applied sciences

MDPI

Article

Detection and Evaluation of Machine Learning Bias

Salem Alelyani [1,2]

[1] Center for Artificial Intelligence (CAI), King Khalid University, Abha 61421, Saudi Arabia; s.alelyani@kku.edu.sa
[2] College of Computer Science, King Khalid University, Abha 61421, Saudi Arabia

Abstract: Machine learning models are built using training data, which is collected from human experience and is prone to bias. Humans demonstrate a cognitive bias in their thinking and behavior, which is ultimately reflected in the collected data. From Amazon's hiring system, which was built using ten years of human hiring experience, to a judicial system that was trained using human judging practices, these systems all include some element of bias. The best machine learning models are said to mimic humans' cognitive ability, and thus such models are also inclined towards bias. However, detecting and evaluating bias is a very important step for better explainable models. In this work, we aim to explain bias in learning models in relation to humans' cognitive bias and propose a wrapper technique to detect and evaluate bias in machine learning models using an openly accessible dataset from UCI Machine Learning Repository. In the deployed dataset, the potentially biased attributes (PBAs) are gender and race. This study introduces the concept of alternation functions to swap the values of PBAs, and evaluates the impact on prediction using KL divergence. Results demonstrate females and Asians to be associated with low wages, placing some open research questions for the research community to ponder over.

Keywords: machine learning bias; bias detection; bias evaluation; KL divergence; explainable models; cognitive bias

Citation: Alelyani, S. Detection and Evaluation of Machine Learning Bias. *Appl. Sci.* **2021**, *11*, 6271. https://doi.org/10.3390/app11146271

Academic Editor: Federico Divina

Received: 22 March 2021
Accepted: 5 July 2021
Published: 7 July 2021

1. Introduction

Machine learning bias has garnered researchers' attention lately [1–4]. Researchers are mostly concerned about the potential bias that machine learning systems may demonstrate against protected attributes such as: gender, race, age, etc. [5]. The interest in this area was initiated by a report published by ProPublica.com [1] that examined a judicial risk assessment; researchers found that the system exhibits bias toward black people. Later on, many other news reports and research papers have raised the concern of bias in data. Another more recent example of machine learning bias is Amazon's hiring system. The system did not like women a Reuters report claims (https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G, accessed on 1 February 2021). The report stated that the hiring system was fed ten years of hiring data experience from the company. However, the system gave less weight to the CVs that included women indicators.

The aforementioned typical machine learning examples work by extracting patterns from the training data. Due to the fact that data is collected from historical humans' practices, the element of bias is prevalent within it. For instance, intelligent hiring systems learn the behavior from the hiring practices embedded in the training data fed to the model. It is a natural phenomenon that humans have a cognitive bias naturally existing in their judgment and decisions. Human cognitive bias is a well-studied fact in the psychology field [6,7]. In order to make a decision, the enormous amount of information residing in a human's brain is filtered, and only relevant information is used for decision-making.

In our quotidian hiring decisions, generally a gender bias is observed, where usually a male employee is preferred over a female one. For instance, technical positions at

Amazons are male dominated, as mentioned by the same report. Therefore, the company's historical data fed to the intelligent hiring system as training data is biased toward male candidates. A gender bias is observed in many other disciplines; thus the datasets also follow a biased pattern in which the data is more inclined toward a certain gender [8–11]. Similarly, the recidivism prediction in the judicial system extracts convicted data from a set of 137 questions; then, an equation is applied to score the likelihood of a convict to commit another crime. The equations consist of some parameters that are weighted in association with each question. The parameters of the equation are either obtained by training the model using real-world data or chosen by a domain expert. In either case, it contains bias introduced by humans.

Since the dawn of artificial intelligence, we have deliberated to build machines that mimic humans' ability to think and behave [12]. As consequence, we usually set our intelligence and cognitive abilities as the bar for machine intelligence in spite of the fact that humans suffer a cognitive bias. The machine learning models are trained using the data that represents human behavior [13,14]. The data fed to the machine learning models for training is considered as the *ground truth*, enabling the model to learn behavior from this data. Nevertheless, considering the bias in human nature and cognitive thinking, the training data also depicts such biased characteristics, which ultimately impact the model's learning by inducing bias [15]. Objecting to this bias, Amazon shut the hiring system off to stop it. Machine bias will not stop showing up everywhere until we stop making it: *bias in, bias out*.

The question remains: Why do we endure cognitive bias? To answer such a question, we need to emphasize the decision and judgment attributes considered during the process of the decision and judgment [16]. The judge knows the defendant's race during the trial, which may induce a cognitive bias or even worse, preconceived judgment. Bias against gender or race in our life is not easy to eliminate due to the presence of the protected attributes. Therefore, we aim to mitigate it. However, this is inexact in machine learning, in which we can hide these attributes in the model building process, resulting in a total elimination of bias. In contrast to human judgment, in machine learning we should not aim to mitigate bias since it may lead to more unjustifiable human intervention, which may cause unfairness. We can, in fact, eliminate protected attributes from training data in order to wipe out bias.

In [17], the authors proposed a solution to eliminate model bias by re-weighting data samples without changing the class label. This approach gave almost equal learning performance to the true labels' classifier. However, in the real-world, we cannot simply assume the existence of bias without detecting it first. Similarly, Ref. [18] argues that the bias is introduced from humans who are responsible of labeling training data. Yet, the authors argue that humans are not the only sources of bias, but algorithms are as well. For instance, recommender systems that provide humans with their preferred content are responsible for such issues. The bias generation in this scenario is iterative. They concluded that the iterative bias negatively impacts the learning performance, with iterated filter bias having the greatest impact. In addition, Agarwal et al., in [19], defined the unbiased classifier as the classifier that can predict the class label independently of the protected attribute. They proposed to two reductions to improve fairness in binary classifiers.

The research community are aiming to propose techniques to mitigate algorithmic bias. However, there is limited work in detecting and evaluating bias in the first place. Thus, in this paper, we propose a technique to detect and evaluate machine learning bias. Our contributions in this paper are three-fold:

- Firstly, we aim to investigate machine learning bias in relation to human cognitive bias. Some important philosophical arguments will be discussed in this section.
- Secondly, we intend to propose a wrapper bias detection technique based on a novel alternation function to detect machine learning bias.

- Lastly, we propose an evaluation method to determine the bias in data using KL divergence. This may help in creating more reliable and explainable machine learning models. We will conduct several experiments to validate our contribution.

2. Bias and Unfairness

In the machine learning literature, researchers use the terms bias and unfairness interchangeably [3,16,17,20–23]. We believe this is a fundamental mistake that may lead to more human intervention in machine decisions. For example, assume that in an automated hiring system, males are preferred over females. The system learns the criteria of hiring automatically from the historical training data we provided. Therefore, the algorithm is not intrinsically unfair, yet the data has a historical bias due to the fact that in this position, males are dominant [24,25]. If we want to eliminate or mitigate model bias, we need to tune the model's parameters to obey our desires of expected outcomes [26]. Tuning the parameters is equivalent in this case to changing the hiring criteria to accredit one gender over another, which is not fair [27,28].

From the aforementioned example, a question arises: Who has the right to distort the model in real-world application that will impact human lives? How can we rationally justify the model's outcomes, if we intentionally tweaked it? Although we might mitigate bias, this may lead to model unfairness.

