JOURNAL OF
EMERGING INVESTIGATORS

Article

Addressing and Resolving Biases in Artificial Intelligence

Akshat Choudhari[1], Niren Choudhari[2]
[1]Oak Grove High School, San Jose, California
[2]Cisco, San Jose, California

SUMMARY

This paper hopes to determine what are the best strategies through which programmers can reduce bias in artificial intelligence (AI). Alongside this goal, the paper covers fairly introductory content and can be helpful for beginners to start their journey in AI and utilize its powers. To determine the best strategies to mitigate bias, we looked at diverse datasets, hyperparameter optimization, in-processing techniques, and post processing techniques which are all used in the industry. The results that we got were particularly favoring diverse datasets and hyperparameter optimization which means that adjusting weights and the initial data of a model have the biggest impact on accuracy rather than adjusting the final outputs. The best overall strategy was keeping a diverse dataset and allowing the algorithm to independently set its weights based on the best model possible. These factors show that the setup for your model is just as important if not more than actually being fed an output and adjusting the answer through a series of processes like in-processing or post-processing.

INTRODUCTION

With the rapid growth of artificial intelligence (AI) and its integration into various sectors, bias in AI models has emerged as a significant concern. Biases can stem from multiple sources, primarily data and algorithms, and can lead to skewed, unfair, or unjust outcomes. This paper provides an overview of the origins and implications of bias in AI and offers comprehensive strategies to mitigate these biases. Biases in AI can best be reduced by modifying data sets to account for a wide variety of people based on their race/origin, gender, and age.

AI models, especially those in decision-making roles, have shown evidence of biases, leading to discussions on their fairness and ethics (1). From recruitment tools to criminal justice applications, the implications of bias are vast and varied. Biases can affect these results in many ways and how people live their lives, whether by providing fake news or instilling hidden biases into customers. As a result, biases must be reduced in AI. We undertook a comprehensive evaluation of different methods consisting of diverse datasets, hyperparameter optimization, in-processing techniques, and post-processing techniques to reduce bias in our algorithms. By implementing these methods in consistent scenarios, we ensured a fair basis for comparison. While each method we explored bolstered the diversity of outcomes, the impact on accuracy varied. In subsequent discussions, we'll delve into the specific outcomes and improvements each method brought.

Models learn from historical data. If these datasets contain societal biases or do not represent the whole population, the AI will inherit these biases. A common example could be the bias that most nurses are women while most programmers are men. However, some women are programmers, and some men are nurses. The job of an AI program is to pick up these trends and output them to the user. As a result, a lot of the datasets get blended in, and the model assumes career profiles based on gender. Amazon's recruitment algorithm, which ignored a large majority of women for software programming roles, is a large-scale example of this bias (1). Biases based on other demographic factors such as race and age, could similarly occur in an algorithm such as Amazon's recruitment algorithm (1). The project was canned in less than a year since the algorithm discriminated against women for technical jobs such as software engineers. The root cause for this was found to be that men kept the large majority of technical jobs, and this information was fed into the algorithm. As a result, it hired men at a much higher rate compared to their women counterparts. This is a real-life example that we could relate to in our lab experiment. Luckily, there are many ways to prevent these societal biases in our data which will be discussed later in the paper.

AI-driven tools are used to interview and screen job seekers, many of which pose enormous risks for discrimination against people with disabilities and other protected groups (2). Rather than help eliminate discriminatory practices, AI has worsened them — hampering the security of marginalized groups that have long dealt with systemic discrimination. Facebook's algorithm, utilized for targeted advertising, faced scrutiny from the US Department of Housing and Urban Development due to its bias towards certain demographic groups, leading to legal action (6). Research from Northeastern University revealed that the algorithm disproportionately directed housing and employment ads toward white and Asian individuals while predominantly delivering other content to Black demographic groups, highlighting the significant impact of algorithmic biases on opportunities and the relevance of addressing such issues (6).

