of work on the analysis of the presence of unintended bias in 1D time series healthcare systems from popular ECG and EEG datasets and observe the effect of bias amplification.

3 The Bias Problem

Bias in machine learning is observed as the prejudice in favor or against someone or something in a way considered unfair [22, 23]. It has been long established in the ML domain that bias-free learning is futile. Machine learning algorithms need to make a priori assumptions, without which there would be no intuition behind classifying test data. While bias is important for learning, it is crucial to mitigate unwanted and derogatory biases such as bias against protected groups, measurement bias, and exclusion bias responsible for model drift in real world. Bias mitigation allows for fair evaluation and performance improvement in ML models. In literature [22–25], derogatory biases are broadly classified for healthcare as follows:

- Algorithmic bias: This bias is induced by algorithmic calculations leading to model bias amplification.
- Sample bias: This bias arises due to small datasets or imbalanced datasets leading to improper learned feature space.
- Prejudice bias: The datasets used to train the ML systems are prejudiced and have stereotypes such as gender stereotypes in working professionals.
- Measurement bias: This bias arises due to errors in measurement instruments or assessment.
- Exclusion bias: This type of bias is introduced due to excluding consequential data samples.

3.1 Sample and Prejudice Bias

Sample and prejudice bias both primarily concern protected groups. Protected groups are a group of people with characteristics that require legal protection from discrimination. There are 10 federally identified protected groups: race, color, ancestry, sex, age, citizenship, physical or mental disability, genetic information, and veteran status [26]. For our purposes concerning time series healthcare systems, protected groups such as citizenship, physical or mental disability, and genetic information, veteran status is ignored due to a lack of available information. For time series healthcare databases,

we focus on sex, age, and race as the protected groups. The sample bias with respect to the protected classes of interest is aggravated by privacy laws such as HIPAA, cost of collection, processing, and annotation sometimes leading to highly imbalanced datasets.

3.2 Algorithmic Bias

The involved mathematical calculations of learning from data have been shown to induce and amplify data bias, termed algorithmic bias [27, 28]. Model bias amplification due to algorithmic bias affects the performance of the deployed models in the real world also known as model drift.

3.3 Measurement Bias

Healthcare time series data are captured with various distributed calibrated devices which are sensitive to external influences. Calibration errors and a multitude of external stimuli such as sudden emotions are prominent sources of noise in bio-signals leading measurement bias.

3.4 Exclusion Bias

The exclusion bias is introduced in processing the data such as removing noisy samples which are introduced from multiple collection sources. The removal of noisy samples often leads to a loss of information due to removal of important samples. This loss of knowledge can be minimized by proper exclusion bias analysis. Selection of important and noisy samples to reduce exclusion bias and improve performance are active research areas termed active learning and core-set selection.

3.5 Combination-Protected Group Bias

Bias is usually discussed for protected groups, overlooking the bias problem in combinations of protected groups at a granular level. We introduce a new way of interpreting bias, *Combination-Protected Group Bias* (CPGB) as the bias that persists in the combination of protected groups. CPGB highlights the importance of granularity in bias analysis. This fine-grained analysis yields interesting insights into the public datasets prevalent in the healthcare domain and draws attention to some combinations which are also prone to learning bias by classification systems. We analyze and discuss potential bias in such combinations. The concept of CPGB is important, as it allows bias investigation at a granular level of classification. The mitigating of bias in protected groups does not guarantee bias mitigation at the granular levels. For example, while gender is a protected group whose distribution is balanced for a fair system, the distribution for age-wise gender groups may remain imbalanced. Superficially, the system might seem bias-free with respect to gender and age independently, while the system might be biased towards a granular group such as women in the age group of 20–30 years. We further illustrate this in Section 5 with public ECG datasets used for training ML classification models.

4 BAHT Framework

In this paper, we propose a bias analysis framework for healthcare time series data-based decision support systems, BAHT, as shown in Fig. 2. BAHT establishes the existence of the different types of biases from the metadata associated with prominent ECG and EEG time series healthcare datasets. Our proposed framework serves as a semi-automatic tool to aid experts in interpreting the bias in any given dataset and deep learning models. The semi-automatic nature of the tool is owing to its manual input of testing dataset, trained model, and expert interpretation of the biased result for model capacity improvement. BAHT consists of two modules:

(1) Graphical interpretation module
(2) Bias amplification module

Graphical interpretation module The module outputs bar plots with ratios for protected variables and combined protected group bias (Section 3.5). The selection of the protected variables involved in a dataset and the interpretation of the graphical results presented in the form of class bias score is benefitted from the presence of an expert.

Bias amplification module This module evaluates the decision model and the effect of bias amplification on the test dataset. The module generates *class-amplified bias score* for each class by subtracting the bias score of the test dataset from the bias score of the predicted classes.

We illustrate BAHT modules and provide a detailed analysis showing the presence of bias, its amplification, and effect of bias mitigation strategies on the popular time series ECG and EEG research datasets. We identify two prominent ECG and one EEG time series dataset for demonstrating the presence, effect of bias, and mitigation strategies. We use the following three datasets:

(i) Dataset A: "1000 Fragments" dataset by P. Plawiak [29] which is processed and curated from the MIT BIH database. This well-established research ECG dataset has normal sinus rhythm, pacemaker rhythm, and 15 classes of abnormal ECG signals collected from 45 patients (19 female and 26 male).
(ii) Dataset B: MIT BIH arrhythmia database [30] with 17 annotated classes of beats labelled 0–16. The dataset has 48 half-hour ECG excerpts from 47 subjects. No additional information has been provided in the metadata.
(iii) Dataset C: The EEG alcoholic dataset [31] is an experimental study conducted by the University of New York Health Center to identify the effect of the sequence of images shown to genetically predisposed alcoholics. The dataset was collected from 122 subjects across 120 trials per subject. The metadata does not provide any additional information about the dataset.

5 Graphical Interpretation of Bias

In this section, we analyze the presence of bias in the identified protected variables and CPGB. An expert identifies the various protected variables associated with the training dataset and the module provides interpretable bias statistics in a graphical format that is often overlooked in decision support system development. The identification allows for mitigating disparate impact. Disparate impact analysis measures selective treatment towards protected groups, not following the 80% or 4/5th rule of participation (https://www.justice.gov/crt/fcs/T6Manual7). The 80% rule states that if for a protected group, the rate of selection is less than 80% of the group with the highest number of samples then there is disparate impact. The bias score for each class is calculated as a ratio of the count of group samples with the group with maximum samples as shown in Eq. 1, where "sample count" represents the number of samples in the intended group and *"i"* represents the groups. Each class or protected variable can be represented as a group.

$$Bias\ Score_i\ (BS) = \frac{Sample\ Count_i}{max(Sample\ Count)} \tag{1}$$

5.1 ECG Dataset A

As Dataset A, we consider the "1000 Fragments" dataset as illustrated in Section 4. This well-established research ECG dataset has normal sinus rhythm, pacemaker rhythm, and 15 classes of abnormal ECG signals. The metadata for this dataset does not provide any further information. The observations from the graphical interpretation module as shown in Figs. 3 and 4 can be listed as:

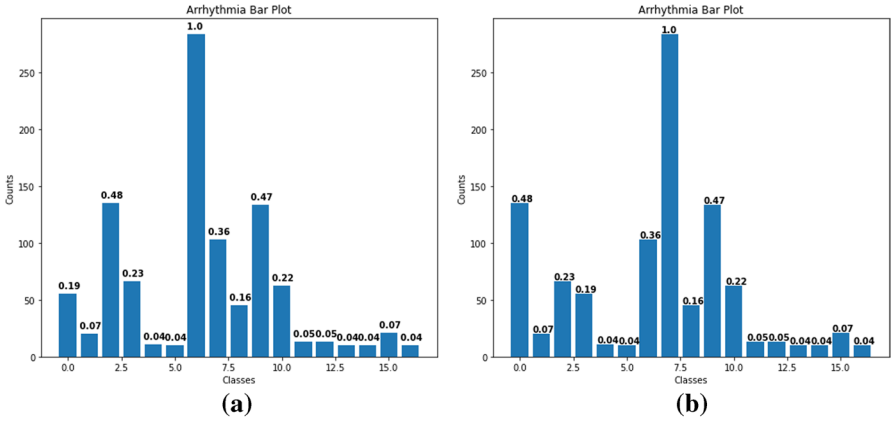


Fig. 3 Bias score for the classes of ECG signals in Dataset **a** (left) and Dataset **b** (right). Dataset **a** and Dataset **b** have 1–16 classes