Humans tend to deviate from rational decisions or judgments to irrational ones. This is a known fact called cognitive bias [6,7]. Individuals make their own irrationality in decision-making due to preconceived beliefs about the topic, usually created from a cumulative subjective reality, not from the input [29,30]. Cognitive bias is not always bad. It was found that it may expedite the decision-making process [31]. Regardless of the benefit of human cognitive bias, machine learning bias is not as desirable and thus impacts the decision-making ability of the algorithm [32]. This is due to the machines' nature. First, machines do not need this kind of expedited decision-making process. They expedite it by either increasing processing power and/or decreasing algorithmic computational complexity. Second, machine learning bias cannot be easily distinguished from error unless the data says otherwise. It is worth mentioning here that human cognitive bias is inherited in the collected data utilized for training machine learning models. Hence, it will learn bias from training data. This is the entire concept surrounding machine learning models.

In this paper, we will tackle the bias in machine learning models, not the unfairness. Therefore, we define machine learning bias as the difference in the underlying distribution of the model learning outcome with respect to certain group(s) influenced by their affiliation to the specific group. The group could be gender, race, age, or any other protected attribute.

3. Machine Learning Bias: Detection and Evaluation

In most real-world cases, as we mentioned in the introduction, we humans tend to decide whether the attribute can be biased or not. Usually, bias attributes are the protected attributes by law. For instance, gender, race, religion, etc., are protected attributes by law; thus, these correspond to some ethical implications [33,34]. In our cognitive biases, we have certain beliefs such as: bias against females in highly paid jobs, bias against black people in the judicial system, and so forth. These beliefs could be true. They could be mitigated over time, yet we are still sensitive about them.

The problem is: How can we be sure about the presence of bias until we detect it and quantify it [35]? In this section, we propose a technique to find out whether an attribute can be PBA toward the classes or not. Furthermore, we will quantify the amount of bias in PBA. To prove the concept, we will conduct the experiment in this paper on certain protected attributes, namely: gender and race.

As we discussed above, we can confidently claim that machine learning bias comes from the training data, which is inherited from the cognitive bias in our decisions and judgments [36]. Furthermore, most attributes are unbiased by nature, yet they might implicitly or indirectly inherit bias. For instance, the *degree* attribute for job applicants

might be biased due to the indirect impact of the *gender* attribute. It is known that fewer females major in engineering or technology, and even fewer have graduate degrees in these fields. In this example, we need to find the quantity of the bias that the gender attribute introduces on the learning model, not the degree attribute.

We believe that bias in the model comes from the statistical priory in the learning data. As one may notice in the degree–gender example, the prior probability of a specific gender should not be controlled in the learning model. If we want to mimic human intelligence, we need to accept bias. Nonetheless, it is a privilege to be able to detect and quantify bias for better explainable models.

4. Our Contribution

In this section, we propose a novel technique for detecting and evaluating potential machine learning bias. As we mentioned above, we believe data is biased by nature due to the cognitive bias of human brains. Therefore, evaluating the potential bias of machine learning models would be valuable for better model explainability. In other words, if we understand model bias, we will better justify the model's behavior. The proposed technique is explainable. Thus, we will be able to detect biased attributes with a certain confidence level and determine which category of the attribute is causing the bias against which.

4.1. Notations

We assume we have a dataset $\mathbf{D} \in \mathbb{R}^{n \times m}, \mathbf{D} = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_i, \ldots, \mathbf{d}_n\}$, where $\mathbf{d}_i$ is the ith instance (i.e., data sample) of $\mathbf{D}$. We also represent $\mathbf{D}$ as a set of attributes' vectors. Thus, $\mathbf{D} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_j, \ldots, \mathbf{a}_m\}$. Since we mainly tackle supervised techniques, we will be using y to denote the target (i.e., class label), where it is a vector of labels with size n; hence, $\mathbf{y} = \{y_1, y_2, \ldots, y_i, \ldots, y_n\}$ is the actual class label. The predicted target, which is obtained from the learning process, is denoted as $\hat{\mathbf{y}}$. We believe the proposed technique will generalize well with both discrete and continuous values. Therefore, $f(\mathbf{D}) \rightarrow \hat{\mathbf{y}}$ is either a regression or a classification model that takes dataset $\mathbf{D}$ and assigns each data sample $\mathbf{d}_i$ to a specific target $\hat{y}_i$.

4.2. Problem Statement

In this paper, we aim to define, evaluate, and detect bias in machine learning models. Formally, we want to detect a subset of $\mathbf{D}$; $\mathbf{a}_{Bias} \subset \mathbf{D}$ that may introduce bias according to a specific evaluation metric. In this problem, and without loss of generality, we assume all potentially biased attributes to be categorized, and the class labels are discrete.

Definition 1. *Potentially Biased Attribute (PBA). Without loss of generality, we assume that $\mathbf{a}_j$ is a categorical attribute. We can say that $\mathbf{a}_j \in \mathbf{D}$ is a potentially biased attribute if $f(\mathbf{D}) \not\sim f(\varphi(\mathbf{D}))$.*

Where $\varphi(\mathbf{D})$ is an alternation function; see Definition 3. We will denote the prediction of $f(\varphi(\mathbf{D}))$ model as $\hat{\mathbf{y}}_\varphi$. For the remainder of this paper, we call $\hat{\mathbf{y}}_\varphi$ the alternative prediction. It is not the actual prediction. Instead, it is the prediction when we change the PBA's values using $\varphi()$.

For example, let $\mathbf{D}$ be a dataset of n applicants, and $\mathbf{a}_j$ is a gender attribute, while $\mathbf{y}$ represents whether the applicant is qualified 1 or not qualified applicant 0. The problem in this scenario is to predict whether an applicant $\mathbf{d}_i$ is qualified for a job or not. If $\mathbf{a}_j$ is able to change the prediction by only changing the gender while maintaining the values of the remaining attributes, then the attribute $\mathbf{a}_j$ is said to be a *potentially biased attribute* (PBA).

Bias Evaluation: KL Divergence

Machine learning model is considered biased when the hypothesis prediction diverges with respect to one or more specific values of PBA. Formally, we can define machine learning bias as:

Definition 2. *A predictor is considered biased if it is dependent on one or more PBA given the class label [16,19].*

Assume the model, for instance, predicts the wage of a male instance with a specific amount. Then, we assume the same instance is passed to the model after changing the gender to a female. If the model's prediction changes dramatically by changing the gender only, this can be considered as model bias. We propose to evaluate bias amounts by quantifying the divergence between the densities of the predicted wages for the two predictors. We introduce a new concept to evaluate and quantify machine learning bias. We call this concept "*alternation*" function.

Definition 3. *Alternation is a function that alternates between attributes' values so the values are swapped each time.*

Alternation takes an attribute and changes the instance's identity. The purpose is to test the predictability consistency if the identity changes. If a female instance becomes male, is there any effect on the prediction? Similarly, if the race of an instance changes, does the prediction change accordingly? Alternation changes the identity of an instance. In another word, we aim to check the dependency of the predictor on the attribute.

In this paper, we denote the alternation function as $\varphi(\cdot)$. The function takes a dataset as an input and returns the alternative dataset, which is the dataset with alternative values of a specific attribute. $\varphi(\mathbf{D}) = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \neg\mathbf{a}_j, \ldots, \mathbf{a}_m\}$. In this context, $\varphi(\cdot)$ is a function that switches the protected attribute's values in a way that $\mathbf{a}_j$ becomes $\neg\mathbf{a}_j$. Thus, the female values become male, and vice versa.