Another group of researchers did a similar topic on biases in AI (2). The paper mainly focused on the different methods through which the researchers could accurately determine the bias in an algorithm. However, they failed to address how we could reduce the biases. Realizing this gap in their research about Artificial Intelligence, we, therefore, decided to highlight some of the most popular methods like diverse datasets, hyperparameter optimization, in-processing techniques, and post-processing techniques for minimizing

DOI: https://doi.org/10.59720/23-249

bias in the industry and do a side-by-side comparison on how much of an impact each one made for the algorithm. The topic we wish to address is how to optimally reduce error and bias in our algorithms. While there are many different ways to determine the amount of bias in the algorithm, we used a traditional method where we compared the number of correct results to the total results. The next step was to research a variety of methods that data scientists utilize to improve the accuracy of their models. According to an online book, some of the most prominent methods of optimizing an algorithm were using diverse datasets, hyperparameter optimization, in-processing techniques, and post-processing techniques (4).

The overall purpose of this study was to help beginners in AI understand how to effectively achieve results that make their models more accurate. We hypothesized that organizing training datasets to be diverse and varied is the best strategy to achieve optimal results for the accuracy_score method. The accuracy_score method scores results by dividing the number of correct results by the total number of results. It is important to acknowledge that there can be various other tools used to determine different types of accuracy; however, for simplicity's sake, we used the accuracy_score method.

RESULTS

In the experiment, we utilized a dataset of information about a person's age, hours they worked, and gender. Based on that, we determined whether a person was making more than a $50,000 salary. We tested this scenario with different methods of minimizing bias to determine which method yielded the highest accuracy.

Hyperparameter optimization is similar to how we normally train our data; however, we provide a list of possible values that our model tries out for weights and other constraints to determine under what said-values we get the best results . This means that the data trains itself by comparing each possible variation of the list we provided (**Figure 1**). Whichever result shows the best accuracy stays as the weight for the model. Then, for in-processing techniques, adversarial training can be used to ensure an algorithm does not rely too heavily on a potentially biased feature, like gender in our earlier example with the UC Irvine Adult dataset (8). However, it was important to note that ignoring one attribute did not necessarily get rid of the entire bias in the model and only had a minor impact.

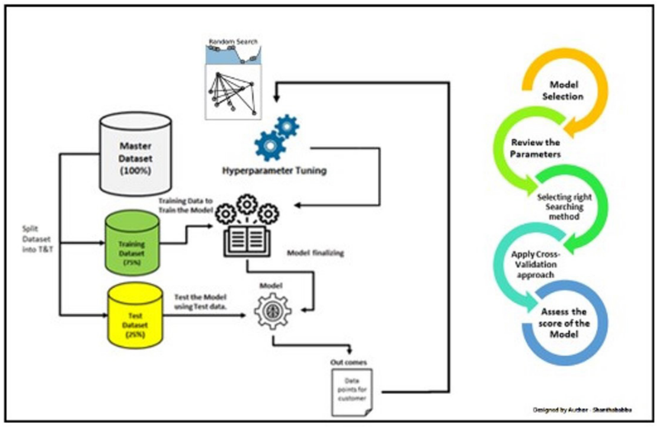

Figure 1: Method of Hyperparameter Optimization. Schematic of the process for hyperparameter optimization. The programmer provides their own list of desired values for certain parameters. Then, the program runs through them either using GridSearch or RandomSearch to find the one bringing the highest accuracy.

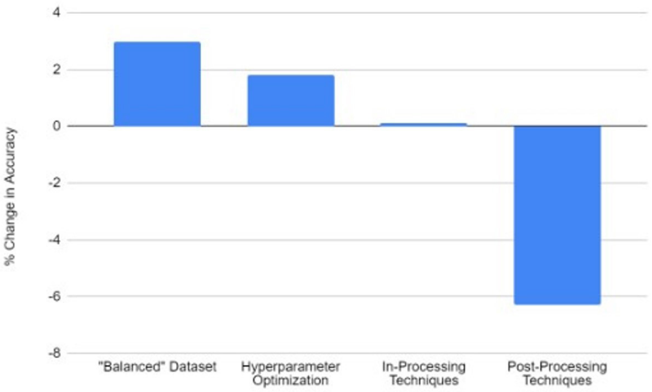

Figure 2: Percent Change in Accuracy. The percent change from the original algorithm to the new one utilizing one of the strategies shown on the x-axis. A side-by-side comparison of the data illustrates the difference in accuracy between the strategies. The largest positive is with a balanced dataset while the lowest is with post-processing techniques.

