Algorithmic Fairness in Healthcare Data with Weighted Loss and Adversarial Learning

Pronaya Prosun Das[1], Marcel Mast[2], Lena Wiese[1,3], Thomas Jack[4], Antje Wulff[2], and ELISE STUDY GROUP[5]

[1] Fraunhofer Institute for Toxicology and Experimental Medicine, Hannover, Germany.
pronaya.prosun.das@item.fraunhofer.de
[2] Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover, Germany.
[3] Institute of Computer Science, Goethe University Frankfurt, Frankfurt a. M., Germany.
[4] Department of Pediatric Cardiology and Intensive Care Medicine, Hannover Medical School, Hannover, Germany.
[5] ELISE STUDY GROUP: Louisa Bode [a]; Marcel Mast [a]; Antje Wulff [a, d]; Michael Marschollek [a]; Sven Schamer [b]; Henning Rathert [b]; Thomas Jack [b]; Philipp Beerbaum [b]; Nicole Rübsamen [c]; Julia Böhnke [c]; André Karch [c]; Pronaya Prosun Das [e]; Lena Wiese [e]; Christian Groszweski-Anders [f]; Andreas Haller [f]; Torsten Frank [f]
[a]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover, Germany.
[b]Department of Pediatric Cardiology and Intensive Care Medicine, Hannover Medical School, Hannover, Germany.
[c]Institute of Epidemiology and Social Medicine, University of Muenster, Muenster, Germany.
[d]Big Data in Medicine, Department of Health Services Research, School of Medicine and Health Sciences, Carl von Ossietzky University Oldenburg, Oldenburg, Germany.
[e]Research Group Bioinformatics, Fraunhofer Institute for Toxicology and Experimental Medicine, Hannover, Germany.
[f]medisite GmbH, Hannover, Germany.

Abstract. Fairness in terms of various sensitive or protected attributes such as race, gender, age group, etc. has been a subject of great importance in the healthcare domain. Group fairness is considered as one of the principal criteria. However, most of the prevailing mitigation techniques emphasize on tuning the training algorithms while overlooking the fact that the training data may possibly be the primary reason for the biased outcomes. In this work, we address two sensitive attributes (age group and gender) with empirical evaluations of systemic inflammatory response syndrome (SIRS) classification for a dataset extracted from electronic health records (EHRs) for the essential task of improving equity in outcomes. Machine learning (ML)-based technologies are progressively becoming prevalent in hospitals; therefore, our approach carries out a demand for the frameworks to consider performance trade-offs regarding sensitive patient attributes combined with model training and permit organizations to utilize their ML resources in manners that are

aware of potential fairness and equity issues. With the intended purpose of fairness, we experiment with a number of strategies to reduce disparities in algorithmic performance with respect to gender and age group. We leverage a sample and label balancing technique using weighted loss along with adversarial learning for an observational cohort derived from EHRs to introduce a "fair" SIRS classification model with minimized discrepancy in error rates over different groups. We experimentally illustrate that our strategy has the ability to align the distribution of SIRS classification outcomes for the models constructed from high-dimensional EHR data across a number of groups simultaneously.

Keywords: Neural networks, adversarial Learning, fairness, bias, SIRS, healthcare, EHR

1 Introduction

Machine learning (ML) can be utilized to identify statistical patterns from the data that is produced by thousands of physicians and millions of patients. Determining statistical patterns is important to train computers to carry out specific tasks with incredible efficiency from time to time, such as diagnosing eye diseases in diabetic patients, to the extent of an experienced and knowledgeable specialist [1]. However, historical data might contain patterns of disparities regarding health care. Therefore, these inequities can be perpetuated in a ML model which was trained on those data. It is a significant task to promote fairness in the healthcare domain. Therefore, the American Medical Association passed the policy recommendations to "promote the development of thoughtfully designed, high-quality, clinically validated health care AI (artificial or augmented intelligence, such as machine learning) that identifies and takes steps to address bias and avoids introducing or exacerbating health care disparities including when testing or deploying new AI tools on vulnerable populations" [2].

In this work, we have focused on disparity in gender and age groups for the diagnosis of Systemic inflammatory response syndrome (SIRS). It is defined as an excessive defense response of the body to a noxious stressor e.g., infection, acute inflammation, trauma, surgery, reperfusion, ischemia, etc. to localize and subsequently terminate the external or endogenous cause of the insult. Professionals are usually led by SIRS identification criteria which were proposed in 1992 [3]. The advancement from sepsis to septic shock can increase the mortality rate significantly. Study [4] showed a 28-day/in-hospital mortality in serious sepsis and septic shock of 10%-40% and 30%-60%, respectively. Fluid resuscitation and early treatment with antibiotics were highly correlated with a higher survival rate [5]. A number of studies [6–8] demonstrated the applicability of Machine Learning algorithms to predict the diagnosis of a disease. The patients in those studies were from different age groups and genders, and these categories were not equally distributed most of the time – a fact that will indeed have an impact on the classification result. Such misestimation causes considerable harm to SIRS diagnosis in a sense that incorrect classification can endanger patients as

a result of both over- or undertreatment leading to avoidable sepsis side effects or incidents from unwanted treatments, respectively. It is our belief that the future selection of patients can be benefited from a much better understanding of the various patient subcategories for specific treatments.

A significant amount of attention has been drawn to fairness and bias in ML and it has become a prominent area of research for the ML students, researchers, and industry professionals [9]. To ensure a impartial future for AI, the Ethics and AI communities have aspired to decrease biases in ML [1]. By utilizing data analytics, empirical research has been carried out to evaluate fairness with respect to race groups [8]; yet, a very small number of works have focused on enhancing fairness in healthcare from the AI viewpoint. This study proposes techniques intended for investigating the fairness pertaining to the SIRS classification model regarding gender and age groups. Our experimental results are based on the SIRS dataset provided by the Hanover Medical School. We suggest approaches for enhancing group fairness at the time of both data processing and model training stages despite retaining overall accuracy. Our experimental outcomes show that: (1) Adversarial learning is effective since most marginalized groups display more significant average enhancements compared to other groups across all evaluation metrics, (2) In general, our method produces high scores for fairness while causing only a slight decrease in the overall performance of the classification, and (3) different groups require different model strategies for optimal effectiveness.