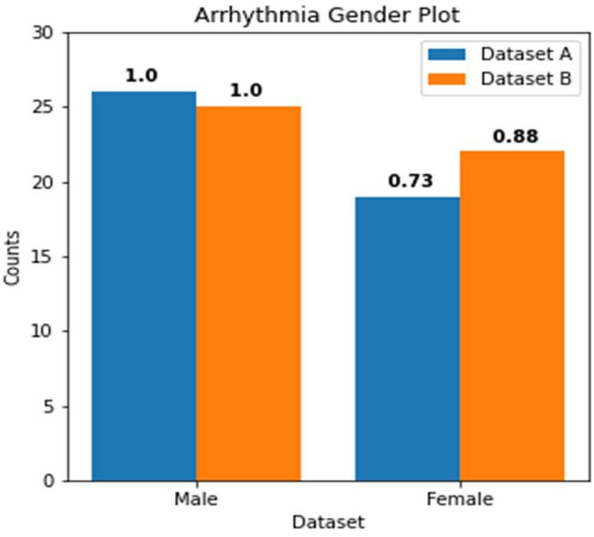


Fig. 4 Class ratio of the protected variable "Gender" for Dataset A and Dataset B. The dataset has a female population of 0.88 times the male population of the Dataset B development as none of the classes satisfies the 80% criteria thus having a disparate impact

- We observe from Fig. 3 that there is a class bias in the dataset with a relatively lower bias score of 0.04 for classes 4, 5, 13, and 14. The 0-indexed class 2 with the highest bias score of 0.48 also does not satisfy the optimal suggested ratio of 0.8 for disparate impact analysis.
- We observe that the protected group gender extracted from the dataset has a bias score of 0.73 as shown in Fig. 4.
- The protected group age is mentioned to be in the range of 23–89 years without the age of each patient and cannot be investigated for disparate impact towards age.
- The lack of meta-information about the protected group's age, race, and measurement instruments renders the study and identification of bias in these groups infeasible.

5.2 ECG Dataset B

The Dataset B we analyze is the MIT BIH arrhythmia database [30]. The graphical interpretations from Dataset B can be summarized as follows:

- The results show a high-class imbalance with a majority of classes having a ratio below 0.05, and the highest ratio is 0.48. This significant class imbalance poses a fundamental problem in the model.
- Figure 4 shows the class ratio bias for the protected variable "Gender." With a bias score of 0.88, it satisfies the 80% rule.
- Furthermore, we investigate the existence of bias in the protected variable age as shown in Fig. 5. We bin the protected group age into 5 bins representing values in the range 1–100 with each bin of width 20 for uniform distribution. We are unable to analyze the disparate impact in age bins 1–20 with no data points.

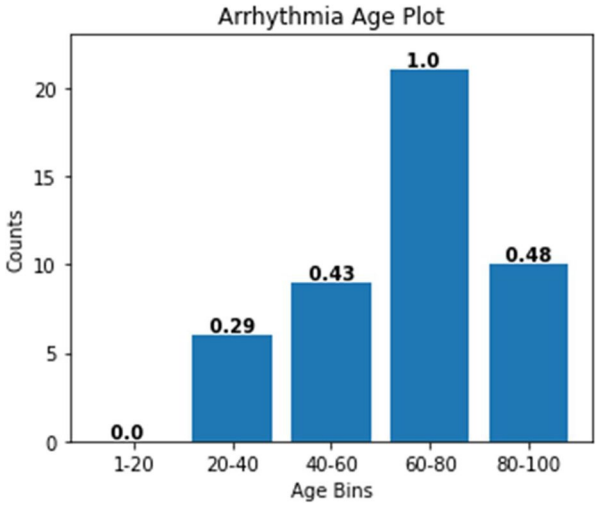


Fig. 5 Bias score for the protected variable "Age." The age of the dataset population is grouped into 5 bins of 20 years width for Dataset B

The bins 20–40 have a bias score of 0.29, 40–60 have a bias score of 0.43, and 80–100 have a bias score of 0.48. All the age bins with data samples fail to satisfy the 80% rule of disparate impact with respect to the bins 60–80 which have the majority of samples.

- We also identify age-wise gender group to analyze combined protected group bias (CPGB). The ECG signals were collected from 25 men and 22 women setting a ratio of 0.88 for the protected group gender which satisfies the 80% or 4/5th rule of disparate impact analysis. This problem multiplies if we look at our proposed CPGB for gender-wise age bins ignoring bins of 1–20 due to a lack of data samples. As shown in Fig. 6, while the CPGB for female have bias score of 0.71, 0.43, 1.0, and 1.0 for bins of 20–40, 40–60, 60–80, and 80–100 respectively. The disparate impact ratio gets worse for gender-wise male bins with a

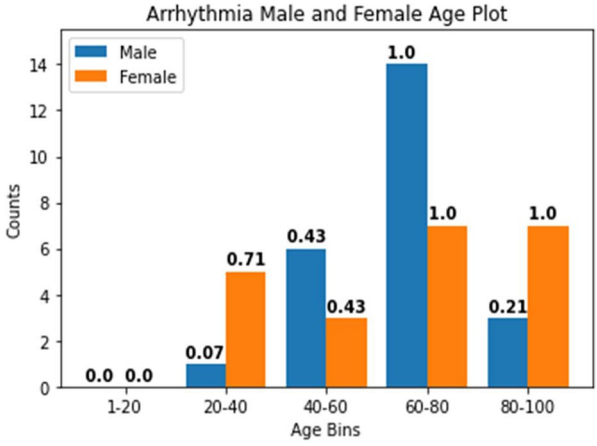


Fig. 6 Bias SCORE for the combined protected group "Female Age" and "Male Age." The age of the population is grouped into 5 bins of 20 years width for Dataset B

ratio of 0.29, 0.43, 1.0, and 0.48 respectively as shown in Fig. 5. The maximum observable disparate impact is on the males of the age bin of 20–40 which is 73% higher than that of female of the same age bin.

- The MIT BIH provides metadata that 9 measurement instruments used to record the ECG signals. Figure 7 shows the *measurement bias* score with a maximum and minimum of 0.62 and 0.08 respectively.

5.3 EEG Dataset C

This EEG dataset [31] has two groups namely alcoholic and control; we focus only on the alcoholic group for machine learning classifier purposes as classification between the types of predispositions to alcoholism is a more crucial and challenging task. Each participant is shown one stimulus or two stimuli, where either two same or different images were presented, giving us a total of 3 classes of annotation. This dataset highlights the importance of proper documentation for collection, annotation, and pre-processing for proper data bias analysis with no information about any protected groups. Figure 8 shows the class balance of the 3 classes of EEG classification as illustrated above. We observe that the classes are well balanced and satisfy the 4/5th rule of disparate impact. The lack of metadata is also a serious concern for widely used datasets to train deep learning models leaving a gap in the fairness analysis of decision support systems.