Initially, we conducted a statistical analysis to ascertain the magnitude of change necessary to influence our algorithm. Following numerous iterations of the default algorithm, we determined that alterations below 0.5% are inconsequential and may merely be attributed to random variation (**Figure 2**). The statistic at the bottom of every file in the Git Repository shows the percentage accuracy we received by running it 10 times and how all the results were within 0.5% of one another.

The algorithm estimated a person's salary correctly 79.1% of the time (**Figure 3**). We received this value with the information of the person's age, gender, and the hours they worked. Then, we processed the data by excluding any instances of women making over $50,000. As a result, our dataset was very skewed and the effects were seen in our algorithm's predictions. Compared to our original diverse dataset, the accuracy_score was 3 percent lower coming in at 76.2% which had the lowest accuracy_score for all the models tested (**Figures 2-3**).

For hyperparameter optimization, our data averaged 78.1% on the data with a default algorithm (**Figure 2**). After figuring out the accuracy of our default data with the pre-set weights, we optimized all our parameters such as n_estimators, max_depth, and min_samples_split by utilizing hyperparameter optimization. After rerunning the experiment with these new parameters, we found an increase of 1.8% in the accuracy_score from our default dataset (**Figure 2**). The data shows how the parameters such as weights in our algorithm need to be constantly tuned to improve accuracy.

We can see that in-processing techniques had a very small effect on improving the bias compared to previous methods as the accuracy only increased by 0.1% in comparison to the default test (**Figure 2**). This is smaller than the change of 0.5% meaning that this method was a fluke and had no real impact on the accuracy.

Post-processing techniques resulted in 72.8% accuracy which was lower than our default setting of 79.1% and showed that this method allowed for more equality among all groups of people (**Figure 3**). However, the issue was that when trying

DOI: https://doi.org/10.59720/23-249

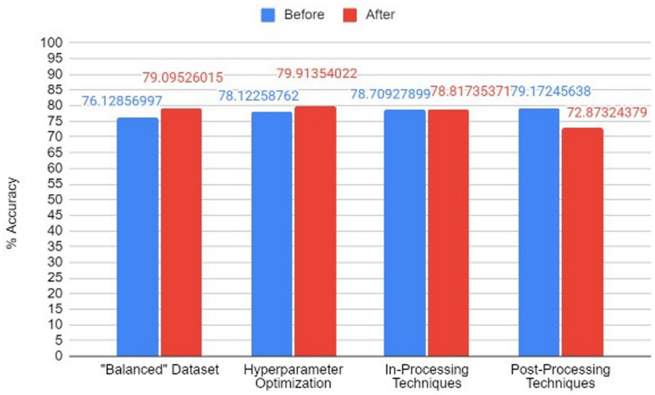

Figure 3: Percent Accuracy of each strategy. The percent of the original dataset in comparison to the new strategy's accuracy rate. The data allows us to compare the original dataset values with their new ones utilizing the strategy to show how much of an impact they individually had. The largest change we see between the original algorithm and one of the new strategies is with the post-processing technique and the least change is with the in-processing technique.

to achieve inclusiveness, the algorithm tended to forget the main objective: determining whether a person was making more or less than $50,000.

DISCUSSION

As we saw in the results section, filtering out our datasets was the most appropriate approach to getting the highest accuracy rate. However, that doesn't mean we can't implement all the other methods, such as hyperparameter optimization, in-processing techniques, and more while running our algorithm. Combining the forces of all these methods can ensure an even higher accuracy rate in our code. In the end, an algorithm can only achieve results within the scope of the provided data. In a perfect world, a data scientist would be able to access every detail of a person to accurately determine a result. However, that brings to question many other issues which might arise relating to privacy.