We need to find the divergence between the distribution of the original class $\mathbf{y}$ and the predicted classes $\hat{\mathbf{y}}$ and $\hat{\mathbf{y}}_\varphi$. Measuring the variation of information might indicate some differences between classes. However, bias should not be symmetric. The bias might be against or with certain attributes' values. For example, we usually believe that there is a bias against female or black people in gender and race attributes, respectively. Thus, the bias is with male and white, for instance. Thus, the measure of bias $\gamma(\cdot,\cdot)$ of vectors $\mathbf{u}$ and $\mathbf{v}$ should obey the following rule:

$$\gamma(\mathbf{u}, \mathbf{v}) \neq \gamma(\mathbf{v}, \mathbf{u}) \tag{1}$$

Another important property of $\gamma(\cdot,\cdot)$ is that the difference between the values from Equation (1) indicates the amount of bias. The larger the difference is, the larger the bias with or against a certain value.

In order to satisfy this asymmetric property, we use Kullback–Leibler (KL) divergence to estimate the distribution difference between the wage prediction for each value in PBA. We aim to find the impact of the PBA on the model's prediction using the alternation function. In the interest of generating different training and testing datasets to build models that will generalize on real-world situations, we applied a cross-validation (CV) sampling technique. It is known for its statistical ability of generating less biased datasets, also known as data folds. In CV, we split the dataset into *k*-folds where each fold contains training and testing samples. In this work, we set $k = 10$. To reduce variability of training sets, each training fold will consist of 90 percent of the data, while the remaining 10 percent will be for testing. Each sample of the dataset will be used for testing in just one fold. Unlike training folds, where each sample will appear in training folds *k-1* times. Following that, the model will be trained and tested on each fold. Thus, after predicting the wage in each fold, we predict it again using the same instance, but by changing the value of the PBA. In the gender attribute example, we change the value of female instance to be male, and vice versa.The KL divergence estimates the difference between the distribution of two

populations based on its information [37]. Assuming we have the densities p and q, we calculate the divergence between them using the following equation:

$$D_{KL}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx \geq 0. \tag{2}$$

We use Equation (2) to observe the difference between the distribution of the predicted class label $\hat{\mathbf{y}}$ and the alternative prediction $\hat{\mathbf{y}}_\varphi$ with respect to each PBA's value. If the result of D_{KL} is zero, the two distributions are identical. Hence, there is no bias caused by this attribute. Otherwise, the distributions differ; hence, there is bias. The larger the result is, the greater the difference between the distributions would be, which indicates the presence of bias. Nevertheless, the result cannot be negative. The way we avail of the KL divergence is simply by applying it to the distributions of the wage. For instance, the distribution of the female wage is denoted as p, while the distribution of female wage after applying the alternation function is denoted as q. The divergence between the two distributions (p and q) represents how much the wage changes when changing the gender only. Since the wage is a continuous random variable, p and q will be probability density functions (PDFs). The variables of p and q are assumed to be drawn from $\sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\sim \mathcal{N}(\mu_2, \sigma_2^2)$, respectively. From Equation (2), we derive the following equation:

$$D_{KL}(p||q) = -\int p(x) \log q(x) dx + \int p(x) \log p(x) dx$$

The two terms will be as follows:

$$-\int p(x) \log q(x) dx = -\int p(x) \log \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp -\frac{(x-\mu_2)^2}{2\sigma_2^2} dx,$$

and

$$\int p(x) \log p(x) dx = -\frac{1}{2}(1 + \log 2\pi\sigma_1^2).$$

After putting the two terms together again, we obtain:

$$D_{KL}(p||q) = \frac{1}{2} \log(2\pi\sigma_2^2) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}(1 + \log(2\pi\sigma_1^2))$$

Finally, $D_{KL}(p||q)$ is going to be:

$$D_{KL}(p||q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \tag{3}$$

The KL divergence is evaluated twice using Equation (3) for each binary protected attribute (e.g., gender).

5. Methodology

Our empirical evaluation intends to find out if the dependency of the hypothesis on the protected attribute. If the predictor predicts significantly different labels with different density if the alternation function is applied, then this is an indicator that the hypothesis is dependent on that PBA. In contrast, if the hypothesis is independent of the PBA (i.e., the two predicted labels densities are similar), the hypothesis is not biased. Figure 1 illustrates the proposed framework to quantify and detect bias. We consider the density of the predicted class label of the original dataset **D** to be the original distribution in the divergence evaluation denoted as p. While the density prediction of the alternated dataset is the other density denoted as q. D_{KL} quantifies how much p is deviated from q. In other words, it evaluates the divergence between the density of predicted class label $p_{\hat{\mathbf{y}}}$ and the

density of the predicted class label $q_{\mathbf{y}}$ after applying the alternation function with respect to the PBA's values.

To evaluate bias in the gender attribute, we should do the following:

1. Train the model $f(\cdot)$ on the dataset $\mathbf{D}$.
2. Predict the class label for each data point in $f(\mathbf{D}) \rightarrow \hat{\mathbf{y}}$.
3. Apply the alternation function on the gender attribute $\varphi(\mathbf{D}) \rightarrow \mathbf{D}_{\varphi}$.
4. Train the model $f(\cdot)$ on the alternative dataset $\mathbf{D}_{\varphi}$.
5. Predict the alternative predicted label $f(\mathbf{D}_{\varphi}) \rightarrow \hat{\mathbf{y}}_{\varphi}$.
6. Using Equation (3), evaluate KL Divergence between the distributions of $\mathbf{y}$ and $\hat{\mathbf{y}}_{\varphi}$ for each gender distinctively across all folds. That is, $D_{KL}(p_{\mathbf{y}_f}||q_{\hat{\mathbf{y}}_{\varphi f}})$ and $D_{KL}(p_{\hat{\mathbf{y}}_m}||q_{\hat{\mathbf{y}}_{\varphi m}})$.
7. The difference between the D_{KL} with respect to the gender represents the bias. The larger D_{KL} is, the larger the bias will be.

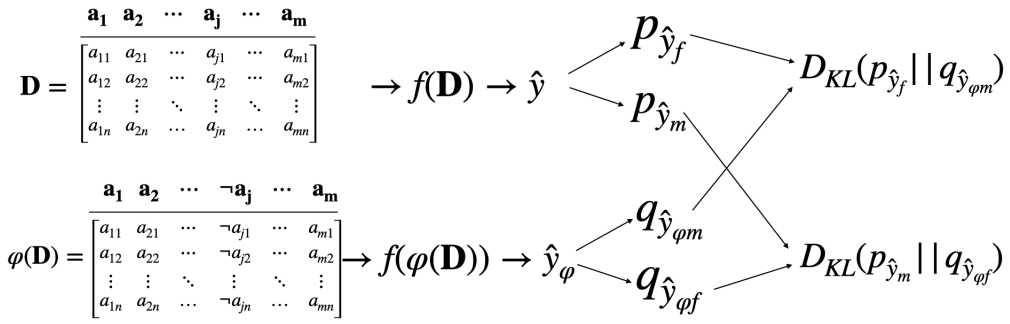

Figure 1. KL divergence for bias evaluation.

The proposed methodology to evaluate bias in machine learning models is a wrapper approach. Thus, it is a model-specific approach and is more expensive in terms of computational complexity. However, it is known to be more accurate with respect to the model. On the other hand, the alternation is performed on the PBAs with respect to the number of values in the attribute $|\mathbf{a}_i|$. The number of the sets that would be generated for the PBAs' attributes is equal to:

$$1 + \sum_{j=1}^{|\mathbf{a}_i|} (|\mathbf{a}_i| - j),$$

which seems large; yet, usually the PBAs' number of categories $|\mathbf{a}_i|$ is small like *gender* or *race*.