The article is structured as follows. In the next Section 2, we provide the used definitions of fairness and a brief literature review on various related works. We describe our datasets and analysis in Section 3. Different strategies and weighted loss are depicted in Section 4. In Section 5, we demonstrate the proposed adversarial learning approach. Results analysis and discussion are provided in Section 6. Finally, we derive conclusions of the work in Section 7.

2 Related Works

2.1 Fair Prediction

In general, supervised learning can be utilized to approximate the conditional distribution $p(Y \mid X)$ for a function $f(X)$ where N samples $\{x_i, y_i, z_i\}_{i=1}^{N}$ are taken from a given distribution $p(X, Y, Z)$. Usually, a vector representation $X \in \mathcal{X} = R^m$ of the medical history is extracted from the Electronic Health Records (EHRs). A binary label $Y \in \mathcal{Y} = \{0, 1\}$ that represents the outcome observed in the EHR for patient i, is used to obtain the outcome. Sensitive attributes, for example, gender, race, or age, with k groups, is indicated by $Z \in \mathcal{Z} = \{0, ..., k-1\}$. To render a prediction $\hat{Y} \in \{0, 1\}$, the output of the learned function $f(X) \in [0, 1]$ is thresholded with a value T.

Demographic parity [10] is one of the popular metrics to evaluate the fairness of a classifier regarding a sensitive attribute Z. The demographic parity criterion assesses the independence between the prediction $\hat{Y}$ and Z, formalized as

$$p(\hat{Y} \mid Z = Z_i) = p(\hat{Y} \mid Z = Z_j) \forall Z_i, Z_j \in \mathcal{Z} \tag{1}$$

Nevertheless, optimizing a ML model for demographic parity is inadequate for the prediction of clinical risk or diagnosis, as it may prevent the model from contemplating pertinent clinical features affiliated with the outcome and the sensitive attribute. Therefore, it can reduce the overall performance of the model for all protected groups [9].

The equality of odds [11] is another metric for evaluating fairness where it specifies that, for the given true label Y, prediction $\hat{Y}$ is conditionally independent of Z. Equality of odds is formally defined as

$$p(\hat{Y} \mid Z = Z_i, Y = Y_k) = p(\hat{Y} \mid Z = Z_j, Y = Y_k) \quad \forall Z_i, Z_j \in \mathcal{Z}; Y_k \in \mathcal{Y} \tag{2}$$

The definition states that if it is possible to accomplish equality of odds, then both false negative rates (FNR) and false positive rates (FPR) will be equal for a certain threshold T over all pairs of protected groups specified by Z. Therefore, equality of odds is more suitable in a clinical background in contrast to demographic parity [11].

2.2 Reducing the Impact of Algorithmic Bias

There are various strategies that can be utilized to reduce algorithmic bias. These strategies can be designed and carried out in different phases of a usual ML pipeline: during the construction of the dataset, model training, and inference (i.e., prediction). Removing sensitive features from the training data during the dataset construction phase is a simple and uncomplicated solution. However, due to different feature-class correlations, prediction outcome inequity may still be maintained. Poor model performance can also be observed as a result of removing features directly [12]. Additional techniques to reduce biases during the data construction phase aim to address imbalanced data related to predicted class and group. Predicted group and class size can be balanced by updating the loss function in terms of re-weighting every label and designating distinct weights to training samples, respectively [13]. Nevertheless, even in the case of balanced training data, ML models might still learn correlated information regarding sensitive features like gender and race from the provided intermediate representations [14]. Correlated information relating to sensitive features can be removed from the intermediate representation which is fed as input for the predictive models by utilizing adversarial learning [15,16]. A predictor (classifier) and an adversarial network are trained concurrently during adversarial learning. The primary goal of a predictor is typically to ensure that the intermediate representations used by the model remain highly informative for the prediction task. In contrast, an adversarial network's purpose is to hinder the predictor's capability to anticipate sensitive features [17]. Thus, by eliminating the biased

information concerning the sensitive features, a fair representation of model input can be learned using adversarial learning. The mitigation of the bias in ML can also be carried out at the inference phase. The main concept is to detect and turn off the portions of the model, that have learned the sensitive features. Therefore it eliminates the correlation between model output and those sensitive features [17].

In this work, we will focus on data construction, and model training phases for the mitigation of bias.

2.3 Different Approaches for Mitigating Bias

A considerable amount of interest has been observed in healthcare [18,19] regarding the ethical implications of applying ML algorithms. However, comparatively, little work exists that represents the applicability of satisfying formal fairness constraints while developing risk prediction or classification models trained with the EHR's data. We have seen a number of adversarial learning-based approaches in the non-healthcare domains to satisfy fairness constraints, especially in the form of demographic parity. In the situation of image anonymization, one approach [20] showed that a predictive model can be substituted by an autoencoder and an adversarial component to accomplish demographic parity. The adversarial learning technique was further inspected with a gradient reversal objective [21] for the imbalanced data in terms of the sensitive attributes as well as the outcome. It was also demonstrated that a small amount of data is needed to train adversarial networks. Alternatively, the use of equality of odds was presented in another work [11] to deal with the limitations of demographic parity. In that work, post-processing techniques were developed to attain equality of odds for the fixed-threshold classifiers. Recently, equality of odds was achieved for an adversarial framework by giving the discriminator access to the outcome values [15].

Equality of odds and demographic parity are called group fairness criteria as they are mainly involved with evaluating quantities at a group level, generally recognized as sensitive attributes such as age, gender, ethnicity, etc. The reasoning and computation of these metrics are straightforward. However, during optimization, they might generate models which are biased towards certain subgroups over groups of sensitive attributes [22]. By utilizing the notion of individual fairness [10], it may be possible to handle these issues. In this metric, a model is assessed whether it generates similar outputs for similar types of individuals. Nevertheless, this notion has limited practical use, as the domain-specific similarity metric is needed to be developed to encode the preferred criteria of fairness. A recent work [23] has explored an alternative to both individual and group fairness with a technique where subgroups are discovered, for which the model is performing poorly, and subsequently improves the performance of the model for those subgroups. This approach is model oriented; hence it mainly relies on model tuning for mitigating the bias. Another related work in healthcare [24], examined the fairness of risk prediction models for the context of predicting the mortality of patients in intensive care units. They argued that it

is undesirable to carry out a trade-off between the performance of the model and fairness across sensitive attributes.