6 Bias Amplification Study

In this section, we demonstrate the presence of algorithmic bias by showing bias amplification [32]. Bias amplification, as the name suggests is the tendency of machine learning models to increase the bias present in the training ground truth data. This means that for a specific class, there is a significant difference in the distribution of protected variables in its output predictions versus the distribution

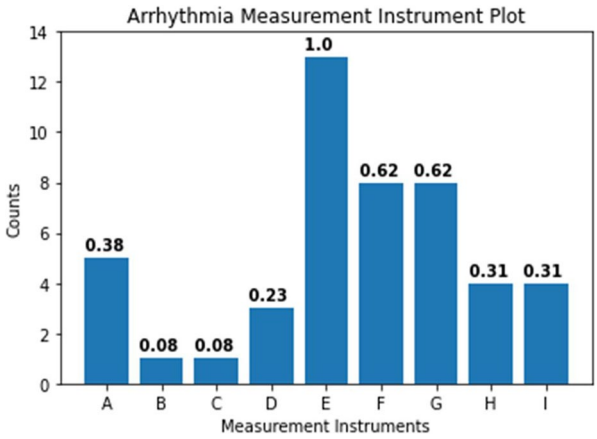


Fig. 7 Measurement bias score for the 9 separate measuring instruments for Dataset B

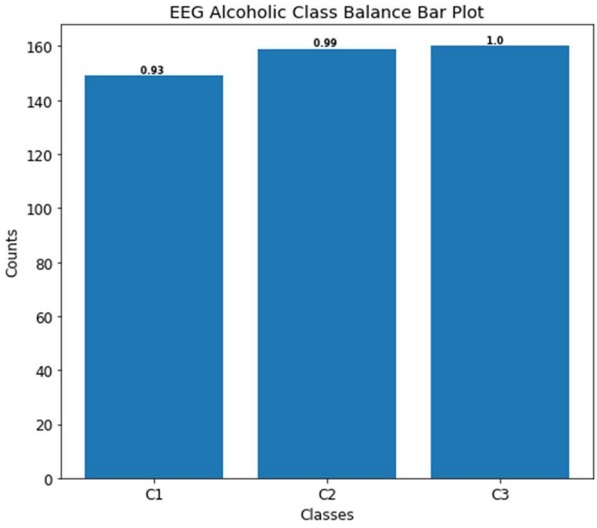


Fig. 8 Class ratio for the classes of EEG signals in Dataset C

in the input data. Class imbalance is one of the common reasons attributed to bias amplification, but it is not the only reason as some of the recent works have traced it to some specific features of the machine learning models [24].

Our proposed bias amplification measurement module can be used as a tool to calculate bias amplification on the predicted outputs of a model in a semi-automatic manner. On comparing group bias score (GBS) between the input test data (*test bias score*) and predicted output (*predicted bias score*) for each class, three conditions can be identified namely:

- Bias amplified
- Bias reduction
- Bias unchanged

These three conditions can be defined for comparison of bias per class as well as for comparison of two or more models where each class is treated as an individual group. The increase in bias score for a class denotes a reduction in bias amplification and consequently reduces disparate impact towards that class thus reducing bias. As illustrated above, a bias score greater than 0.8 satisfies the 80% criteria of participation and shows acceptable disparate impact. Consequently, a decrease in bias score leads to bias amplification. Furthermore, two deep learning models can be compared on the basis of the bias score for each of their classes. We identify the bias amplification, reduction, or no change between the two models using the three conditions illustrated below. The models being compared should be tested on the same test dataset. For each model, we compare the predicted bias and the test bias scores for each of the models under comparison. The bias amplification, reduction, and unchanged conditions for a class i, where GBS stands for group bias score, can be used to compare two models as follows:

Condition 1 (bias amplification):

$$\left(GBS_{Max}^{Test} - GBS_{i}^{Test}\right) < \left(GBS_{Max}^{Predicted} - GBS_{i}^{Predicted}\right)$$

Condition 2 (bias reduction):

$$\left(GBS_{Max}^{Test} - GBS_{i}^{Test}\right) > \left(GBS_{Max}^{Predicted} - GBS_{i}^{Predicted}\right)$$

Condition 3 (bias unchanged):

$$\left(GBS_{Max}^{Test} - GBS_{i}^{Test}\right) = \left(GBS_{Max}^{Predicted} - GBS_{i}^{Predicted}\right)$$

Of the three conditions, amplification of bias is most concerning as it negatively affects the performance of the model in the real world. The research on the effect of bias reduction and bias remaining unchanged is nascent and requires investigation in detail.

6.1 ECG Bias Amplification on Dataset B

In this section, we present the experiment and results for the identification of bias amplification with a 1D convolutional deep neural network on ECG Dataset B. We choose Dataset B only as ECG Dataset A does not have information to evaluate bias in depth.

6.1.1 Deep Learning Model and Training

Time series signals have both spatial and temporal features making it challenging for deep learning models. In this experiment, we used a 11-layer 1-dimensional CNN proposed by Maweu et al. [33] as shown in Fig. 9. We used a *Tanh* activation

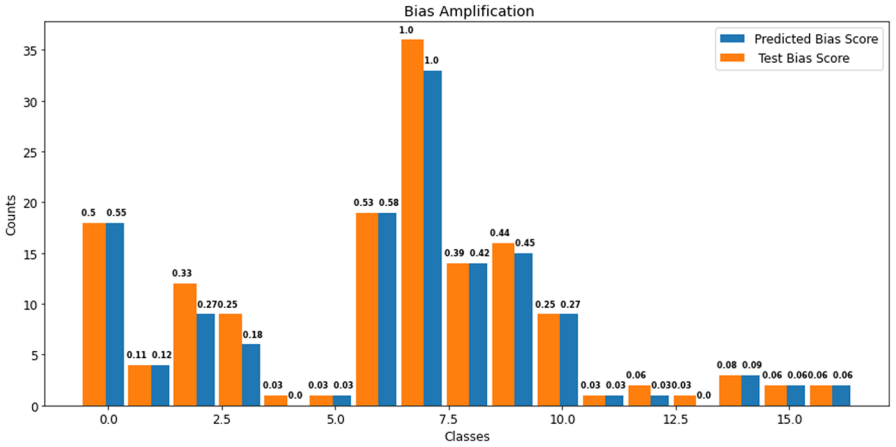


Fig. 9 Blue: Calculated Class Bias Score (Test Bias Score) of each class on Test Dataset B before prediction and Orange: Calculated Class Bias Score (Predicted Bias Score) of each class on predicted output of the model [33]

function for consistent superior results with ECG signals as demonstrated by Maweu et al. [33]. The Dataset B was split in a 80, 20 ratio to obtain a test set. Our model was trained for 200 epochs with an Adam optimizer, a learning rate of *0.0*1, and achieved a testing accuracy of 89.6%.

6.1.2 Amplification Results on Dataset B

In Fig. 10, we record the *test bias score* on the test data and the *predicted bias score* on the predicted output for each class. We observe that there is a significant change in the class bias scores for some of the classes on the predicted output in comparison to the input test data.

Table 1 reports the percentage of bias change for each of the classes based on the three conditions. We observed a maximum of 9.33% amplification in bias for class 3 and the maximum of 10.64% reduction in bias for class 6. The classes 0 and 6 which

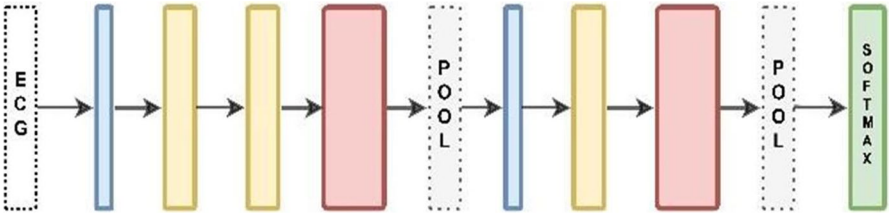


Fig. 10 Blue: calculated class bias score (test bias score) of each class on test Dataset B before prediction and orange: Calculated class bias score (predicted bias score) of each class on the predicted output of the model

Table 1 Categorizing each data class into groups representing bias amplification, reduction, or no change for interventional model improvement by expert. The values represent the percentage change in bias score for each class for Dataset B and "X" denotes the class without any samples

	Bias unchanged	Bias reduced	Bias amplified
Class 0		10%	
Class 1		1.12%	
Class 2			8.96%
Class 3			9.33%
Class 4			3.09%
Class 5	0%		
Class 6		10.64%	
Class 7	0%		
Class 8		4.92%	
Class 9		1.79%	
Class 10		2.67%	
Class 11	0%		
Class 12			3.19%
Class 13			3.09%
Class 14		1.09%	
Class 15	0%		
Class 16	0%		

had the highest bias scores showed the highest reduction in bias in prediction. Our experimental results highlight that for most of the classes, there can be bias amplification which requires mitigation at the source to reduce disparate impact. No direct relational mapping could be observed between the bias scores and the effect of bias amplification or reduction, and this warrants further investigation.