6. Experiment Setup

We conducted the experiment to detect and evaluate the bias introduced by different PBAs to the predicted wage. The PBAs in this experiment are gender and race. We aim to predict the wage of the instance using the original gender and race. Then, we alternate the gender and race to see how the predicted wage is changing. As we mentioned above, we will evaluate the divergence of the mean of the prediction corresponding to each PBA's value to detect where bias occurs and to evaluate the amount and direction of bias.

6.1. Dataset

In this experiment, we used a publicly available dataset from UCI Machine Learning Dataset Repository called: Census-Income Database. It originally consisted of approximately 300,000 instances representing the US census extracted from the 1994 and 1995 population surveys. The data originally contained more than forty demographic- and employment-related attributes.

For the sake of proving the concept of the proposed technique, the data was cleaned according to some specific criteria. First, a class label was chosen from the list of attributes. We will be trying to predict the wage for each instance. Therefore, the wage will be the label and will be denoted with **y**. Then, we manually selected 10 attributes only, which we believe is relevant to the class label. Among the selected attributes, we will be measuring the amount of bias in the learning model with respect to PBAs, namely gender and race.

The instances were further cleansed according to some values in the attributes. For instance, in the *Education* attributes, we kept five categories only: *High School Graduate, Some College However, No Degree, Bachelor's Degree, Master's, and Doctorate Degree*. Instances with missing data were eliminated as well, and after the cleaning process, 14,864 instances from the dataset were included for this study.

Such data might be helpful in predicting any financial guarantees from banks, insurance companies, and so forth. Therein lies the problem. If the decision makers will be taking such data as a source of evidence toward any decision that might harm a human, then we need to make sure that all attributes are not biased toward a specific gender or race.

6.2. Algorithm and Model Selection

Since the label uses continuous values, we are going to use regression to predict it. A polynomial regression algorithm is applied to build the model throughout the experiment. Any other machine learning algorithm can be used similarly. Here, we apply one algorithm since the experiment is not meant to compare algorithms. We want to prove the concept we are proposing.

To ensure consistency in our results, we apply the *10-fold cross-validation* model selection technique. For each fold, we train the model using 90 percent of the data. Then we predict the target using the model in each fold. After that, we apply the alternation function and predict the target again for the same folds. It is necessary to point out the fact that D_{KL} is evaluated for each gender using the gender's instances which are not the same instances for the other gender. Thus, the asymmetrical results are not due the asymmetrical property of D_{KL}.

6.3. Experimental Findings

Figures 2 and 3 illustrate the results of an experiment to predict the wage. We conducted the experiment as explained earlier using the 10-folds cross-validation model selection. The plots show the results of the average predicted wage for both male and female beside the average predicted wage when we apply the proposed alternation function on the gender attribute. The plot on the left, Figure 2, shows the female average predicted wage in green circle markers. The results show a considerable improvement in the predicted wage after applying alternation on the gender attribute. This indicates the prediction of females' wage increase if their identity changed to male. In other words, a male instance with the same profile would earn more wage than a female instance. This indicates the presence of bias against females.

The plot on the right-hand side, Figure 3, shows the average predicted wage for males in green, which degraded substantially after applying the alternation function. In contrast to the female results above, this indicates bias with males. In summary, we can say that the prediction of the wage is obviously influenced by the gender attribute. Both figures show the bias against females. In other words, one profile for one person could be treated differently if we only change the gender.

In real-life scenarios, we usually tend to consider this as a kind of bias. In this experiment, what we have done thus far is simply saying that if this person with the same profile was a male, the wage would be higher than if the person was a female. To quantify this bias, we evaluate the KL divergence between the original and the alternative prediction mean. Figure 4 depicts the KL divergence of the results presented previously. The red line represents the KL divergence of the results in Figure 2, which is the divergence between the predicted wage of the original dataset and the predicted wage when changing female

into male. We can clearly observe bias against females. In contrast, the green line indicates less bias against males. This is the KL divergence between the predicted wage for females and the predicted wage for females after applying alternation. The green line, on the other hand, is the bias against males. The bias against females is significant in these results.

At the moment, we can see the decrease in the wage if the gender is female. Let us now consider the other controversial attribute (i.e., *race*). We have five different races in the dataset; namely: (a) American Indian, Aleut, or Eskimo, (b) Asian or Pacific Islander, (c) Black, (d) Other, and (c) White. We will use the word Indian and Asian to represent the races in (a) and (b), respectively.

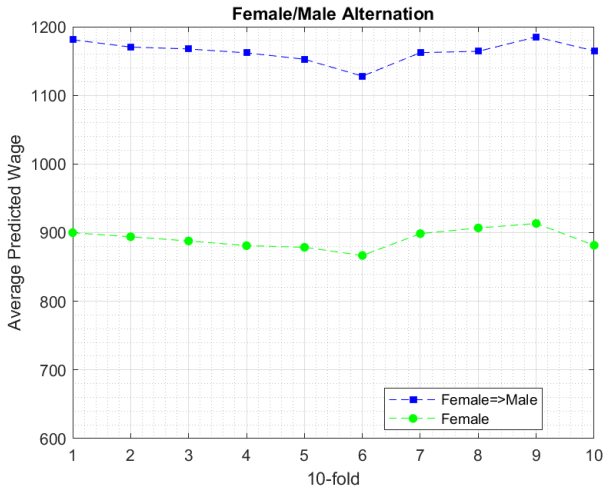

Figure 2. Female to male alternation. It shows that the female wage (in green) improves when the gender is changed to male (in blue).

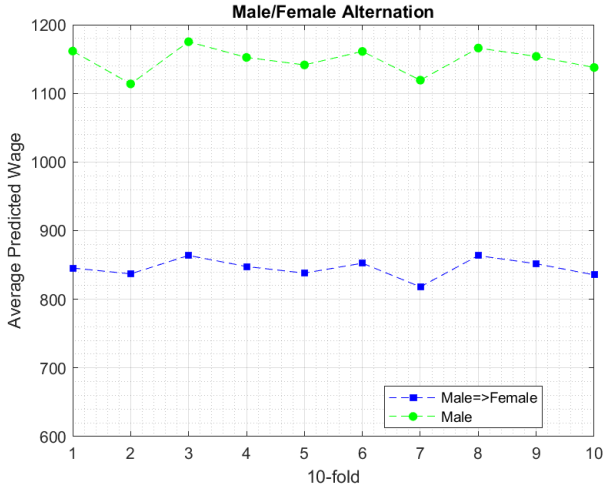

Figure 3. Male to female alternation. The wages of males (in green) decrease when we change the gender attribute to a female (in blue).

We used the same methodology with the race attribute. The results are demonstrated in Figures 5–7. Carefully analyzing each plot, we observe interesting results. Based on Figure 5; plots: Asian/Indian Alternation and White/Indian Alternation, we cannot claim any kind of bias on or against the American Indian race. In all cases, it gives better wage prediction when we change the race from or to American Indian. We find this behavior to be quite confusing; however, this needs further investigation and should be considered when building a real-world model.

Similarly, we did not notice a clear bias against Asian or Pacific races (see Figure 5; plots: Asian/Other Alternation and White/Asian Alternation), except with the black race.

There is a bias against Asians when we change race to black. In other words, when we change the race of an Asian instance to a black, the wage noticeably gets higher. Likewise, when we change the race from black to Asian, the wage gets lower. We believe this is a kind of bias. It is not surprising that the wage in the United States might be higher for black than Asians. However, the surprising result comes in Figure 5 plot White/Black Alternation. The results show a clear bias against the white race when it comes to black. If we change the race from white to black, unexpectedly, the wage gets much higher, and it gets lower when the race is changed from black to white. This can be taken as evidence to our earlier claim about bias. In this case, can we call it bias, even if it is against a very advantageous race? Should we aim here to mitigate bias, so the black race would suffer from reducing their wages or the white race would get more benefits by increasing their wages?