However, none of the works dealt with unbalanced groups and labels. Besides, we have also been able to train a generalized adversarial model that satisfies different fairness constraints which will be discussed in the upcoming sections. From the literature review, we have realized that the fairness of a model should be evaluated in the context of the data [24]. Therefore, it motivates us to build a fair model for the context of EHR data where unbalanced groups and labels will be addressed to make the data bias-free to some extent.

3 Dataset

For the purposes of this work, a dataset of routine data from the pediatric intensive care unit at the Hannover Medical School is utilized. The data, which was obtained from a previously published study [25], has been pseudonymized to protect patient confidentiality. The dataset includes information on 168 pediatric patients, including vital parameters such as temperature, heart rate, respiration rate, and results from laboratory tests, as well as information from medical devices such as cooling blankets, ventilators, and pacemakers. Each patient can be identified by a unique study number, which was generated during the pseudonymization process. The laboratory test results include counts of leukocytes, platelets, and neutrophils, as well as INR values derived from the prothrombin time. Each measurement has a corresponding timestamp, providing a temporal sequence of data. The age of the patients is also recorded, which is crucial for correct diagnosis in the context of pediatric intensive care, particularly for SIRS detection (Figure 1). Blood pressure values have been added to the existing parameters.

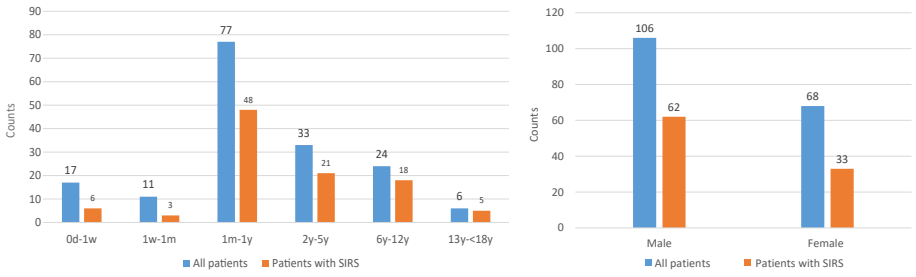

Fig. 1: Distribution of the cohort with classification according to IPSCC.

Along with the described dataset, there is also a gold standard for the existence of SIRS for the patients during the time period of the documented data, which was established by two experienced pediatric intensive care physicians.

These physicians evaluated the patients based on the SIRS diagnostic rules defined by the International Pediatric Sepsis Consensus Conference (IPSCC) [26]. The presence of SIRS was recorded each day that a patient was in the pediatric intensive care unit. Besides the day-based gold standard, the pediatricians also recorded the exact time of SIRS episodes, which provides an additional episode-based gold standard.

3.1 Descriptive Analysis by Age Groups and Gender

The dataset being used is comprised of records from 168 pediatric patients who were admitted to the department of pediatric cardiology and intensive care medicine of the Hannover Medical School over a total of 243 days, resulting in 1,998 days of hospital stay. Out of these 1,998 days, 460 were marked with a SIRS label in accordance with the day-wise gold standard. The age of the patients in the cohort is divided into groups based on the IPSCC criteria. The sex distribution of the patients shows that there were 106 male patients and 62 female patients.

Table 1: Number of observations for each sensitive attribute and label.

Number of observation	Count	%
Male	8710	59.04
Female	6042	40.96
SIRS	5680	38.5
No SIRS	9072	61.5
Newborn (0d-1w)	1969	13.35
Neonate (1w-1m)	2107	14.28
Infant (1m-1y)	5338	36.18
Toddler (2y-5y)	3458	23.44
School-aged (6y-12y)	1469	9.96
Adult (13y-<18y)	411	2.79

However, after compiling all the data, it is revealed that each patient has multiple observations even within a single day. This causes the data to be unbalanced in terms of the number of observations and SIRS labels for certain attributes. The actual number of observations can be found in the observation count table.

3.2 Feature Extraction

Our analysis involves working with a set of six health indicators, including temperature, respiration rate, pulse rate, systolic and diastolic pressures, and leukocyte count, which are chosen based on their relevance to the IPSCC criteria.

Additionally, we use information about the patient's birthdate, gender and disease diagnosis to further investigate the data. Age groups are derived from the birthdates. We extract these features from their corresponding datasets and split them into hourly observations. Afterwards, maximum, minimum, median and mean are calculated from the observations of each hour and created a new dataset. Therefore, the processed dataset contains 24 features (maximum, minimum, median and mean for temperature, pulse, respiration, systolic and diastolic pressures, and Leukocytes) for training with additional 4 dimensions for study number, timestamp, age groups and gender. Here, age groups and gender are sensitive features.

4 Approaches to Reduce Bias

A number of techniques will be used and adapted to alleviate algorithmic bias that have been mentioned in previous research, in the context of the sepsis classification task. A summary of all the techniques discussed in this section is presented in Table 2.

Table 2: A list of techniques that will be employed to reduce bias in the SIRS classification model.

Used techniques	Names
Weighted loss	wgLoss
Fairness through unawareness	default (loss)
Sensitive features (age group, gender) added to input	featureAdded
Adversarial learning with demographic parity	adv˙DP
Adversarial learning with equality of odds	adv˙EO

4.1 Classification using Artificial Neural Network

Artificial Neural Network (ANN) is supposed to be an effective tool to find an association between input and output data. A set of records that consist of input and interrelated output data are needed to train ANN for the accomplishment of this purpose. The typical architecture of ANN comprises three types of layers: (1) an input, (2) a hidden, and (3) an output. The neurons of the input and output layers are linked to the input and output vectors, respectively [27]. In contrast, neurons of the hidden layer are associated with the neurons of the input and output layers. These hidden layer neurons are mainly responsible for transforming the input data into the related output data. In addition, a transfer function is used to transfer a weighted summation of the input data. In this study,

a back propagation network with the Adam optimization algorithm was used to train ANN for the first three strategies: fairness through unawareness (default), weighted loss (wgLoss) and sensitive features (age groups, gender) added to input (featureAdded). The network consists of four layers with rectified linear unit (ReLU) activations and dropouts of 0.5 in between. The hidden layers consist of fixed 32 neurons each. All the implementation in this work is accomplished using Python, while the rest of the characteristics of ANN were set according to those implemented in the previous experiments [28, 29]. The default loss is used as follows,

$$\begin{aligned} Loss_{def} &= Loss_y = BinaryCrossEntropy(y_i, \hat{y}_i) \\ &= -\frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i) \end{aligned} \tag{3}$$

Where, N is the output size, y_i is the ground truth and $\hat{y}_i$ is the model output.