6.2 EEG Bias Amplification on Dataset C

In this section, we present results on our EEG Dataset C using 1D convolutional neural network.

6.2.1 Deep Learning on EEG Dataset C

We train a 105-layer neural network (Fig. 11a) with 3 building blocks. Each building block (Fig. 11b) is a combination of 1D convolution, batch normalization, and activation layers. The final layer is a three-node softmax function for classification. The "add" layer essentially acts as a residual connection between the results from

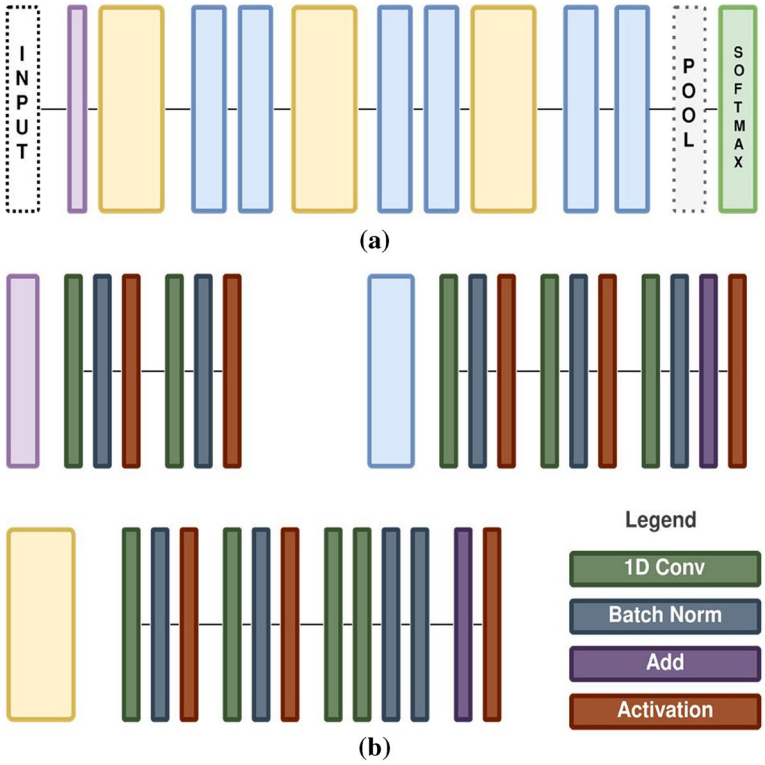


Fig. 11 **a** Neural network architecture of 105 layers with 3 different blocks. Each block is explained in Fig. 11b. **b** Neural network blocks illustrated in blue, yellow, and violet

batch normalization and the activation layer from earlier in the network. The available dataset is split into train and test sets into 80, 20 ratios. The model was trained for 100 epochs with a learning rate of 0.001 to achieve 92.3% testing accuracy. The training dataset analysis is illustrated in Section 5.3 in this section, we show bias amplification of our trained model on the test dataset.

6.2.2 Amplification Results on Dataset C

The *test bias score* and *predicted bias score* results for each class of our experiment are presented in Fig. 12. We observe that there is significant bias amplification of 28%, 66.66% for classes 1 and 2, even though both the classes have the class bias score close to the 80% rule of disparate impact. This observed effect of bias amplification demands the need for improved bias mitigation strategies and their evaluation.

7 Deep Learning Model Drift

To further bolster the practical significance of bias in real-world machine learning tasks, we recreate a focused experiment with 4 classes of ECG signals from Dataset B [30]. S. Dakshit et al. in their work [34] proposed an algorithm for selecting a subset of data to train deep networks efficiently using global explanations for ECG time series signals. Their work showed the possibility of achieving better model performance using a selected subset of incoming data samples. We replicate the focused experiment to verify the correlation between non-selected data samples to the bias ratio of the classes. This experiment demonstrates the problem of exclusion bias. Figure 13 shows the class bias ratios of the four classes namely normal sinus rhythm, periventricular contractions (PVC), left bundle bunch block (LBBB), and atrial fibrillation (AFIB).

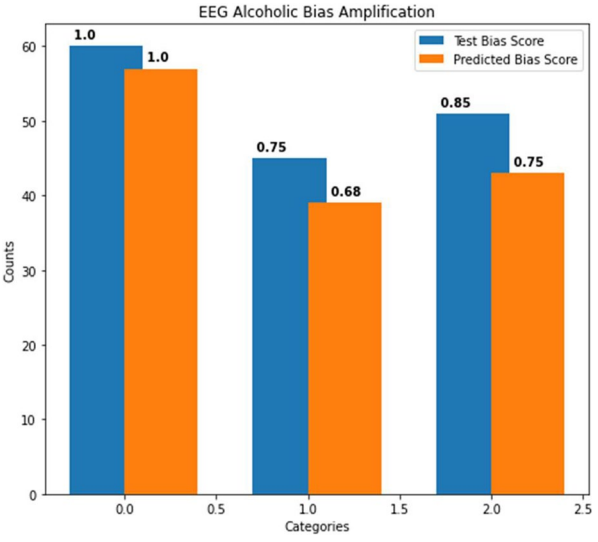


Fig. 12 Blue: calculated test_bias_score of each class on test Dataset C before prediction, orange: predicted_bias_score on the test data of Dataset C

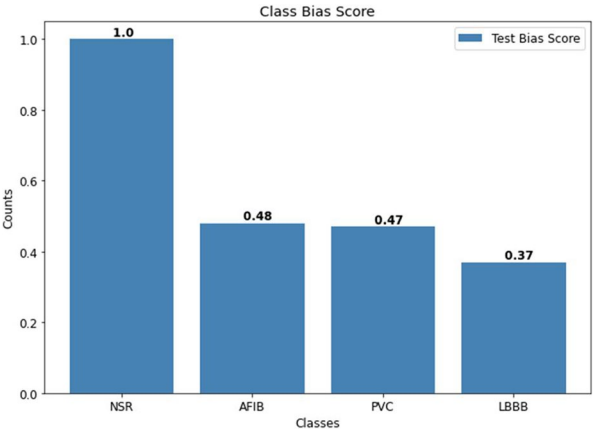


Fig. 13 Blue: class bias ratio in subset of Dataset B used for our focused experiment on 4 classes of ECG [24]; 0: NSR, 1: AFIB, 2: PVC, 3: LBBB

We observe that there is a significant class imbalance with classes 2 and 4 having the highest and lowest number of samples respectively. We ran simulations with hyperparameter budget values of 50% and 98% based on the author's published suggestions. We observe from Table 2 that for the set budget values, the majority of the samples selected were from classes 2 and 3 which have the highest-class bias score. For class 4, which has the lowest *class bias score*, we observe that 0% of the samples were selected in the core set for both budget values. Similarly, we observe only 12.7% selected from class 1 for a budget of 98% and 0% selected with a budget of 50%. The results show that the selection of samples is skewed with lower samples selected from classes of highest and lowest *bias score*. This mapping between the class bias and sample selection demonstrates the importance of accounting for bias in deep learning model development using time series healthcare data. A model trained with the samples removed without the bias analysis would lead to a loss of generalization. In our focused experiment using an established core-set selection strategy to improve model performance, we observe that although the model performance increases, a significant bias is introduced for the classes 1 and 4 leading to model drift in the real world from losing generalization to new data samples.

