Andrew Bahsoun

Mitigating Bias in Machine Learning through In-processing Methods

1 Introduction

Artificial Intelligence has become increasingly popular in the last decade, and people from all different backgrounds have been using AI for downstream tasks [1]. Some examples include self-driving cars, healthcare and medical diagnosis, video surveillance, product and content recommendation, email filtering, chatbots, and various AI assistants [2]. These tools make products smarter and increase user satisfaction [3]. Ostensibly, we should continue to increase AI development and apply it to every facet of our lives, but there are reasons why this is dangerous. Many people who use AI don't know how it comes up with its prediction. The average user only acknowledges the output that the AI gives them, making it a "black box," which is when the internal workings of a system are hidden or not completely understood [4]. AI users tend to actually prefer this idea of a black box; users with less AI literacy are shown to have greater receptivity to AI [5]. This is where AI's dangers come in; people tend to trust AI just about the same, or even more than other humans, even though they don't know how the system came up with their answer [6]. The result of this is that we now have AI making critical decisions in areas like loan management, job applications, and criminal justice, but providing results to people who may not understand how the system works or if a fair decision has been made [7]. Bias AI models have recently made news by falsely predicting heart disease probabilities for lower-income groups, unfairly flagging black defendants to be twice as likely to be future criminals than white ones, assigning black women a higher chance of having postpartum depression, and many more [8], [9], [10]. Bias enters the AI model through biases that exist in datasets and models that are trained without bias mitigation in mind. My FIRE Project will be finding the methods that best address the bias in AI models. Researchers have defined the possible methods for reducing into three groups: Pre-processing, In-processing, and Post-processing [9]. Pre-processing is where the data itself is modified to reduce bias, and then the model is trained on unbiased data. Post-processing is when results are recalibrated after a prediction is given. In-processing is unique as it aims to reduce bias by directly affecting the model training. Research has pointed out that In-processing models are the most capable of reducing bias because bias is often a result of the algorithm and not the data; therefore, the only way to make the model fair is by changing the algorithm [9]. Also, it allows large pre-trained models to be tuned and reweighted without retraining the entire model, which takes a large amount of resources [9].

2 Methodology

MIMIC-IV, which consists of 6 different models: Hosp, ICU, ED, CXR, ECG, and Note [11], allows for our data to have 3 different modalities. The first is structured longitudinal measurements obtained from the Hosp and ICU models. The next is unstructured clinical note data from the Note module. Finally, the CXR module gives access to X-ray images for patients. To obtain a singlar form of data, we will need to fuse these datatypes together into a single representation. In FairEHR-CLP [12], a similar fusion was created using structured longitutional data and notes. This is possible with the PyTorch `torch.cat` function, and is used in FairEHR-CLP. The function concatenates the given sequence of tensors, giving an overall representation of multiple types of data. We will fuse a representation of the longitudinal data, the note data, and the image data. To extract the important features from the notes, we will use a pre-trained LLM model RoBERTa [13]. This model will tokenize the note data and generate context-aware vector embeddings for each token. This will create semantic and contextual nuances in the text. To extract important features from the X-Ray data, we will use a CNN.

We will employ a deep neural network (DNN) model because it can directly incorporate fairness constraints into the training process through loss-function regularization. Other models, like Decision Trees[14] have the ability to have constraints added to the loss function, but complexity can arise, like in the decision tree example by Aghaei et.al, where a mixed-integer linear program (MILP) is required with many constraints to guide data around the tree fairly.

The goal of a deep neural network is to increase accuracy by minimizing the loss function. So we can define the objective function $\mathrm{AP}(\theta)$ of our model as

$$\mathrm{AP}(\theta) = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \mathbf{1}\left\{h_\theta(x_i) = y_i\right\}$$

Where θ signifies the model parameters N the number of training examples, $h_\theta(x_i)$ the prediction made by the model, y_i the true label, and 1 the indicator, which is 1 if the statement is true, and 0 if false. In the next section, we will see how this function changes to account for fairness.

Now we have our objective in mind, we create the constraint optimization problem[15]

$$\max_\theta AP(\theta)$$
$$\text{subject to } |FP(\theta)| \leq \varepsilon$$

Where $\text{FP}(\theta) = f(h_\theta, g_1) - f(h_\theta, g_2)$. Which is essentially the difference between a fairness metric between two groups g_1 and g_2 based on the prediction h_θ. This value must be less than or equal than the user-defined limit of ϵ. However, this problem is hard to solve, since we are hard constraining the accuracy to ϵ. Trying to solve this problem can lead to a drop of accuracy in the accuracy function since it is forced to stay within the limit.

So, like shown in OmniFair and other Lagrange Multiplier literature [15], [16], we will transform the problem into a unconstrained optimization problem, which is a softer, easier to solve function.

$$\max_\theta AP(\theta) + \lambda \cdot FP(\theta)$$

Finally, we can optimize our function by finding the best value for λ and maximizing accuracy. We will use a hill-climbing algorithm, where values λ are updated, and fairness values are checked to find the λ that best optimizes the equation [15]. At the same time, we will use an Iterative Gradient-Based Training method that computes the fused embedding as a forward pass in each iteration. The loss will be computed, and a backpropagation step is performed to calculate the gradients and update the model parameters.