4.2 Data Construction and Weighted Loss

Bias may be introduced into the prediction model due to three main factors during the data construction phase. The first factor is associated with to sensitive patient attributes, where there may be an imbalance in the training samples, as demonstrated in Table 1. This can result in disparities in prediction quality, as outlined in Section 2. In order to address the problem of imbalanced data samples, we can adjust the loss function by assigning weights to training samples, which can help to deal with the under-representation of gender and age groups, and eventually, promote fairness during the data construction phase. The ANN loss function is modified for SIRS classification and defined as follows:

$$L_{wgGroups} = \alpha(g(y_i))\ \beta(a(y_i))\ BinaryCrossEntropy(\hat{y}_i, y_i) \tag{4}$$

where, $g(y_i)$ denotes the gender and $a(y_i)$ denotes the age group of the patient sample that have the SIRS label y_i, and $\alpha(g(y_i))$ and $\beta(a(y_i))$ assign the patient sample with the weights associated with their gender and age group respectively. Normally, the values of $\alpha(\star)$ and $\beta(\star)$ vary, but to ensure that the model learns from under-represented groups, it is necessary to assign smaller weights to the majority groups compared to the minority groups in the data. This is because the model tends to learn more from the group with larger weights than the group with smaller weights.

A second factor is the imbalanced label distribution across groups, which can introduce bias. SIRS label distributions of the patients indicate unevenness, also echoed in our dataset as depicted in Table 1, where somewhere around half of labels (38.5%) are in a positive category (SIRS). A model tends to exhibit bias towards the distribution of the groups that are the largest in size when disparities between group label distributions are in existence, deteriorating the fairness issue regarding SIRS classification. We can actually balance labels in

a similar manner to instance balancing by assigning different weights to the training samples according to their SIRS labels. This results in an adjusted loss function for the ANN classification, which can be expressed as:

$$L_{wgLabel} = -\frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \lambda(y_i)(y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i)) \tag{5}$$

Here, the function $\lambda(y_i)$ assigns a weight to each patient sample based on its diagnosis label. Both label-based and group representation-based instance balancing can be used. In that case, the weighting schemes are merged as follows:

$$L_{wgCombine} = -\frac{1}{\mathrm{N}} \alpha(g(y_i))\, \beta(a(y_i)) \cdot \sum_{i=1}^{\mathrm{N}} \lambda(y_i)(y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i)) \tag{6}$$

Third, it has been demonstrated that the "fairness through unawareness" strategy is ineffective because it does not fully conceal protected (sensitive) attributes which could be inferred from other relatively unrelated features [10]. As an alternative, it is recommended to identify sensitive patient attributes like gender and age group, and use more advanced modeling approaches to reduce any bias resulting from these attributes.

5 Adversarial Learning

5.1 Model Structure

We start with the ANN SIRS classification model, a multi-layer artificial neural network C which outputs a probability distribution, designated as $\hat{y}$, of SIRS for each patient in the intensive care unit. The objective in this situation is to make sure that the ANN can efficiently classify diagnoses while simultaneously exhibiting the highest level of ambiguity or uncertainty regarding the gender and age group of the patient. At this point, we require our output $\hat{y}$ to satisfy the constraints of demographic parity or equality of odds, separately, for different sensitive features (gender and age group). A network can learn a bias even if the sensitive features are not an input to our neural network, as those features may have some correlation with other features. The features that are used as input x are mentioned in Section 3.2. The diagram of our adversarial model is illustrated in Figure 2.

Demographic Parity Model: an adversarial neural network A takes the prediction $\hat{y}$ as input and learns to predict sensitive features, z_{Gender} and $z_{Age\ group}$. If our classification model exhibits bias against z_{Gender} and $z_{Age\ group}$, these sensitive attributes can be predicted from the value of $\hat{y}$. Consequently, A will lead to high classification accuracy.

Equality of Odds Model: an adversarial neural network called A takes the predicted value $\hat{y}$ and the true label y as input and learns to predict sensitive

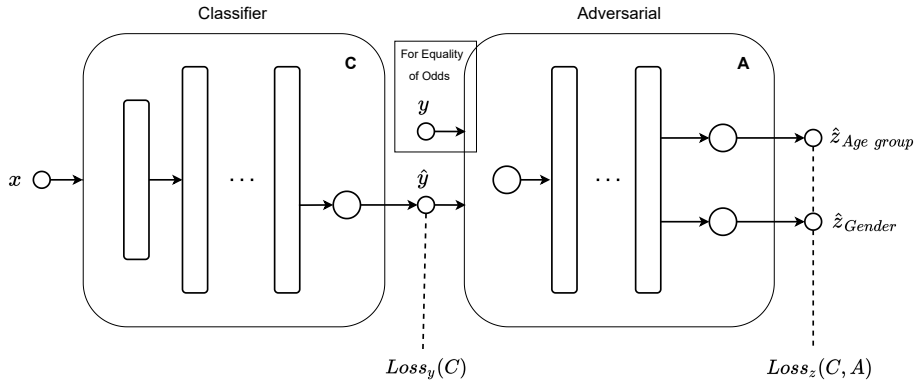

Fig. 2: Diagram of our adversarial model structure.

patient features, specifically z_{Gender} and $z_{Age\ group}$. When the predicted value $\hat{y}$ and the sensitive features z_{Gender} and $z_{Age\ group}$ are not conditionally independent given the true label y, the pair $(\hat{y}, y)$ can reveal the sensitive attributes, and A will have high accuracy in classifying them.

Both classifier C and adversarial A consist of 3 hidden layers with 32 neurons in each layer. Rectified linear unit (ReLU) activations and dropouts of 0.5 are used between hidden layers.

5.2 Model Training

The approach of adversarial learning has been utilized to find out unbiased representations trained out of data that contains inherent biases [15]. The principal idea should be to use deep representations that are highly informative for the primary task of prediction or classification, although being as minimally discriminative as possible with regard to predicting sensitive attributes [17]. Our objective is for classifier C to predict $\hat{y}$ correctly and for adversarial network A to predict age group $z_{Age\ group}$ and gender z_{Gender} poorly. For simplicity, we mention $z_{Age\ group}$ and z_{Gender} as z_a and z_g, respectively, in the equations. If it is possible to accomplish this, the model will output an accurate, unbiased $\hat{y}$. Particularly, binary cross-entropy losses are used for C and A, which are referred to as $Loss_y$ and $Loss_z$, respectively. $Loss_z$ represents the loss from $z_{Age\ group}$ and gender z_{Gender}. Here, we can treat gender as a binary feature and use a binary cross-entropy loss.

$$Loss_{z_g} = -\frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} z_{g_i} \cdot \log \hat{z}_{g_i} + (1 - z_{g_i}) \cdot \log (1 - \hat{z}_{g_i}) \tag{7}$$

Algorithm 1 Training procedure of Adversarial Network to mitigate biases.