8 Bias Mitigation Strategies

We have discussed the problem of bias in general and in ECG and EEG healthcare signals specifically. Our framework allows for the identification of biases in prevalent public datasets. In this section, we discuss the most prevalent bias mitigation

Table 2 Categorizing each data class into groups based on a percentage of selected and not-selected samples for core-set from Dataset B

Budget		Class 1	Class 2	Class 3	Class 4
Budget 50%	Selected	0%	72.7%	25.5%	0%
	Not Selected	100%	27.3%	74.5%	100%
Budget 98%	Selected	12.7%	100%	100%	0%
	Not Selected	87.3%	0%	0%	100%

strategies and their effect on various time series ECG classification works. Bias mitigation consequently helps mitigate the disparate impact and improves model performance on all groups.

In this section, we demonstrate bias mitigation using only ECG Dataset B. Due to a lack of metadata, the presence or absence of bias in the ECG Dataset B for protected variables could not be adequately established. Our considered EEG dataset C already satisfies the 80% rule of participation and is not the best candidate for demonstrating bias mitigation. In the development of intelligent decision support arrhythmia classifiers, ECG signals are prevalently used in both modalities of 1D time series [33, 35] and 2D images [36–39]. For a comprehensive evaluation of bias mitigation strategies on ECG datasets, we present results on both data modalities. We compare and discuss three primary bias mitigation strategies namely:

(1) Synthetic data augmentation: We experiment with the bias mitigation strategy of augmenting the dataset with synthetic data on a 1D time series.
(2) Under-sampling: We experiment with the bias mitigation strategy of under-sampling the dataset on 2D images
(3) Oversampling: We experiment with the bias mitigation strategy of oversampling the dataset on 2D images.

8.1 Experiment A: Bias Mitigation on Time Series Data

In this experiment, we study the effect of balancing the time series classes by augmenting them with synthetic data. The recent success of generative models [35] on 1D time series ECG, over 2D ECG image generative models, motivates our choice of using time series signals for the demonstration of synthetic data augmentation as a bias mitigation strategy. We replicate the deep learning architecture proposed by Maweu et al. [33] as shown in Fig. 14 to record the bias score of each class on the models trained on imbalanced and synthetic data balanced data. As shown in Figs. 3 and 4, all the classes of datasets B have a ratio of less than 0.5. This implies

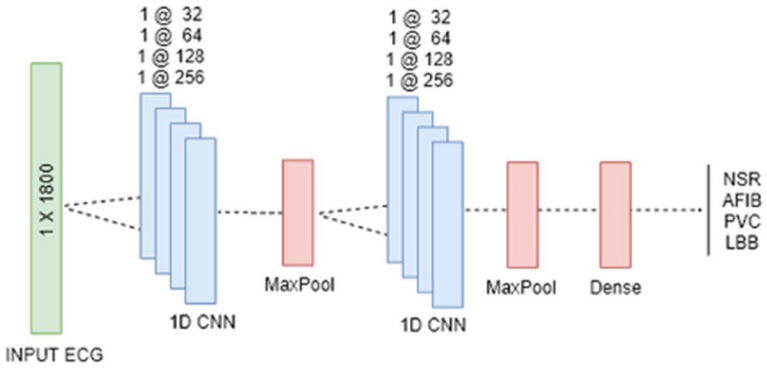


Fig. 14 Convolutional neural network architecture with two blocks of 8 1D Conv layers of 32, 64, 128, and 256 filters intertwined with max-pooling layers. The last layer is a softmax over the four time series classes [33]

that if the lower end of the distribution is augmented with synthetic data, sixteen of the seventeen classes would have at least twice as many synthetic samples as real samples. This justifies against using synthetic data to augment all classes of ECG arrhythmias. To observe performance using synthetic data as an augmentation strategy, we create a classifier for the four classes with class bias ratios closest to 0.5, namely normal sinus rhythm (NSR), premature ventricular contraction (PVC), atrial fibrillation (AFIB), and left bundle branch block (LBBB).

The dataset is first split into separate training and testing set with 83 NSR, 40 AFIB, 39 PVC, and 30 LBB samples in the test set. We use the synthetic generator proposed by Maweu et al. [35] to generate synthetic ECG samples of classes PVC, LBBB, and AFIB to augment our training dataset only. We balance the classes with 198 samples for each class to observe the effect of balancing the data using synthetic data. We do not augment the testing dataset with synthetic data and test only on real samples for proper evaluation of performance in the real world. We train two models with the same hyperparameters and architecture on the imbalanced original and synthetic data to augment the training dataset to ensure no additional algorithmic bias is added allowing for a fair comparison.

In Fig. 15a and b, we present the confusion matrix of both the models to compare their performance on the common real test data. We observe that both the models have comparable performance with the augmented model performing better on class 0, while there is no change in performance for class 2 and a slight drop in performance for classes 1 and 3. In Fig. 16, we present the *test bias score* and the

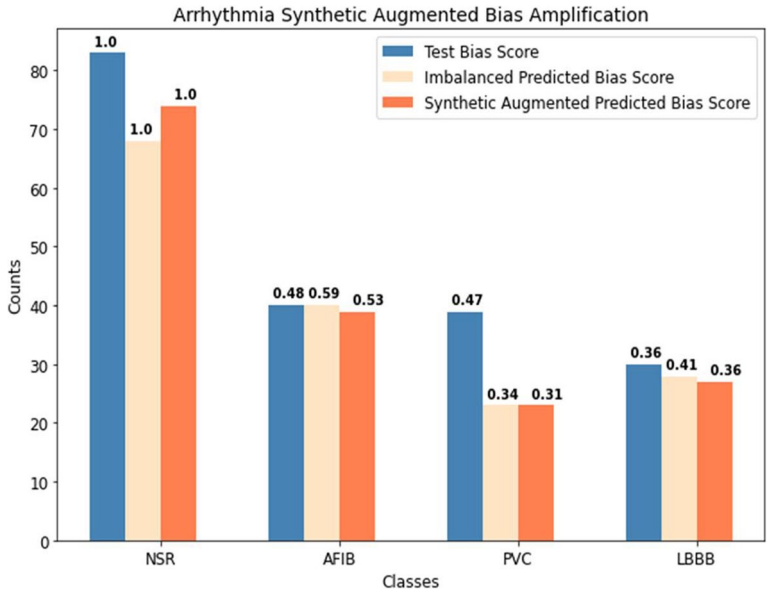


Fig. 15 The Test bias score and predicted bias score for four classes of time series 1D arrhythmia on the imbalanced model, and model balanced with 4 classes NSR, AFIB, PVC, and LBB augmented with time series synthetic data