Many papers have found similar results in these types of experiments which were that "balanced" datasets provide the best results. One paper regarding this topic tests out various algorithms; however, there is a big flaw where they track the accuracy of an algorithm on subjective results (5). What that means is that their grading criteria for accuracy was a subjective task and not quantifiable. The emphasis on what is the best method can't be effectively proven with that form of research. As a result, we wished to follow up on their work, but using a methodology that could provide us with a True-or-False answer and achieve a more accurate understanding of the role that each method has.

The accuracy dip in our programs can largely be attributed to our consistent choice of model. Across all experiments, we exclusively employed the RandomForestClassifier. Some fallbacks in our study might have been the size of our data. Oftentimes, companies use much larger quantities of data in developing their algorithms compared to our smaller amount for the experiment. However, that should not have a significant effect on the findings of the study because that was one of the variables we kept consistent with all programs. We mitigated this pitfall in our program by making sure that no specific program would have the advantage over the other in terms of the data they were fed for training and unexpected variables were kept minimal. While this decision ensured uniformity, it's plausible that alternative models might have offered more optimal results. A lacking point in our study was not accounting for the other various methods used in the AI world to reduce biases. These other methods that may have yielded better results than a balanced dataset could be model selection, ethical guidelines, or fairness constraints.

With this newfound understanding, we could utilize some of these strategies in notable examples of recruitment tools such as Amazon's failed project in automating its hiring process (1). Furthermore, the same concepts can be applied to the decision-making in algorithms for Facebook where the datasets can be readjusted to not factor in race for deciding what gets post visibility (6). Similar to how our program determined whether people made $50,000 or not, these algorithms determine factors that can shape their company. The similarity between these two use cases of AI is that both algorithms had gender or racial biases. In our situation, the most optimal way of resolving the data was diversifying the data. The programmers at Amazon should have implemented a program where the representation of men and women in technical jobs had equal representation in the data. This would allow women to get equal representation and not be discriminated against in the automated hiring process. The same goes for Facebook where users who should be allowed to see the post should not have been determined on factors like race.

Bias in AI is a problem arising from various sources, from historical data to algorithm design. While the challenge is substantial, through rigorous efforts in data collection, algorithm design, and regular audits, biases can be effectively mitigated, paving the way for fair and trustworthy AI systems. The best way to ensure this is by teaching beginners from the start about how to reduce their biases in models to achieve better results.

Future possibilities of "purging" bias lie within quantum computing. Quantum computing's feature to store qubits allows quantum systems to process vast volumes of information concurrently. Moreover, the principle of entanglement, where qubits are interconnected in a manner that the state of one instantly influences the state of another, can enable multiple tasks to happen simultaneously. This is particularly advantageous for specific AI endeavors, such as optimization problems tackled via quantum annealing. As researchers combine the principles of neural networks with quantum mechanics, we witness the emergence of Quantum Neural Networks (QNNs), which aim to outpace their classical counterparts. To explore the potential of Quantum Computing more, you should visit various research papers where they discuss the potential with more detail and calculations (7).

MATERIALS AND METHODS

In the experiment, we ran a program that determined whether a person was making over $50,000 as a yearly salary based on factors like gender, age, and how many hours a week they worked. In the first scenario, our data had already been organized which was targeted to reduce overfitting and not discriminate based on factors like gender.

All of the programs provided in the Git Repository (https://github.com/A-Choudhari/highschoolresearch) were run on a Python Interpreter with some additional libraries that are

DOI: https://doi.org/10.59720/23-249

commonly utilized for Artificial Intelligence. These libraries are pandas, sklearn, torch, and numpy. All of the libraries mentioned there will be required to run the code and achieve the results shown in the experiment. The dataset used was the UCI Adult Dataset for all the programs which were found on Kaggle (https://www.kaggle.com/datasets/wenruliu/adult-income-dataset).

Hyperparameter optimization is a methodical process where different values are tested for hyperparameters, such as learning rate or regularization strength, to find the combination that results in the best performance of a machine learning model on a given dataset. This is typically done through techniques like grid search, which systematically explore the hyperparameter space to identify the optimal configuration. Grid search involves defining a grid of hyperparameter values to be evaluated exhaustively. Each combination of hyperparameters is tested, and the performance of the model is assessed using a predefined metric, such as accuracy or loss. This method can be computationally expensive, especially when dealing with a large number of hyperparameters or a wide range of values.