Require: $x,\ z_*,\ y,\ \delta,\ \omega,\ \tau,\ K,\ L,\ N$
Ensure: *Fair* $\hat{y}$

Sample R data samples x_i, y_i, z_{*i} *where* $i = 1, ..., R$ ▷ Minibatch sampling

for K *epochs* **do** ▷ Pretrain classifier, C
  for $i \in R$ **do**
    $\nabla_C[Loss_y(C \mid x_i, y_i)]$ ▷ Update C using loss defined in equation (3)
  end for
end for

for L *epochs* **do** ▷ Pretrain adversarial network, A
  for $i \in R$ **do**
    $\hat{y} = C(x_i)$ ▷ Get the prediction $\hat{y}$ from C
    $\nabla_A[Loss_z(A \mid \hat{y}_i, y_i, z_{*i}, \delta, \omega)]$ ▷ Update A using loss defined in equation (9)
  end for
end for

for N *epochs* **do** ▷ Train adversarial and classifier networks concurrently.
  for $i \in R$ **do**
    $\hat{y}_i = C(x_i)$
    $\nabla_A[Loss_z(A \mid \hat{y}_i, y_i, z_{*i}, \delta, \omega)]$
  end for

  for *a random batch index* $r \in R$ **do**
    $\hat{y}_r = C(x_r)$
    $\hat{z}_r = A(\hat{y}_r, y_r)$
    $\nabla_C[\tau Loss_y(C \mid x_r, y_r) - Loss_z(A \mid \hat{y_r}, y_r, z_{*r}, \delta, \omega)]$ ▷ Update C using loss defined in equation (10)
  end for
end for

However, the age group is categorical, as it contains six different age groups. Therefore, Multi-Class binary cross entropy is suitable for $z_{Age\ group}$.

$$Loss_{z_a} = -\frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{\mathrm{M}} z_{a_{i,j}} \cdot \log \hat{z}_{a_{i,j}} \tag{8}$$

Now, after combining $Loss_{z_g}$ and $Loss_{z_a}$, we get average $Loss_z$ as follows,

$$\begin{aligned} Loss_z = -\frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} (\sigma \cdot (z_{g_i} \cdot \log \hat{z}_{g_i} + (1 - z_{g_i}) \cdot \log (1 - \hat{z}_{g_i})) \\ + \omega \sum_{j=1}^{\mathrm{M}} z_{a_{i,j}} \cdot \log \hat{z}_{a_{i,j}}) \end{aligned} \tag{9}$$

$Loss_z$ is back-propagate through A to train A. Still, we have to train C so that it becomes good at predicting $\hat{y}$ which is not highly correlated with $z_{Age\ group}$ and z_{Gender}. If we subtract $Loss_z$ from $Loss_y$, C will be instigated to maximize $Loss_z$, and as a result, it will produce a $\hat{y}$ which simply cannot be utilized to predict sensitive features. Therefore, $\hat{y}$ values will be closer to obtaining parity. However, it is essential to note that the model must also retain its capability to accurately classify SIRS. Thus, the weights of the losses should be adjusted accordingly to avoid poor SIRS classification performance. The overall loss function can be expressed as follows:

$$\begin{aligned} Loss = -\frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} (\tau \cdot (y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i)) \\ - (\sigma \cdot (z_{g_i} \cdot \log \hat{z}_{g_i} + (1 - z_{g_i}) \cdot \log (1 - \hat{z}_{g_i})) \\ + \omega \sum_{j=1}^{\mathrm{M}} z_{a_{i,j}} \cdot \log \hat{z}_{a_{i,j}})) \end{aligned} \tag{10}$$

where, τ, σ and ω are the coefficients that control the importance of the loss functions. N and M are the numbers of observations and groups for the protected feature age group. The overall adversarial training procedure is shown in Algorithm 1.

6 Result Analysis

We assess the effectiveness of the proposed techniques in ensuring model equity and fairness by measuring positive rate (PR), true negative rate (TNR), true positive rate (TPR), and accuracy (ACC) metrics. These metrics are used to report demographic parity and equity of odds as discussed in Section 2. Accuracy is a metric that determines the overall capability of a model to make accurate predictions. TPR and TNR reflect the probability of correctly identified SIRS

and NO SIRS patients, respectively, by the model. FPR is actually calculated mathematically from TNR using the following equation,

$$FPR = 1 - TNR \tag{11}$$

From the definition, we can say, demographic parity will be satisfied when all the groups have same PR. Similarly, the equality of odds will be satisfied if TPR and FPR are the same for all sensitive groups. Experiments were carried out using the datasets described in Section 3. The test, validation, and training data split in a 4:3:13 ratio. The reported metrics were calculated as averages over all the approaches mentioned in Table 2.