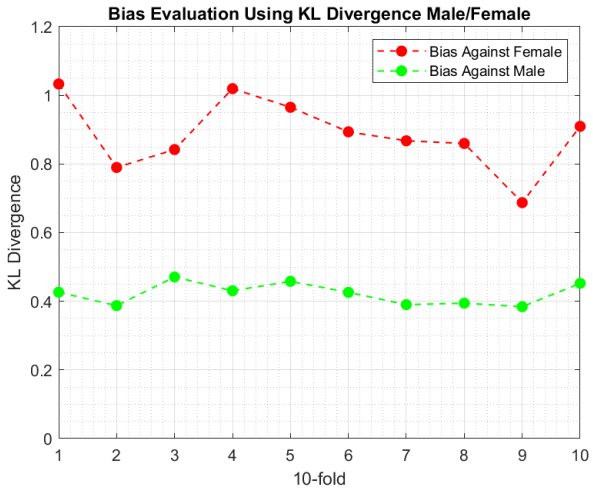

Figure 4. The KL divergence of male and female shows the bias against females.

Figure 5 plot White/Asian Alternation illustrates the evaluation of bias in the race attribute. In most cases, the amount of bias is not significant and is inconsistent in all training folds. There are some folds of the dataset that show some tendency toward more bias against some races, yet it was not concerning since it was not in all folds. This might be attributable to some outliers in the fold itself.

These results can be investigated more when we aim to build a real-world model that would affect people's lives. In this paper, we intended to detect and evaluate bias, not to justify the model. In the future, we will aim to interpret the model.

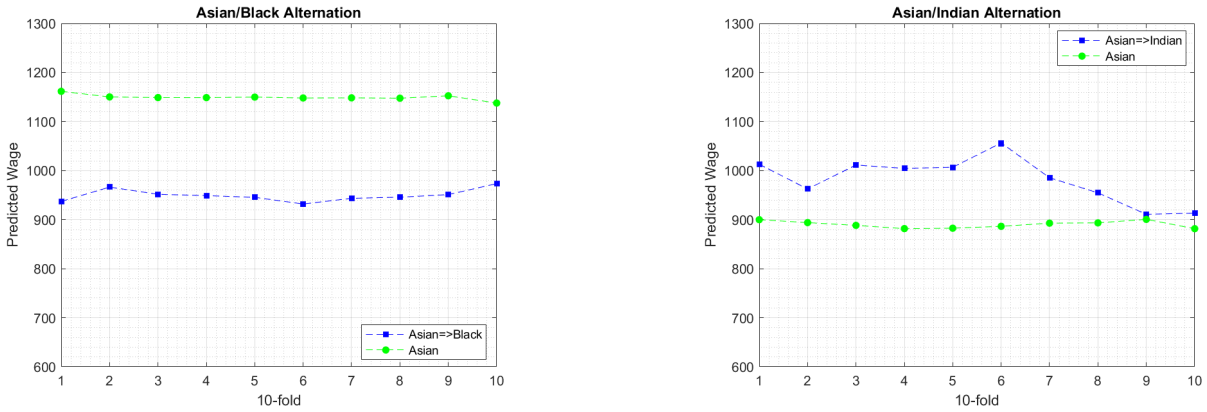

Figure 5. *Cont.*

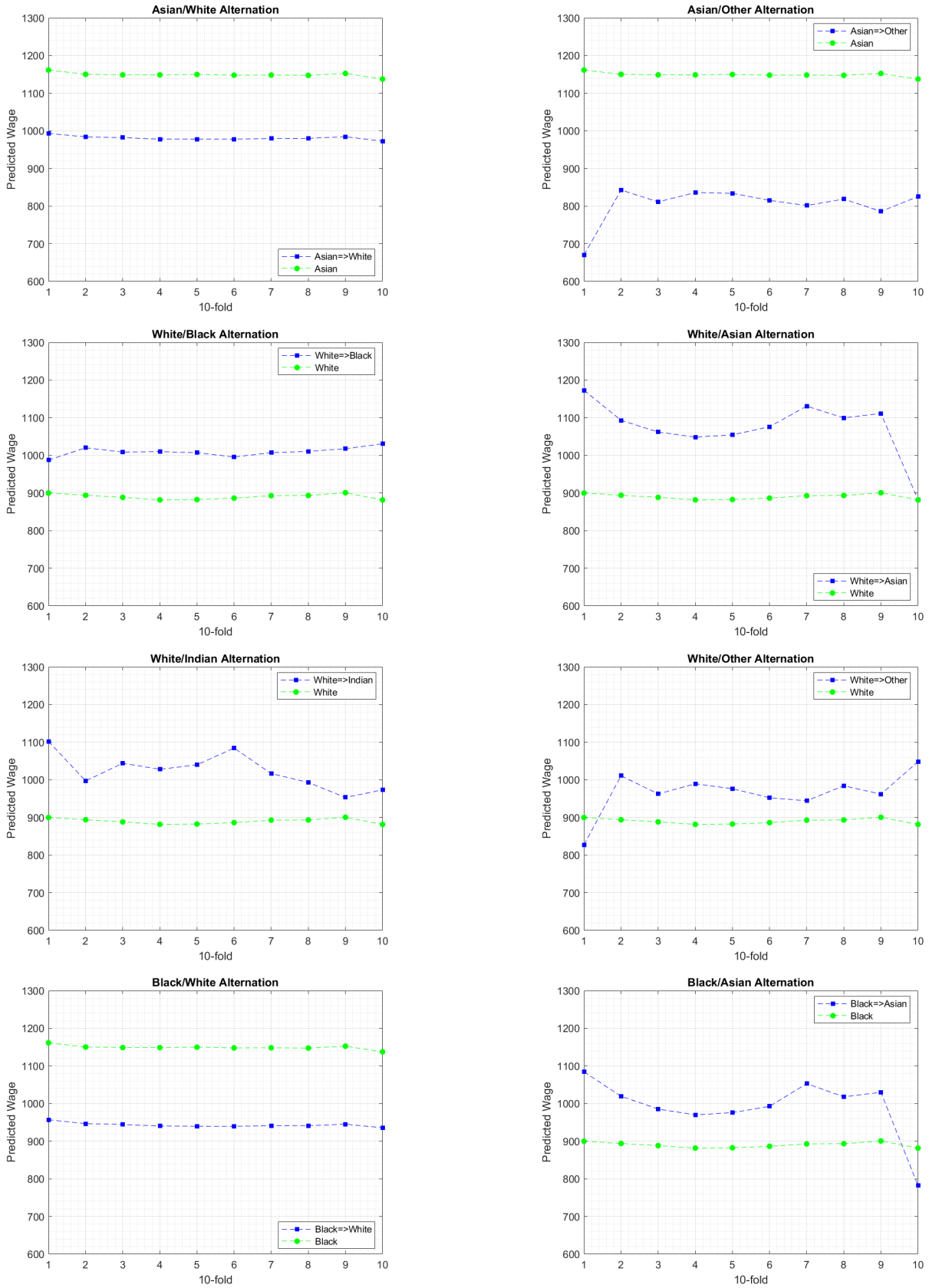

Figure 5. Bias in wages in terms of Race1.