In adversarial training for fairness, a model (the primary classifier, e.g., neural network) is trained to predict the target variable (e.g., income). Another model (the adversary) is trained to predict the sensitive attribute (e.g., gender) from the output of the primary classifier. If the adversary can predict the sensitive attribute with high accuracy, it means the primary classifier's predictions are biased. In a training loop, the primary classifier tries to maximize its accuracy on the main task while minimizing the adversary's accuracy. This process ensures that the primary classifier's predictions do not carry information about the sensitive attribute.

Primary Classifier: A model (e.g., neural network) is trained to predict the target variable (e.g., income).

Adversary: Another model is trained to predict the sensitive attribute (e.g., gender) from the output of the primary classifier. If the adversary can predict the sensitive attribute with high accuracy, it means the primary classifier's predictions are biased.

Training Loop: During training, the primary classifier tries to maximize its accuracy on the main task while minimizing the adversary's accuracy. This process ensures that the primary classifier's predictions do not carry information about the sensitive attribute.

For post-processing techniques, we are setting different classification thresholds for each group to achieve equalized odds. These methods are beneficial as they can be applied without needing to retrain the model. One common post-processing method is adjusting classification thresholds to achieve a fairness criterion like equalized odds.

Train a Classifier: Train your preferred model on the training data without considering fairness.

Determine Thresholds for Equalized Odds: Compute the true positive rate (TPR) and false positive rate (FPR) for each protected group on a validation set. For each group, find the threshold that equalizes the TPR (or FPR) across groups.

Apply Thresholds to Test Data: Once you've found the thresholds for each group, you can adjust the model's predictions on the test data using these thresholds.

Received: September 28, 2023
Accepted: February 11, 2024
Published: May 29, 2024

REFERENCES

1. Goodman, Rachel. "Why Amazon's Automated Hiring Tool Discriminated against Women: ACLU." American Civil Liberties Union, 27 Feb. 2023, www.aclu.org/news/womens-rights/why-amazons-automated-hiring-tool-discriminated-against. Accessed 15 Aug. 2023.
2. Akselrod, Olga. "How Artificial Intelligence Can Deepen Racial and Economic Inequities: ACLU." American Civil Liberties Union, 3 July 2023, www.aclu.org/news/privacy-technology/how-artificial-intelligence-can-deepen-racial-and-economic-inequities. Accessed 14 July 2023.
3. Agarwal, Avinash, et al. "Fairness Score and Process Standardization: Framework for Fairness Certification in Artificial Intelligence Systems", 19 Jan. 2022, arxiv.org/pdf/2201.06952.pdf. Accessed 15 Aug. 2023.
4. Seni, Giovanni, and John Elder. Ensemble Methods in Data Mining: Improving Accuracy through Combining Predictions. Morgan and Claypool, 2010. https://doi.org/10.1007/978-3-031-01899-2.
5. Tomalin, Marcus, et al. "The Practical Ethics of Bias Reduction in Machine Translation: Why Domain Adaptation Is Better than Data Debiasing." Ethics and Information Technology, vol. 23, no. 3, 2021, pp. 419–433, https://doi.org/10.1007/s10676-021-09583-1.
6. "Roughly Six-in-Ten Americans Believe It Is Not Possible to Go through Daily Life without Having Their Data Collected." Pew Research Center: Internet, Science &amp; Tech, 12 Nov. 2019, www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/pi_2019-11-14_privacy_0-01/. Accessed 29 July 2023.
7. Mangini, S., et al. "Quantum Computing Models for Artificial Neural Networks." Europhysics Letters, vol. 134, no. 1, 2021, p. 10002, https://doi.org/10.1209/0295-5075/134/10002.
8. Omargowaily. "Adult Income Dataset by Omar Gowaily." Kaggle, 18 Apr. 2024, www.kaggle.com/code/omargowaily/adult-income-dataset-by-omar-gowaily. Accessed 25 Apr. 2024.