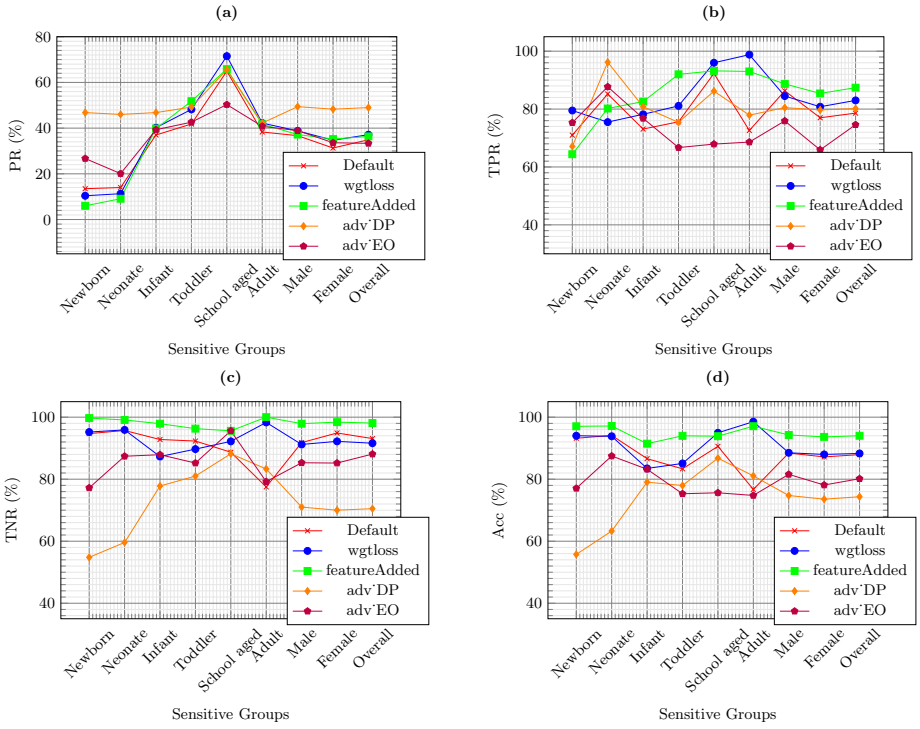

Fig. 3: (a) Positive Rate (PR) (%), (b) True Positive Rate (TPR) (%), (c) True Negative Rate (TNR) (%) and (d) Accuracy (Acc) (%) of the models.

6.1 The Effect of Adding Sensitive Attributes

To assess how age group and gender affect the fairness and predictive performance of the SIRS classification model, we initially incorporated these sensitive

Table 3: The performance evaluation of four fairness-based techniques and their comparison with the default strategy (no strategy) is presented in terms of PR, TPR, TNR, FPR, and ACC, for each sensitive group. Additionally, the standard deviation (STD) is also reported as a measure of group fairness.

		Newborn	Neonate	Infant	Toddler	School aged	Adult	Male	Female	Overall	STD
PR (%)	Default	13.5	14	37	41.8	64.9	38.3	36.7	31.3	34.91	15.31
	wgLoss	10.4	11.3	40.2	48.2	71.5	42.2	38.8	34.7	37.13	18.47
	featureAdded	6	9.1	39.8	51.8	65.8	41.7	37.2	35.2	36.37	18.74
	Adv˙DP	46.8	46	46.8	49.2	65.5	42.2	49.4	48.3	48.98	**6.49**
	Adv˙EO	26.7	20.1	39.3	42.6	50.3	40.8	38.9	33.6	33.34	**8.96**
TPR (%)	Default	71	85.2	73.1	75.6	92.3	72.6	86.3	77	78.6	7.28
	wgLoss	**79.5**	**75.5**	**78.2**	**81.1**	**96**	**98.8**	**84.5**	**80.8**	**83**	**7.98**
	featureAdded	64.4	80.2	82.6	92	93.2	93	88.7	85.4	87.4	9.03
	Adv˙DP	67.1	96.2	80.8	75.4	86.2	77.9	80.5	79.7	80.2	7.84
	Adv˙EO	75.3	87.7	76.8	66.7	67.9	68.6	75.9	65.9	74.6	**6.91**
TNR (%)	Default	94.7	95.7	92.8	92.3	88.7	77.5	91.8	94.9	93.1	5.57
	wgLoss	**95.2**	**95.9**	**87.3**	**89.7**	**92.2**	**98.3**	**91.2**	**92.2**	**91.6**	**3.35**
	featureAdded	99.7	99.1	97.9	96.3	95.6	100	97.9	98.4	98.1	1.45
	Adv˙DP	54.8	59.6	77.8	81	88.2	83.3	71	70	70.5	10.92
	Adv˙EO	**77.2**	**87.4**	**87.9**	**85.2**	**95.6**	**79.2**	**85.3**	**85.2**	**88.1**	**5.32**
FPR (%)	Default	5.3	4.3	7.2	7.7	11.3	22.5	8.2	5.1	6.9	5.57
	wgLoss	4.8	4.1	12.7	10.3	7.8	1.7	8.8	7.8	8.4	3.35
	featureAdded	0.3	0.9	2.1	3.7	4.4	0	2.1	1.6	1.9	1.45
	Adv˙DP	45.2	40.4	22.2	19	11.8	16.7	29	30	29.5	10.92
	Adv˙EO	22.8	12.6	12.1	14.8	4.4	20.8	14.7	14.8	11.9	5.32
ACC (%)	Default	93.20	94.12	86.62	83.34	90.60	76.70	88.43	87.19	87.92	5.26
	wgLoss	**94.01**	**93.83**	**83.47**	**85.08**	**94.96**	**98.54**	**88.52**	**87.98**	**88.30**	**5.02**
	featureAdded	97.06	97.15	91.45	93.99	93.87	97.09	94.26	93.64	94.01	1.96
	Adv˙DP	55.74	63.28	79.05	77.96	86.79	81.07	74.74	73.55	74.34	9.38
	Adv˙EO	**77.06**	**87.48**	**83.21**	**75.30**	**75.61**	**74.76**	**81.56**	**78.12**	**80.15**	**4.25**

attributes to the input by joining their one-hot representation. It is solely done for validation purposes. The results of the evaluation (Table 3 / Figure 3) indicate that incorporating sensitive attributes into the model input led to an improvement in classification accuracy for most of groups. This finding aligns with the prior research that the overall accuracy can be enhanced by including the sensitive features [30]. On the other hand, the model suffers from the increased discrimination in terms of TPR, TNR, and FPR. Specifically, adding age group and gender to the model input are likely to improve TPR for infant, toddler,

school-aged, adolescent, male and female, while decreasing TPR for Newborn and equal TPR for Neonate compared to TPRs produced by the default model.

On the contrary, the inclusion of sensitive attributes resulted in an increase in TNR for all groups. However, the opposite relationship between TNR and TPR was observed with the addition of more attributes to the model input highlighted the tradeoff and conflict between them. This phenomenon of significant differences in TNR and TPR between underrepresented and majority groups is consistent with the findings of Yu et al. [31]. Our findings validate the prior studies that have shown the possibility of identity-based biases being introduced in predictive analytics due to the awareness of sensitive attributes [32].