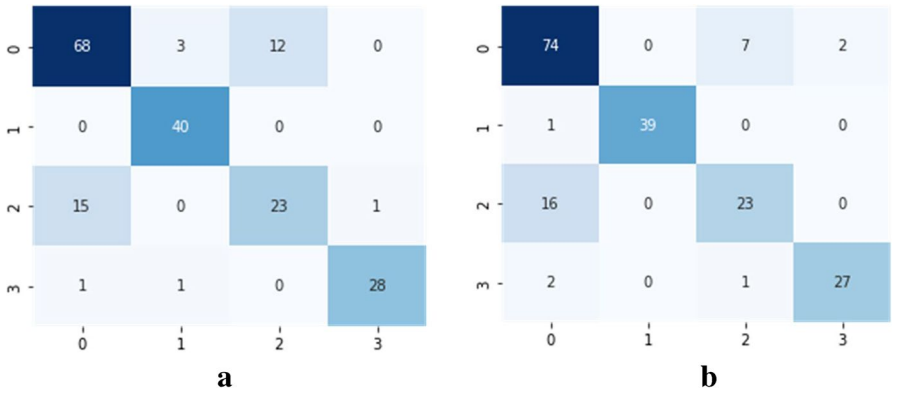


Fig. 16 **a** Imbalanced signal classification confusion matrix over 192 test samples; Class 0: NSR, Class 1: AFIB, Class 2: PVC, Class 3: LBB. **b** Synthetic data augmentation confusion matrix over 192 test samples; Class 0: NSR, Class 1: AFIB, Class 2: PVC, Class 3: LBB

predicted bias score on the test data and the predicted output for both the imbalanced and synthetic data–augmented model. We observe that there is a bias reduction in the model trained on the augmented synthetic data. We observe for the imbalanced model, the predicted bias scores show a 21.1% reduction in bias for Class 1 (AFIB), 7.8% for Class 3 (LBB), and an amplification of bias of 24.5% for Class 2 (PVC). In comparison to the test data bias scores, the model trained on balanced data shows a 9.6% bias reduction for class 1, 30.1% amplification for class 2, and no change in bias for class 3. The results show that the predicted bias for both the imbalanced model and the synthetic augmented model performs similarly by reducing bias for classes 1 and 3 but differently for class 2. This warrants further experiments on synthetic data as augmentation method with a larger dataset and across multiple models, which are beyond the scope of this paper.

8.2 Bias Mitigation 2D ECG Images

In this section, we observe the effect of under-sampling (experiment B) and over-sampling (experiment C) on 2D image modality ECG. Our choice of 2D image modality and the use of ECG beats instead of long rhythm signals for the demonstration of under-sampling and over-sampling is motivated by the skewed distribution of our dataset. As shown in Fig. 3, 2 classes, namely NSR and AFIB, have four classes with a ratio less than 0.5, a ratio close to 0.5, studying under-sampling on Dataset B, and Fig. 4, only motivating their selection for under-sampling such as a small number of samples per class. For a fair comparison of over-sampling, and under-sampling strategies, we use the same data modality of 2D images and the same deep learning architecture; hyperparameters are shown in Fig. 17 to classify the two classes of NSR and AFIB. Following the works [36–39], we preprocess the ECG by splitting the 1D time series signals into individual beats and creating their 2D image plots.

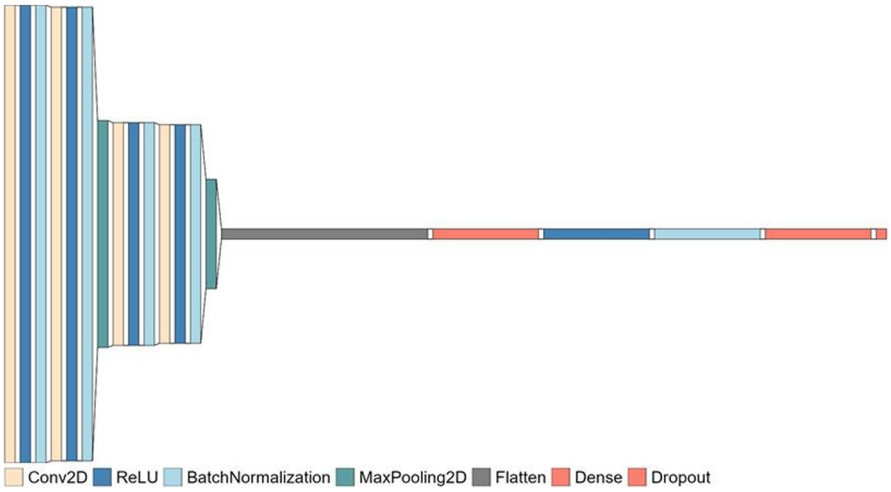


Fig. 17 Convolutional neural network used for 2D image under-sampling and oversampling with blocks of 1D Conv layers of 32 filters with ReLU activation and batch normalization intertwined with max-pooling layers

We use an Engzee segmenter to identify the R peaks to split each ECG signal into its corresponding beats. The processed beat dataset results in 574 NSR to 1398 AFIB samples. Since we demonstrate both under-sampling and oversampling, for a fair comparison, we choose to use the same dataset. The dataset is split into 80, 20 ratios giving us a total of 142 NSR and 352 AFIB testing samples. We keep this testing dataset fixed for both our following experiments B and C. The splitting of the time series signals into individual beats increases the sample counts of NSR and AFIB allowing us to compare under-sampling and over-sampling strategies.

8.2.1 Experiment B: Under-Sampling Strategy

In this experiment, we study the effect of under-sampling the data as a bias mitigation strategy. Under-sampling allows us to balance the number of samples for each data class by randomly selecting data subsets for each class. We under-sample the dataset to balance the classes with 574 samples each and use the derived test set as explained above. In Fig. 18a and b, we present the confusion matrix of both the models to compare their performance on the common test data. We observe that the under-sampled model's performance deteriorates significantly for class 0 (AFIB) but improves significantly for class 1 (NSR). This can be explained by the test and predicted bias ratios presented in Fig. 19. We observe that for both the imbalanced model, there is a 35.59% bias amplification on the predicted output.

For the under-sampled model, we observe a 37.3% bias reduction. Comparing the models in terms of bias, we can clearly see that the under-sampled model has a 72.89% less disparate impact. This shows that though the bias mitigation strategy of under-sampling yields mixed results with a decrease in performance for one class

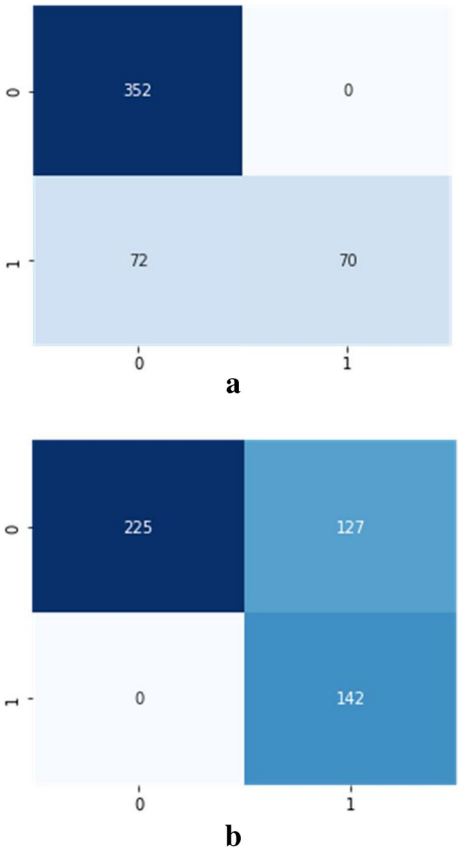


Fig. 18 **a** Imbalanced 2D Image Beats Model Confusion Matrix on the 494 test samples; Class 0: NSR, Class 1: AFIB. **b** Under-sampled 2D Image Beats Model Confusion Matrix on the 494 test samples; Class 0: NSR, Class 1: AFIB

and an increase in performance of the other, it removes disparate impact which is crucial for deployable healthcare decision support systems.

8.2.2 Experiment C: Over-Sampling Strategy

In this experiment, we focus on studying the effect of bias amplification on over-sampling to balance classes as a bias mitigation strategy. We use the same classes of NSR and AFIB, with 1398 samples each after over-sampling NSR. We present results for random oversampling as well as synthetic minority oversampling technique (SMOTE) on the test data. Random over-sampling is an oversampling method where the minority class is over-sampled by picking samples at random with replacement and adding them to the dataset. SMOTE is an improved method of dealing with imbalanced classes in data. In Fig. 20a and b, we present the confusion matrix of our model on the 2D image test data of 494 samples (142 NSR and 352 AFIB).