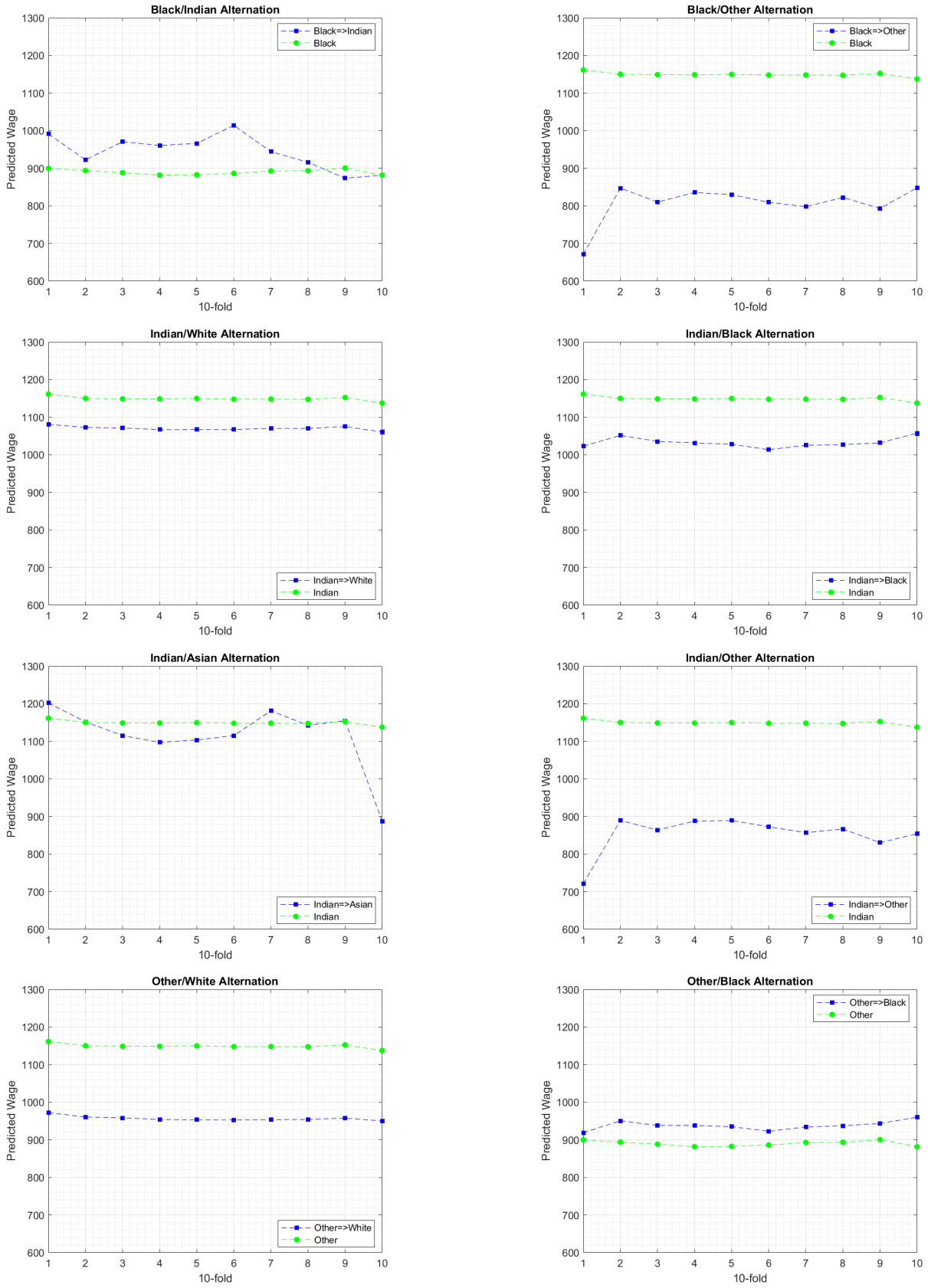

Figure 6. *Cont.*

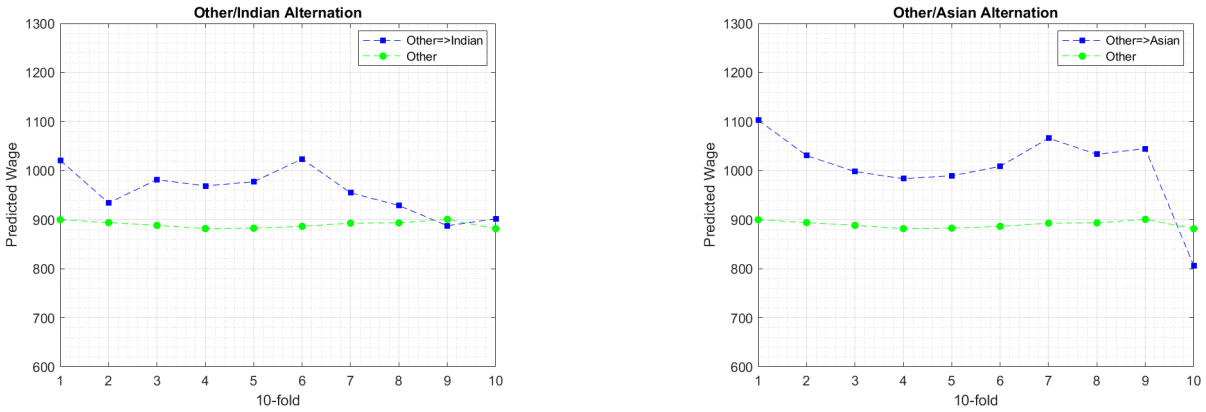

Figure 6. Bias in wages in terms of Race2.