6.2 Mitigating Imbalanced Labels and Sensitive Group Disparities

From table 1, it is evident that the distribution of SIRS labels in the entire dataset is uneven, with 38.5% of the population having SIRS. Additionally, there are disparities among sensitive groups (gender and age groups). Three different weights α, β, λ have been introduced where α and β have been used for balancing sensitive attributes and λ is used for balancing SIRS label to mitigates the prediction quality disparity issues. Sample re-weighting based on sensitive attributes is a technique that aims to increase the representation of underrepresented groups in the training set by assigning more weight to their samples in the loss function. In the absence of such weighting, all samples are given equal importance in the loss function, which results in the model focusing more on the majority groups that have more samples than the underrepresented groups. This approach is called "fairness through unawareness" and is specified as the default strategy, which is mentioned in Section 4. Default strategy uses the loss function described in Equation 3.

The weighting functions α, β are defined to assign higher weights to underrepresented groups and lower weights to overrepresented groups in equation 6 as $\alpha(v_{gender}) = 1/v_{gender}$ and $\beta(v_{age\_group}) = 1/v_{age\_group}$, where v_{gender} and v_{age_group} are the proportion vectors of each gender and age group in the data. λ is calculated in the same way. The key distinction is that it is calculated from mini-batches and defined as $g(v_{lbl}) = 1/v_{lbl}$, where v_{lbl} refers to the proportion of each label type (i.e., SIRS or NO SIRS) in the mini-batch, represented as a vector. Therefore, the weight of each sensitive attribute and label in the loss function will be nearly equal on average after re-weighting.

The evaluation results of different models based on four metrics are depicted in Figure 3. The results are also presented in Table 3. This section compares the result between unweighted loss (default) and weighted loss (wgLoss). In general, the models trained with the unweighted loss function (default) had a lower TPR compared to TNR on average (78.6% v.s. 93.1%). This is due to the fact that the model was trained on more samples with the No SIRS label, according to Table 1. However, the difference between overall TPR and TNR grew smaller (83% v.s. 91.6%) after using the weighted loss function. The default model still had a lower TPR than TNR for all groups except school-aged children when examining the population by gender and age groups. This is probably because

of the fact that the school-aged group contains higher proportion of patients that have SIRS. The strategy of using weighted loss improved both TPR and TNR for almost every sensitive groups, indicating that it effectively addresses the unfairness issue among different age groups and gender to some extent. It is also observed that weighted loss increased accuracy for the overall population as well as specific subgroups such as Newborn, Toddler, School-aged, Adolescent, Male, and Female, without sacrificing much accuracy for Neonate and Infant. Hence, we conclude that utilizing weighted loss during training is an effective approach for reducing bias in imbalanced SIRS label as well as sensitive groups.

6.3 Group Fairness

This paper defines group fairness in terms of both equalized odds [11] and demographic parity [10]. This means that each gender and age group requires same TPR and FPR for equality of odds (equalized odds), and positive rate for demographic parity. Here, group fairness is assessed using PR, TPR, FPR, and accuracy. Therefore, we calculate the standard deviation (STD) of each metric. The lower the STD over all the groups, the less disparity there is among the groups, which means there is greater fairness. If group fairness is fully achieved, the STD will be 0. Therefore, the STD can be considered as a measure of group fairness. The evaluation outcomes of the suggested techniques on all sensitive groups are presented in Table 3. Five different models were constructed in accordance with these techniques. The original ANN SIRS classification model is specified by "default" which is mentioned in the previous section. "wgLoss" is the loss weighting strategies by SIRS label and sensitive groups (gender, age group), which is described in Section 6.2. "featureAdded" represents the approach of incorporating gender and age groups to the input of model discussed in Section 6.1. The terms "adv_DP" and "adv_EO" refer to the models that were trained using adversarial learning introduced in Section 5.

The adversarial learning strategies ("adv_DP" and "adv_EO") achieved comparatively minimum SDs for PR, TPR, FPR, and accuracy, demonstrating the highest level of fairness among all the techniques that were compared. SD of TPR and FPR are the lowest for "adv_EO", hence, we can say, this strategy has produced a fair model considering the equality of odds. According to the definition of demographic parity, "adv_DP" should be the fairest model if we consider Positive Rate (PR). PR has increased significantly among different groups compared to other strategies. However, it also has the overall lowest accuracy. As well, demographic parity has some limitations as discussed in Section 2.1. Therefore, "adv_EO" is more suitable for the nature of our dataset. "wgLoss" technique has also demonstrated encouraging outcomes regarding TPR, FPR and accuracy. While maintaining fairness constraints, adversarial models suffer from predictive performance in terms of accuracy. In that sense, if accuracy is the primary concern, "wgLoss" could be considered as a viable strategy in this work.

The adversarial learning for SIRS classification aims to minimize discrimination in the model's hidden states with respect to gender and age group, thereby

allowing it to learn unbiased representations from biased data. Although this technique did not perform the best when considering TPR, TNR, FPR, or accuracy, it did not suffer substantially in terms of these metrics. None of the strategies was consistently the best for all these metrics. The method that frequently performed the best was the one that incorporated sensitive features in the input. However, this strategy was also most frequently the worst in terms of group fairness and often worse as opposed to the default strategy. This highlights the unavoidable trade-offs that must be made when considering fairness [9, 33].

7 Conclusions

In this work, we worked with demographically imbalanced and biased data, and focused on data construction, and model training phases for the mitigation of biases. We demonstrated a general strategy for training unbiased adversarial models which are able to apply constraints of different fairness definitions. As anticipated, the "Fairness through unawareness" strategy was not successful in attaining group fairness. Nevertheless, explicitly providing the sensitive attributes as input to the model resulted in unfair outcomes compared to all other strategies in terms of TPR. Adversarial learning with equality of odds obtained the highest fairness scores on TPR, second and third-best scores on Accuracy and FPR, respectively. Adversarial learning with demographic parity attained the best fairness scores while considering PR. However, it exhibits poor scores for all other metrics.

We found SIRS label balancing and re-weighting underrepresented groups to be a compelling strategy for boosting TNR and TPR (reducing FPR) among these groups. It was also successful in enhancing the prediction accuracy for many of the marginalized groups, specifically Newborn, Toddler, School aged, Adolescent. This discovery highlights a basic yet significant observation that minority groups are likely to be predicted poorly than the majority group. Weighting loss function mitigates this effect to some extent.

Additional work is required to establish further recommendations to tackle fairness and equity in the plethora of healthcare circumstances where unfortunately the ML could otherwise broaden performance gaps and disparities. The directions for future work can be taken to address fairness with some other fairness constraints such as false positives parity, false discovery rate parity, recall parity, etc. We also want to work on an approach for the mitigation of the bias at the inference phase.