In comparison to the imbalanced model (Fig. 18a), we observe both oversampling strategies improve the performance of Class 1 (NSR) significantly and deteriorate the performance of Class 0 (AFIB). These results can be explained by the test bias score

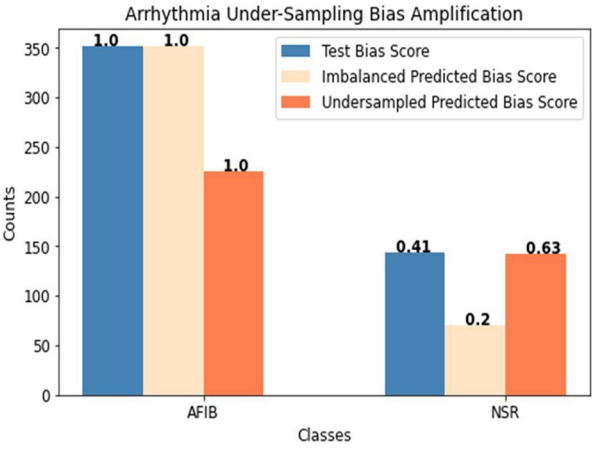


Fig. 19 The common test bias score and predicted bias score for the two classes of NSR and AFIB arrhythmia on the imbalanced model, and model balanced with under-sampling

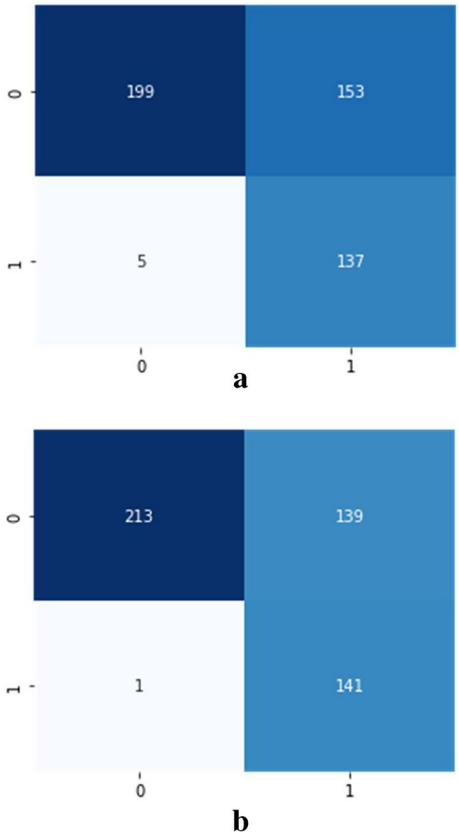


Fig. 20 **a** Random 2D Image Beats Model Confusion Matrix on the 494 test samples; Class 0: AFIB, Class 1: NSR. **b** SMOTE Oversampled 2D Image Beats Model Confusion Matrix on the 494 test samples; Class 0: AFIB, Class 1: NSR

and the predicted bias scores as shown in Figs. 21 and 22 for random and SMOTE oversampling respectively. On oversampling for both the strategies, the test bias score is significantly improved in balancing the training data. For the random over-sampling strategy, we observe that there is a bias reduction of 10% in comparison to the 35.59%

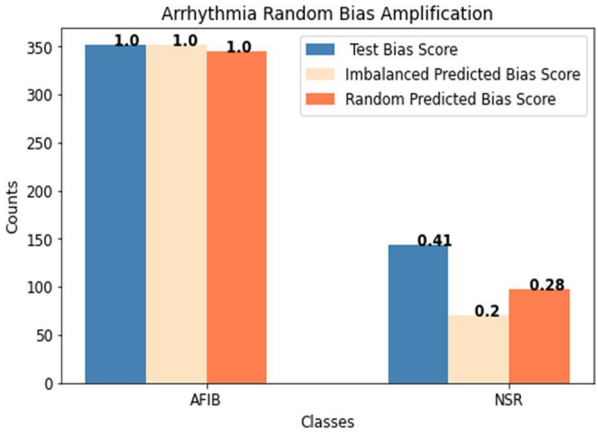


Fig. 21 The common test bias score and predicted bias score for the two classes of NSR and AFIB arrhythmia on the imbalanced model, and model balanced with random oversampling

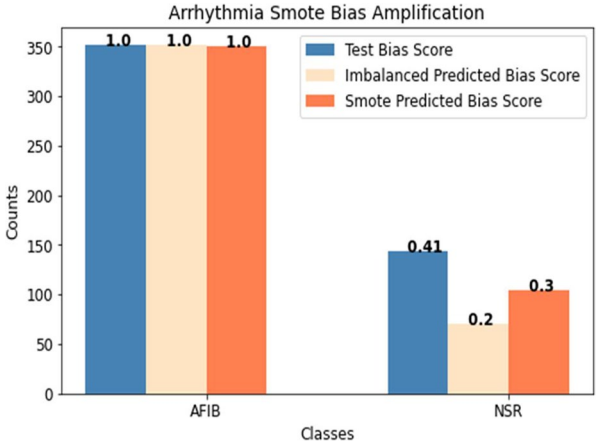


Fig. 22 The common test bias score and predicted bias score for the two classes of NSR and AFIB arrhythmia on the imbalanced model, and model balanced with SMOTE oversampling

bias amplification on developing models with an imbalanced dataset. Similarly, for SMOTE over-sampling, we observe a 12.5% bias reduction. These results confirm that both over-sampling strategies considered help reduce dataset bias and prevent the addition of algorithmic bias consequently reducing the disparate impact towards any class. In comparison to the model trained on imbalanced data, which amplified bias by 37.3% the random and SMOTE strategy is better by 47.3% and 49.8% respectively.

9 Conclusion

Bias analysis is crucial for applications that have a social impact such as healthcare decision support systems. Bias prejudices intelligent healthcare decision support systems impacting their quality of decisions and performance in the real world. As observed in our literature survey, bias analysis methods studies for time series healthcare signals are limited, with a focus on bias studies on 2D images such as

X-ray, MRI, and 3D scans such as OCT. We propose the use of the metadata of ECG and EEG healthcare time series datasets to analyze and identify diverse types of biases through BAHT. The identification of the presence of bias allows mitigation of the same at the source before data sharing. This is crucial as metadata needs to be obfuscated for privacy before sharing the dataset hindering the bias analysis and interpretation in later stages. BAHT has two modules allowing the graphical interpretation and analysis of bias, and measurement of the bias amplification for supervised algorithms. The bias amplification can satisfy one of the three identified conditions of bias reduction (negative bias amplification), unchanged bias condition, or a positive bias amplification. On analyzing three prominent time series ECG and EEG datasets, we were able to establish the presence of extensive bias and combined protected group bias (CPGB) in protected groups. These existing biases affect the model learning and capacity, and their analysis can potentially aid strategize model performance improvement by reducing disparate impact towards any group. To observe the effect of sample bias, we present results for under-sampling, over-sampling, and synthetic data augmentation as strategies for bias mitigation. We observe that though the performance results are mixed, with improvement for some classes and degradation for some classes, all the methods lead to a model bias reduction. None of the methods improves the bias score to greater than 0.8, to satisfy the 80% rule of disparate impact but a significant improvement of 72.89% in the bias score is observed.

While BAHT allows for the interpretation of bias graphically independent of the learning paradigm, our bias amplification module has been tested only on only supervised algorithm. Our experimental results demonstrate the presence of bias and highlight the importance of metadata for the analysis of bias in datasets. Through our bias amplification module, we demonstrate the effect of bias amplification on small time series healthcare datasets and consolidate if any bias mitigation strategy is superior to others, further experiments on larger datasets are required.

In future work, we want to develop a framework for de-biasing the time series healthcare datasets and further experiments on bias mitigation strategies to ensure fair model hyperparameters and test data constant. We also plan to conduct tests on unsupervised algorithms to quantify bias amplification. Bias analysis is crucial for applications that have a social impact such as healthcare decision support systems.

Author Contribution Sagnik Dakshit (primary author) has been responsible for designing the framework and experiments, writing manuscripts, conducting experiments, and creating figures. Sristi Dakshit (second author) and Ninad Khargonkar (third author) have been responsible for conducting experiments, creating figures, and reviewing the paper. Dr. Balakrishnan Prabhakaran (fourth author) has been responsible for reviewing the paper and helping design the framework and structuring the experiments and manuscript.