3 Expected Outcome

In all fairness model evaluation, some sort of fairness metric is used. There are too many to list all of them, but two of the most popular are Equalized Odds (EO) and Overall Accuracy Equality (OAE). EO measures if different sensitive groups (race, gender, age, etc..) have equal true positive (TPR) and false positive rates (FPR) [9]. OAE measures if accuracy among different groups is the same. Based on the previous literature review, we have seen that DfC [11] achieved an EO score of 4.9±0.6. This means the largest gap between TPR and FPR in all groups is around 5%. We will also aim for an EO of **5**. With OAE, we aim for an accuracy gap between the two groups to be a maximum of **4%**; this comes from the benchmark in FairEHR-CLP [12]. Finally, an F1 or accuracy score must be considered, as with any prediction model. Setting these goals will help us quantify our model success.

If our goals are met, and the model can achieve reduced levels of bias while maintaining a similar or improved F1 score, then we can be sure that we have created a model that can mitigate bias in a healthcare setting. Our code and model can be evaluated for future research or directly used for predictions.

4 Timeline and Deliverables

Given this project will take place for 11 weeks in the summer, I have created a timeline.

- **Weeks 1-3** will be for setup and preprocessing. I will finalize data cleaning and preprocessing for structured (Hosp, ICU), text-based (Note), and imaging (CXR) modules. I will implement CNN-based feature extraction for X-rays and create RoBERTa embeddings for clinical notes. Finally, I will train an initial (unconstrained) baseline model and record performance/fairness metrics.

- In **Weeks 4-7** I will implement and tune the model. I will implement the Lagrangian into the baseline model and iteratively test the best values for λ. I will tune hyperparameters such as learning rate, batch size, network depth on a validation set, and finally document the trade-offs between accuracy and fairness for different constraint settings.

- In **Weeks 8–11** I will finalize the best model that meets fairness goals while maintaining acceptable accuracy. Then, I will compile my findings into a final report with methodology, results, and discussion and release my source code in preparation for my showcase.

5 Qualifications

This research project is feasible due to my knowledge of the subject, the preprocessed EHR dataset, my knowledge of Python and ML, and my personal motivation. The MIMIC-IV[11] dataset consists of 6 different models: Hosp, ICU, ED, CXR, ECG, and Note. This allows for our data to have 3 different modalities. The first is structured longitudinal measurements obtained from the Hosp and ICU models. The next is unstructured clinical note data from the Note module. Finally, the CXR module gives access to X-ray images for patients.

I have been working alongside Dr. Yuxin Wen since last September. During this time, we spent time doing research on machine learning bias. Next, my experience in Python and ML (Machine Learning) has been highlighted by various activities at Chapman. The first was my GCI project, which utilized Python and Machine Learning to create a hand-washing recognition AI model for medical supplies company LayerJot. This experience honed my Python and ML skills since I used PyTorch to create the model. The project ended up winning first place in the

GCI Symposium. Last summer, I assisted Dr. Wen in teaching workshops on ML and Python to local high school students. We taught students how to create simple AI models and explained how neural networks work. This experience gave me a deeper understanding of ML since I had to learn to explain it to an audience who had never used it before. Finally, I am personally motivated to complete this project because mitigating bias will allow for the safer use of AI. I also want to improve my machine learning and research skills, which will be useful to my EECS Masters Thesis and my PhD in the future.

This is also an important project to take on since there is a research disparity between efforts to improve AI models and efforts to address fairness in AI models. It is important that we build intelligent AI models, but we must slow down and make sure they are safe.

References

[1] Rodney C. Richie. Basics of artificial intelligence (ai) modeling. *Journal of Insurance Medicine*, 51(1):35–40, 2024.

[2] Jitendra Parmar, Satyendra Chouhan, Vaskar Raychoudhury, and Santosh Rathore. Open-world machine learning: Applications, challenges, and opportunities. *ACM Computing Surveys*, 55(10):205, 2023.

[3] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. Automatic online evaluation of intelligent assistants. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*, pages 506–516, 2015.

[4] W.J. Von Eschenbach. Transparency and the black box problem: Why we do not trust ai. *Philosophy & Technology*, 34:1607–1622, 2021.

[5] S. Tully, C. Longoni, and G. Appel. Express: Lower artificial intelligence literacy predicts greater ai receptivity. *Journal of Marketing*, 0:ja, 2025.

[6] Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*, pages 763–777, 2022.

[7] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 17(3):35, 2023.

[8] Yikuan Li, Hanyin Wang, and Yuan Luo. Improving fairness in the prediction of heart failure length of stay and mortality by integrating social determinants of health. *Circulation: Heart Failure*, 2022.

[9] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. In-processing modeling techniques for machine learning fairness: A survey. *ACM Trans. Knowl. Discov. Data*, 17(3), March 2023.

[10] Y. Park, J. Hu, M. Singh, and et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA Network Open*, 4(4):e213909, 2021.

[11] A. Johnson, L. Bulgarelli, T. Pollard, B. Gow, B. Moody, S. Horng, L. A. Celi, and R. Mark. Mimic-iv (version 3.1), 2024.

[12] Yuqing et al. Wang. Fairehr-clp: Towards fairness-aware clinical predictions with contrastive learning in multimodal electronic health records, 2024.

[13] Yinhan et al. Liu. Roberta: A robustly optimized bert pretraining approach, 2019.

[14] S. Aghaei, M. J. Azizi, and P. Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of AAAI Conference on Artificial Intelligence*, volume 33, pages 1418–1426, 2019.

[15] Hantian Zhang, Xu Chu, Abolfazl Asudeh, and Shamkant B. Navathe. Omnifair: A declarative system for model-agnostic group fairness in machine learning. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21)*, pages 2076–2088, 2021.

[16] Google. *Constrained Optimization and Lagrange Multiplier Methods.* Google Books, 2025.