8 Acknowledgement

The ELISE project is partially funded by the Federal Ministry of Health; Grant No. 2520DAT66A. This work was also partially supported by the Fraunhofer Internal Programs under Grant No. Attract 042-601000. Ethics approval for use of routine data was given by the Ethics Committee of Hannover Medical School (approval number 9819_BO_S_2021). We would like to thank our colleagues from

the MHH Information Technology (MIT) from the Hannover Medical School for their support.

References

1. Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G.S., Peng, L., Webster, D.R.: Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 125(8), 1264–1272 (2018)
2. Association, A.M., et al.: Ama passes first policy recommendations on augmented intelligence. 2018. Accessed at www. ama-assn. org/ama-passes-first-policy-recommendations-augmented-intelligence on 6 (2018)
3. Bone, R.C., Balk, R.A., Cerra, F.B., Dellinger, R.P., Fein, A.M., Knaus, W.A., Schein, R.M., Sibbald, W.J.: Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Chest 101(6), 1644–1655 (1992)
4. Shapiro, N., Howell, M.D., Bates, D.W., Angus, D.C., Ngo, L., Talmor, D.: The association of sepsis syndrome and organ dysfunction with mortality in emergency department patients with suspected infection. Annals of emergency medicine 48(5), 583–590 (2006)
5. Dellinger, R.P., Levy, M.M., Carlet, J.M., Bion, J., Parker, M.M., Jaeschke, R., Reinhart, K., Angus, D.C., Brun-Buisson, C., Beale, R., et al.: Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2008. Intensive care medicine 34(1), 17–60 (2008)
6. Gupta, A., Liu, T., Shepherd, S., Paiva, W.: Using statistical and machine learning methods to evaluate the prognostic accuracy of sirs and qsofa. Healthcare Informatics Research 24(2), 139–147 (2018)
7. Vembandasamy, K., Sasipriya, R., Deepa, E.: Heart diseases detection using naive bayes algorithm. International Journal of Innovative Science, Engineering & Technology 2(9), 441–444 (2015)
8. Piri, S., Delen, D., Liu, T., Zolbanin, H.M.: A data analytics approach to building a clinical decision support system for diabetic retinopathy: Developing and deploying a model ensemble. Decision Support Systems 101, 12–27 (2017)
9. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807 (2016)
10. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
11. Hardt, M., Price, E., Srebro, N., et al.: Equality of opportunity in supervised learning in advances in neural information processing systems (2016)
12. Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 560–568 (2008)
13. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and information systems 33(1), 1–33 (2012)
14. Wang, T., Zhao, J., Yatskar, M., Chang, K.W., Ordonez, V.: Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5310–5319 (2019)

15. Madras, D., Creager, E., Pitassi, T., Zemel, R.: Learning adversarially fair and transferable representations. In: International Conference on Machine Learning. pp. 3384–3393. PMLR (2018)
16. Wu, C., Wu, F., Wang, X., Huang, Y., Xie, X.: Fairrec: fairness-aware news recommendation with decomposed adversarial learning. AAAI (2021)
17. Du, M., Yang, F., Zou, N., Hu, X.: Fairness in deep learning: A computational perspective. IEEE Intelligent Systems 36(4), 25–34 (2020)
18. Cohen, I.G., Amarasingham, R., Shah, A., Xie, B., Lo, B.: The legal and ethical concerns that arise from using complex predictive analytics in health care. Health affairs 33(7), 1139–1147 (2014)
19. Char, D.S., Shah, N.H., Magnus, D.: Implementing machine learning in health care—addressing ethical challenges. The New England journal of medicine 378(11), 981 (2018)
20. Edwards, H., Storkey, A.: Censoring representations with an adversary. arXiv preprint arXiv:1511.05897 (2015)
21. Beutel, A., Chen, J., Zhao, Z., Chi, E.H.: Data decisions and theoretical implications when adversarially learning fair representations. arXiv preprint arXiv:1707.00075 (2017)
22. Kearns, M., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: International Conference on Machine Learning. pp. 2564–2572. PMLR (2018)
23. Hébert-Johnson, U., Kim, M., Reingold, O., Rothblum, G.: Multicalibration: Calibration for the (computationally-identifiable) masses. In: International Conference on Machine Learning. pp. 1939–1948. PMLR (2018)
24. Chen, I., Johansson, F.D., Sontag, D.: Why is my classifier discriminatory? Advances in Neural Information Processing Systems 31 (2018)
25. Wulff, A., Montag, S., Rübsamen, N., Dziuba, F., Marschollek, M., Beerbaum, P., Karch, A., Jack, T.: Clinical evaluation of an interoperable clinical decision-support system for the detection of systemic inflammatory response syndrome in critically ill children. BMC medical informatics and decision making 21(1), 1–9 (2021)
26. Goldstein, B., Giroir, B., Randolph, A., et al.: International pediatric sepsis consensus conference: definitions for sepsis and organ dysfunction in pediatrics. Pediatric critical care medicine 6(1), 2–8 (2005)
27. Niazkar, M., Talebbeydokhti, N., Afzali, S.H.: Novel grain and form roughness estimator scheme incorporating artificial intelligence models. Water resources management 33(2), 757–773 (2019)
28. Niazkar, M.: Revisiting the estimation of colebrook friction factor: a comparison between artificial intelligence models and cw based explicit equations. KSCE Journal of Civil Engineering 23(10), 4311–4326 (2019)
29. Niazkar, M.: Assessment of artificial intelligence models for calculating optimum properties of lined channels. Journal of Hydroinformatics 22(5), 1410–1423 (2020)
30. Kleinberg, J., Ludwig, J., Mullainathan, S., Rambachan, A.: Algorithmic fairness. In: Aea papers and proceedings. vol. 108, pp. 22–27 (2018)
31. Yu, R., Li, Q., Fischer, C., Doroudi, S., Xu, D.: Towards accurate and fair prediction of college success: Evaluating different sources of student data. International Educational Data Mining Society (2020)
32. Barocas, S., Hardt, M., Narayanan, A.: Fairness and machine learning. fairmlbook. org. URL: http://www. fairmlbook. org (2019)
33. Fazelpour, S., Lipton, Z.C.: Algorithmic fairness from a non-ideal perspective. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 57–63 (2020)