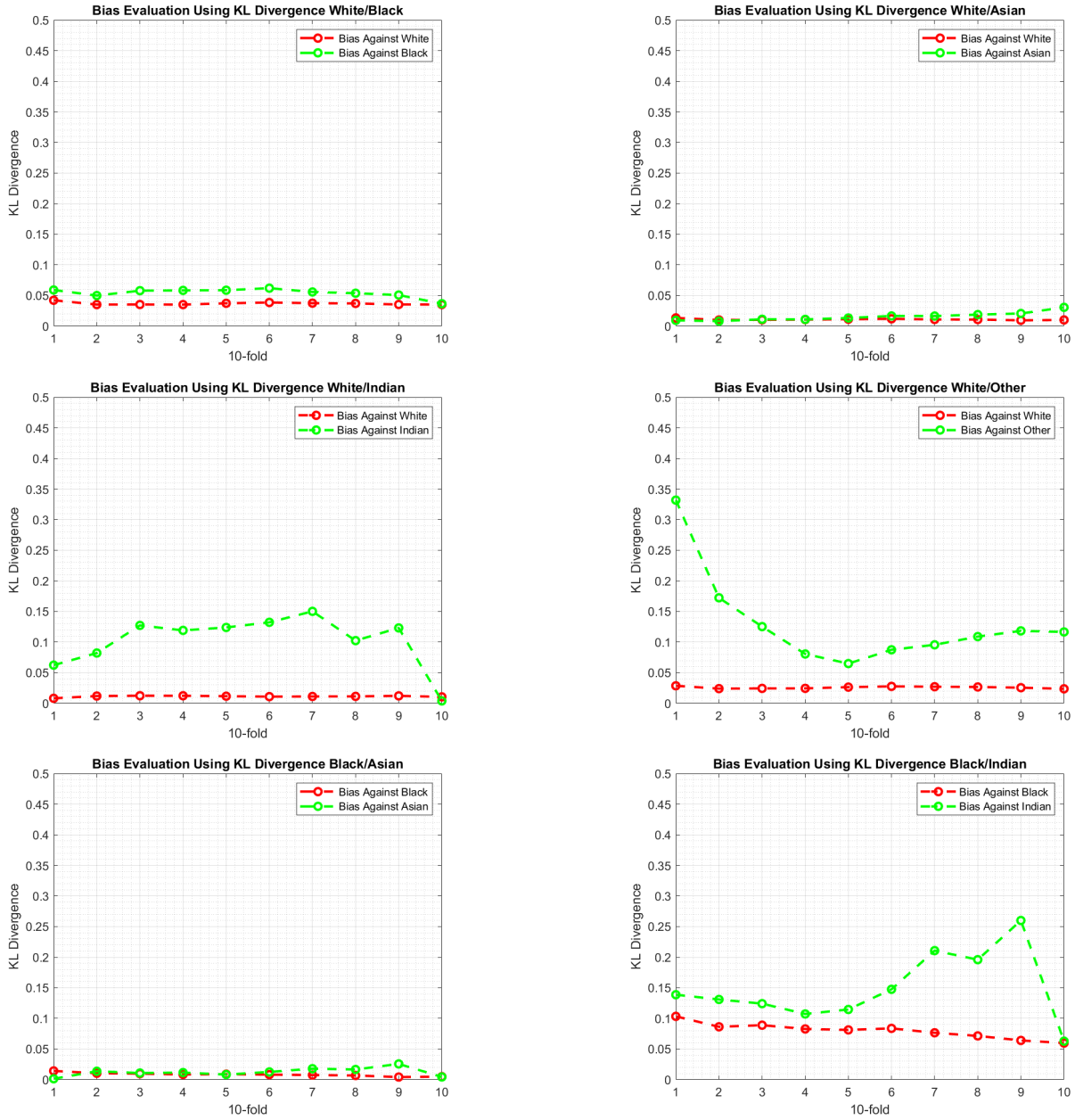

Figure 7. *Cont.*

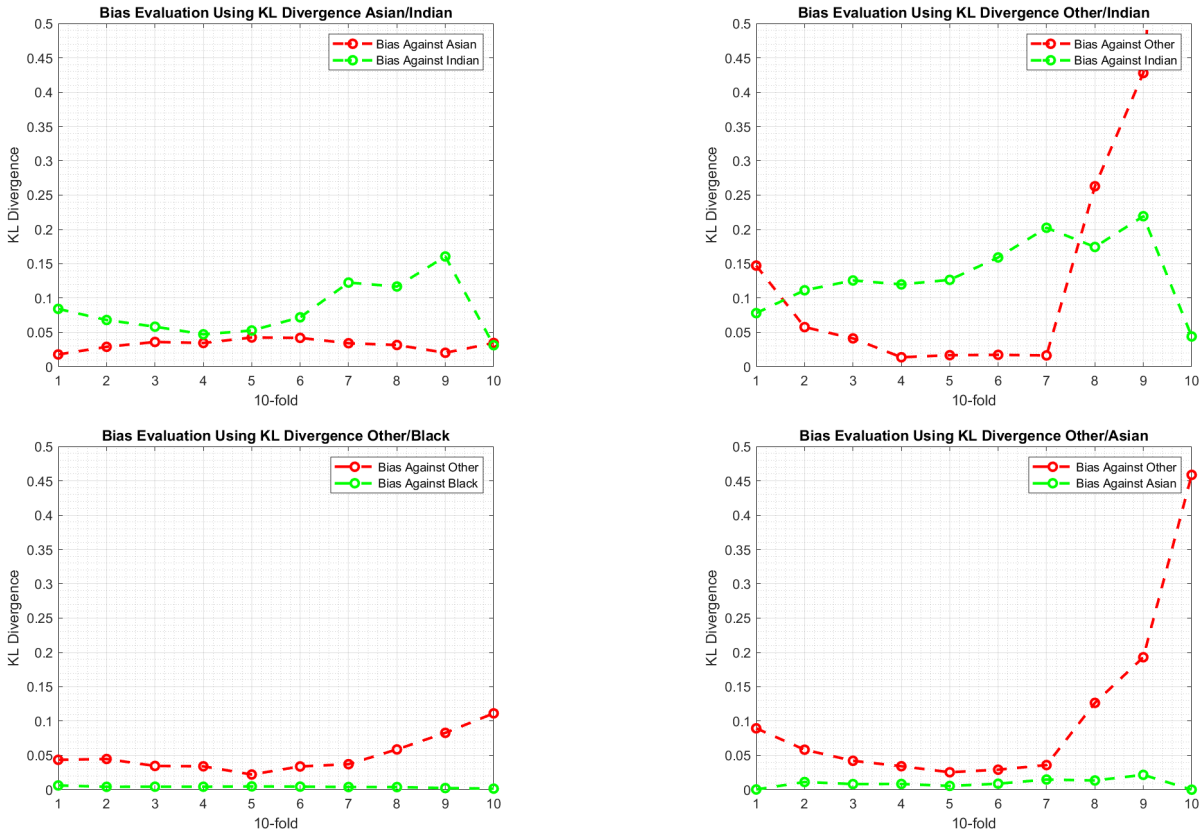

Figure 7. Bias evaluation between different races using KL-divergence.

7. Discussion

Terminology plays a significant role in understanding the context of the scientific domains. Since machine learning bias is a relatively contemporary field, it is essential to inaugurate this section by discussing the terminology.

According to the literature, as we mentioned in Section 2, the terms bias and unfairness are used interchangeably. Nevertheless, we believe bias does not imply unfairness. Bias, in a machine learning model, can be seen as underlying data characteristics inherited from human behavior and practice. The learning models expose bias to the decisions by extracting patterns and hidden relations from the data. While unfairness in a machine learning model is generated by intentionally and prejudicially tuning the model's parameters to satisfy human beliefs or desires: gender, racial, and/or social equality, in this context. In the real world, unfairness can be introduced by altering the hiring criteria to prefer a certain group over another. The criteria in machine learning can be seen as the algorithmic parameters. Thus, unfairness is introduced to the algorithm, while bias naturally exists in data. We are not trying to lessen machine learning bias. We believe it is not desirable, in most cases at least. However, it should be distinguished from unfairness. While unfairness is a very dangerous characteristic of any algorithm, bias is not necessarily as dangerous. We definitely need to prevent and fight unfairness. However, we should be careful when trying to mitigate bias due to the consequences that might appear. We would like to eliminate bias, not mitigate it.

8. Concluding Remarks and Future Work

In this paper, we have proposed a novel technique for the detection and evaluation of potential machine learning bias. We argue that data is biased in nature, reflecting the cognitive bias of human brains, which we have shown by understanding the model's bias and the role of training data on it.

We detect bias by alternating values for PBAs. The attribute that dramatically changes predicted class values after applying the alternation function is a biased attribute. Then, we evaluate the amount of bias by calculating the divergence between the original and the alternated mean of predicted class values with respect to each attribute's value. We can consider this step as a necessary preprocessing step in real-world problems for better model understanding and interpretability. It is worth noting that we need to pay attention to the *curse-of-dimensionality* problem in some datasets. This could degrade the efficiency of the model building.