Data Availability All the data used are open source and can be accessed through the citation links. We will also make the code for our framework on acceptance.

Declarations

Ethical Approval For our novel work in BAHT framework, an ethical approval is not applicable.

Consent to Participate All the authors consent to participate.

Consent for Publication All the authors consent to publish.

Competing Interests The authors declare no competing interests.

References

1. Burlina P et al (2017) Comparing humans and deep learning performance for grading AMD: a study in using universal deep features and transfer learning for automated AMD analysis. Comput Biol Med 82:80–86
2. Oneto L, Silvia C (2020) "Fairness in machine learning." Recent trends in learning from data: tutorials from the inns big data and deep learning conference (innsbddl2019). Springer International Publishing
3. Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In Conference Fairness, Account Trans pages 77–91. PMLR
4. Álvarez-Rodríguez L et al (2022) Does imbalance in chest X-ray datasets produce biased deep learning approaches for COVID-19 screening? BMC Med Res Methodol 221:125
5. Cruz S, Garcia B et al (2021) Public covid-19 x-ray datasets and their impact on model bias–a systematic review of a significant problem. Med Image Anal 74:102225
6. Hague DC (2019) Benefits, pitfalls, and potential bias in health care AI. N C Med J 80(4):219–223
7. Bower JK et al (2017) Addressing bias in electronic health record-based surveillance of cardiovascular disease risk: finding the signal through the noise. Curr Epidemiol Rep 4:346–352
8. Rozier MD, Patel KK, Cross DA (2022) Electronic health records as biased tools or tools against bias: a conceptual model. Milbank Quarter 1001:134–150
9. Bhanot K et al (2021) The problem of fairness in synthetic healthcare data. Entropy 239:1165
10. Zhou Y, Huang S-C, Fries JA, Youssef A, Amrhein TJ, Chang M, Banerjee I et al (2021) "Radfusion: benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr." arXiv preprint arXiv:2111.11665
11. Hague DC (2019) Benefits, pitfalls, and potential bias in health care AI. North Carolina Med J 80(4):219–223
12. Torralba A, Efros AA (2011) "Unbiased look at dataset bias." CVPR 2011. IEEE
13. Hundman K, Gowda T, Kejriwal M, Boecking B (2018) "Always lurking: understanding and mitigating bias in online human trafficking detection." In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 137–143
14. Vasconcelos M, Carlos C, and Bernardo G (2018) "Modeling epistemological principles for bias mitigation in AI systems: an illustration in hiring decisions." Proceed AAAI/ACM Conference on AI, Ethics, Soc
15. Dixon L, Li J, Sorensen J, Thain N, Vasserman L (2018) "Measuring and mitigating unintended bias in text classification." In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 67–73
16. Gurupur V, Wan TTH (2020) "Inherent bias in artificial intelligence-based decision support systems for healthcare." Medicina 56(3):141
17. PPuyol-Antón E, Ruijsink B, Piechnik SK, Neubauer S, Petersen SE, Razavi R, King AP (2021) "Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning based segmentation." In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, pp. 413–423. Springer International Publishing
18. Duprez DA, Jacobs Jr DR, Lutsey PL, Herrington D, Prime D, Ouyang P, Barr RG, Bluemke DA (2009) "Race/ethnic and sex differences in large and small artery elasticity–results of the multi-ethnic study of atherosclerosis (MESA)." Ethnic Dis 19(3):243
19. Kishi S, Reis JP, Venkatesh BA, Gidding SS, Armstrong AC, Jacobs DR Jr, Sidney S, Wu CO, Cook NL, Lewis CE et al (2015) Race–ethnic and sex differences in left ventricular structure and function: the coronary artery risk development in young adults (cardia) study. J Am Heart Assoc 4(3):e001264

20. Moody GB, Mark RG (2001) "The impact of the MIT-BIH arrhythmia database." IEEE Eng Med Biol Mag 20(3):45–50
21. Bhanot K, Qi M, Erickson JS, Guyon I, Bennett KP (2021) The problem of fairness in synthetic healthcare data. Entropy 23(9):1165
22. Gu J, and Daniela O (2019) "Understanding bias in machine learning." arXiv preprint arXiv:1909.01866
23. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G (2018) "Potential biases in machine learning algorithms using electronic health record data." JAMA Int Med 178(11):1544–1547
24. Leino K, Fredrikson M, Black E, Sen S, and Datta A (2019) Feature-wise bias amplification. In Intl Conference Learn Represent (ICLR)
25. Kallus N, Zhou A (2018). Residual unfairness in fair machine learning from prejudiced data. Proceedings of the 35th International Conference on Machine Learning, in Proceedings of Machine Learning Research 80:2439–2448 Available from https://proceedings.mlr.press/v80/kallus18a.html
26. Protected Class: https://content.next.westlaw.com/Document/Ibb0a38daef0511e28578f7ccc38dcbee/View/FullText.html?transitionType=Default&contextData=(sc.Default)
27. Danks D, and London AJ (2017) "Algorithmic bias in autonomous systems." Ijcai. Vol. 17. No
28. Hall M et al (2022) "A systematic study of bias amplification." arXiv preprint arXiv:2201.11706
29. Pławiak P (2018) Novel methodology of cardiac health recognition based on ECG signals and evolutionary-neural system. Expert Syst Appl 92:334–349
30. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals." Circulation 101(23):e215–e220
31. Britton JW, Frey LC, Hopp JLet al (2016) authors; St. Louis EK, Frey LC, editors. Electroencephalography (EEG): an introductory text and atlas of normal and abnormal findings in adults, children, and infants [Internet]. Chicago: Am Epilepsy Soc Intro. Available from: https://www.ncbi.nlm.nih.gov/books/NBK390346/
32. Zhao J, Wang T, Yatskar M, Ordonez V, and Chang K-W (2017) Men also like shopping: reducing gender bias amplification using corpus-level constraints. Proceed Conference Empirical Methods Nat Language Process
33. Maweu BM, Dakshit S, Shamsuddin R, Prabhakaran B (2021) CEFEs: a CNN explainable framework for ECG signals. Artif Intell Med 115:102059
34. Dakshit S et al (2022) "Core-set selection using metrics-based explanations (CSUME) for multiclass ECG." IEEE Int Conference Healthcare Inform (ICHI). IEEE. (Also available at: arXiv:2205.14508)
35. Maweu BM et al (2021) Generating healthcare time series data for improving diagnostic accuracy of deep neural networks. IEEE Trans Instrument Measure 70:1–15
36. Dokur Z, Ölmez T (2001) ECG beat classification by a novel hybrid neural network. Comput Methods Programs Biomed 66(2–3):167–181
37. Nurmaini S, Partan RU, Caesarendra W, Dewi T, Rahmatullah MN, Darmawahyuni A, Bhayyu V, Firdaus F (2019) "An automated ECG beat classification system using deep neural networks with an unsupervised feature extraction technique." Appl Sci 9(14):2921
38. Martis RJ, Rajendra Acharya U, Min LC (2013) ECG beat classification using PCA, LDA, ICA and discrete wavelet transform. Biomed Signal Process Control 85:437–448
39. Yu S-N, Chou K-T (2008) Integration of independent component analysis and neural networks for ECG beat classification. Expert Syst Appl 34(4):2841–2846

Authors and Affiliations

Sagnik Dakshit[1] · Sristi Dakshit[1] · Ninad Khargonkar[1] · Balakrishnan Prabhakaran[1]

Sristi Dakshit
sxd200061@utdallas.edu

Ninad Khargonkar
NinadArun.Khargonkar@utdallas.edu

Balakrishnan Prabhakaran
bprabhakaran@utdallas.edu

[1] Computer Science, The University of Texas at Dallas, Dallas, USA