We conducted an experiment using a publicly available dataset that contains gender and race attributes, which are considered to be prone to bias. We discovered that there is a bias against females in terms of income. The average KL-divergence over all folds was found to be around 0.9, which indicates relatively large divergence. Furthermore, the race attribute showed some bias from some races to others, as we discussed in the findings section. The results show considerable variation in terms of predicted wage when it comes to alternating female to male. The average predicted wage for females, for instance, is around 900 USD where it becomes 1180 USD when alternating to male. On the contrary, the male predicted wages average reduced from around 1150 to 850 USD when alternating male to female. We can conclude that the machine learning model might be biased against females when predicting wages.

The results obtained by our models open new research questions. For the sake of argument, assume the results were completely opposite by chance. Thus, if females have a higher income prediction, can we still call it a bias? Is bias in our minds associated with inequality, unfairness, or simply when females get less benefits than males in some circumstances? Are we looking for equal benefits for all groups? Are we trying to minimize the losses for the least advantaged? Should machine learning algorithms succeed in creating egalitarianism where humans failed? It is important to understand the nature of bias and the differences between bias and unfairness. Datasets contain bias because they reflect human behavior, practice, experience, and actions. Machine learning models are inclined to be biased due to the bias in the training datasets, yet it is important to detect bias.

In the future, we will investigate machine learning bias in depth and distinguish it from the relevance attribute. Furthermore, we will build machine learning techniques, including feature selection, classification, and clustering techniques, that take into consideration the alternation function and the divergence between attributes' values. Additionally, it is important to study the bias from an algorithmic perspective since we might find some algorithms that prefer bias in some ways. In addition, it is our goal to make interpretable models that are able to justify any existing bias.

Funding: This research was funded by The Deanship of Scientific Research at King Khalid University grant number (P.G.P2/100/41).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is publicly available on UCI Machine Learning Repository.

Acknowledgments: This work would not have been possible without the financial support King Khalid University. I would like to express my deepest gratitude to their generous support.

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Angwin, J.; Larson, J.; Larson, S.M.; Kirchner, L. Machine Bias. ProPublica. 2016. Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed on 22 March 2021).
2. Castro, C. What's Wrong with Machine Bias. *Ergo Open Access J. Philos.* **2019**, *6*. [CrossRef]

3. Corbett-Davies, S.; Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv* **2018**, arXiv:1808.00023.
4. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. [CrossRef]
5. Gianfrancesco, M.A.; Tamang, S.; Yazdany, J.; Schmajuk, G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern. Med.* **2018**, *178*, 1544–1547. [CrossRef] [PubMed]
6. Hilbert, M. Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychol. Bull.* **2012**, *138*, 211. [CrossRef]
7. Haselton, M.G.; Nettle, D.; Murray, D.R. The evolution of cognitive bias. In *The Handbook of Evolutionary Psychology*; John Wiley & Sons: Hoboken, NJ, USA, 2015; pp. 1–20.
8. Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.W.; Wang, W.Y. Mitigating gender bias in natural language processing: Literature review. *arXiv* **2019**, arXiv:1906.08976.
9. Amini, A.; Soleimany, A.P.; Schwarting, W.; Bhatia, S.N.; Rus, D. Uncovering and mitigating algorithmic bias through learned latent structure. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019; pp. 289–295.
10. Leavy, S. Uncovering gender bias in newspaper coverage of Irish politicians using machine learning. *Digit. Scholarsh. Humanit.* **2019**, *34*, 48–63. [CrossRef]
11. Prates, M.O.; Avelar, P.H.; Lamb, L.C. Assessing gender bias in machine translation: A case study with google translate. *Neural Comput. Appl.* **2019**, 1–19. [CrossRef]
12. Turing, A.M. Computing Machinery and Intelligence. *Creat. Comput.* **1980**, *6*, 44–53.
13. Cowgill, B.; Dell'Acqua, F.; Deng, S.; Hsu, D.; Verma, N.; Chaintreau, A. Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics. In Proceedings of the 21st ACM Conference on Economics and Computation, Virtual Event, 13–17 July 2020; pp. 679–681.
14. Harris, C.G. Methods to Evaluate Temporal Cognitive Biases in Machine Learning Prediction Models. In Proceedings of the WWW '20: Companion Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020; pp. 572–575.
15. Mehta, P.; Bukov, M.; Wang, C.H.; Day, A.G.; Richardson, C.; Fisher, C.K.; Schwab, D.J. A high-bias, low-variance introduction to machine learning for physicists. *Phys. Rep.* **2019**, *810*, 1–124. [CrossRef]
16. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 3315–3323
17. Jiang, H.; Nachum, O. Identifying and correcting label bias in machine learning. In Proceedings of the International Conference on Artificial Intelligence and Statistics (PMLR), Virtual Event, 26–28 August 2020; pp. 702–712.
18. Sun, W.; Nasraoui, O.; Shafto, P. Evolution and impact of bias in human and machine learning algorithm interaction. *PLoS ONE* **2020**, *15*, e0235502. [CrossRef]
19. Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; Wallach, H. A reductions approach to fair classification. In Proceedings of the International Conference on Machine Learning (PMLR), Stockholm, Sweden, 10–15 July 2018; pp. 60–69.
20. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *arXiv* **2019**, arXiv:1908.09635.
21. Rudin, C.; Wang, C.; Coker, B. The age of secrecy and unfairness in recidivism prediction. *arXiv* **2018**, arXiv:1811.00731.
22. Belkin, M.; Hsu, D.; Ma, S.; Mandal, S. Reconciling modern machine learning and the bias-variance trade-off. *arXiv* **2018**, arXiv:1812.11118.
23. Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; Vasserman, L. Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; pp. 67–73.
24. De-Arteaga, M.; Romanov, A.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; Kalai, A.T. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 120–128.
25. Leavy, S. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In Proceedings of the 1st International Workshop on Gender Equality in Software Engineering, Gothenburg, Sweden, 28 May 2018; pp. 14–16.
26. Bellamy, R.K.; Dey, K.; Hind, M.; Hoffman, S.C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv* **2018**, arXiv:1810.01943.
27. Rajkomar, A.; Hardt, M.; Howell, M.D.; Corrado, G.; Chin, M.H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **2018**, *169*, 866–872. [CrossRef] [PubMed]
28. Zhang, B.H.; Lemoine, B.; Mitchell, M. Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; pp. 335–340.
29. Crowley, R.J.; Tan, Y.J.; Ioannidis, J.P. Empirical assessment of bias in machine learning diagnostic test accuracy studies. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 1092–1101. [CrossRef]
30. Lötsch, J.; Ultsch, A. Current projection methods-induced biases at subgroup detection for machine-learning based data-analysis of biomedical data. *Int. J. Mol. Sci.* **2020**, *21*, 79. [CrossRef] [PubMed]

31. Tversky, A.; Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science* **1974**, *185*, 1124–1131. [CrossRef]
32. McCradden, M.D.; Joshi, S.; Anderson, J.A.; Mazwi, M.; Goldenberg, A.; Zlotnik Shaul, R. Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 2024–2027. [CrossRef] [PubMed]
33. Yapo, A.; Weiss, J. Ethical implications of bias in machine learning. In Proceedings of the 51st Hawaii International Conference on System Sciences, Hilton Waikoloa Village, HI, USA, 3–6 January 2018
34. Barbosa, N.M.; Chen, M. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–12.
35. Barbu, A.; Mayo, D.; Alverio, J.; Luo, W.; Wang, C.; Gutfreund, D.; Tenenbaum, J.; Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 9453–9463
36. Challen, R.; Denny, J.; Pitt, M.; Gompels, L.; Edwards, T.; Tsaneva-Atanasova, K. Artificial intelligence, bias and clinical safety. *BMJ Qual. Saf.* **2019**, *28*, 231–237. [CrossRef] [PubMed]
37. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]