Mitigating Bias in Machine Learning through In-processing Methods

## I: Introduction

Artificial Intelligence has become increasingly popular in the last decade, and people from all different backgrounds have been using AI for downstream tasks [6]. Some examples include self-driving cars, healthcare and medical diagnosis, video surveillance, product and content recommendation, email filtering, chatbots, and various AI assistants [1]. These tools make products smarter and increase user satisfaction [2]. They can also save lives and reduce the time it takes someone to finish a task, increasing productivity. 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 [3]. 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 [4].

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], [7], [9]. Bias enters the AI model through biases that exist in datasets, and models that are trained without bias mitigation in mind. My SURF 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 [7]. 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 [7]. Also, it allows large pre-trained models to be tuned and reweighted without retraining the entire model, which takes a large amount of resources [7].

## III: Goals

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) [7]. 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 [11]. Finally, an F1 or accuracy score must be considered, as with any prediction model. We hope to

achieve an accuracy score of **75-80%**. This number is common for similar healthcare prediction models [11].

Setting these goals will help us quantify our model successes and benchmark our model in each iteration, which will be discussed in the Methodology.

**IV: Methodology & Schedule**

Week 1 & 2: Model Testing and Evaluation

The first part of the Methodology is to start researching existing models and testing our dataset on them. We already have a few in mind, such as FairEHR-CLP [11]. After testing and gathering our results, we will analyze fairness ratings. These metrics will be important for comparing our final model to existing models. We plan to test 4 existing models. Specifically, I will read papers addressing fairness in ML using In-Processing methods containing an open-source code repository. From there, I will download the code and replace the data input and format with our own. Then, I will run the code on our existing Linux server.

Week 3-6: Creating our Model:

To create our final model, we will use an iterative development methodology. To start, we will create a baseline model. We will improve our previous model each week, evaluating our code and what parts could be causing biases or inaccuracies. One way we could do this is to manually evaluate the model weights and see which features are being ranked as more important than others. From there, we can manually change the weights or retrain the model and reward/punish the model based on whether we believe a certain feature should be ranked higher/lower. We can look at various types of methods for mitigating bias in our model. The first is using a regularization $L(D;\theta) + \lambda||\theta||_2^2 + nR(D;\theta)$ or constraint optimization $min\, L(D;\theta) + \lambda||\theta|_2^2 \;\; s.t\, \Omega(D;\theta) < 0$ design. The former decreases the mutual information between the sensitive group and predictions by the penalty term nR(D;0). In the latter, a fairness constraint is applied to force the model to train according to the limits of the fairness metric. Both functions are essentially doing the same thing; they are not allowing the sensitive attribute to influence a decision or prediction being created. However, the latter is a hard constraint, and the model cannot train past the constraint, while the former is a soft penalty. The previous two methods involve changing the loss function itself. However, there are also methods such as Conditional Contrastive Learning (CCL) with a representative learning approach, which can find similarities in subjects without looking at their sensitive features.

Week 7-8: Analysis and Presentation:

Once our model is completed, we can analyze it in the last two weeks. In the first week, I will use matplotlib and other visualization tools like Seaborn to create visually pleasing graphs to present. I can do this by running a series of validation patients to determine how well the model can predict and how fair it is. With this data, I will reevaluate the project's goals and determine where I succeeded and where my shortcomings were. In the last week, I will create the poster and prepare to present my work.

**V: Expected Outcomes**

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.

## VI: Qualification

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.

I have been working alongside Dr. Yuxin Wen since last September; since then, we have already started progressing on this project. When we started, neither Dr. Wen nor I knew much about fairness in machine learning, so in the first 3 months, We spent the time reading, researching papers, and discussing them. This is how we concluded that In-Processing was the most optimal method. This process has given Dr. Wen and I the necessary context and expertise to reach our goals with this project. We also have access to an EHR (Electronic Health Record) dataset called MIMIC-IV. This vast dataset contains 265,000 health records of patients admitted into Beth Israel Deaconess Medical Center in Boston, MA [10]. In the last month, I have cleaned and preprocessed a dataset that can be used to train our model. 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. Finally, this project has a methodology that shows it can be completed in an 8-week period.

## VII: Bibliography


1. Jitendra Parmar, Satyendra Chouhan, Vaskar Raychoudhury, and Santosh Rathore. 2023. Open-world Machine Learning: Applications, Challenges, and Opportunities. ACM Comput. Surv. 55, 10, Article 205 (October 2023), 37 pages. https://doi.org/10.1145/3561381
2. Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic Online Evaluation of Intelligent Assistants. In Proceedings of the 24th International Conference on World Wide Web (WWW '15). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 506–516. https://doi.org/10.1145/2736277.2741669
3. Von Eschenbach, W.J. Transparency and the Black Box Problem: Why We Do Not Trust AI. Philos. Technol. 34, 1607–1622 (2021). https://doi.org/10.1007/s13347-021-00477-0
4. Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. 2022. 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). Association for Computing Machinery, New York, NY, USA, 763–777. https://doi.org/10.1145/3514094.3534150
5. Tully, S., Longoni, C., & Appel, G. (2025). EXPRESS: Lower Artificial Intelligence Literacy Predicts Greater AI Receptivity. Journal of Marketing, 0(ja). https://doi.org/10.1177/00222429251314491
6. Rodney C. Richie; Basics of Artificial Intelligence (AI) Modeling. J Insur Med 1 July 2024; 51 (1): 35–40. doi: https://doi.org/10.17849/insm-51-1-35-40.1
7. Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. 2023. In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. ACM Trans. Knowl. Discov. Data 17, 3, Article 35 (April 2023), 27 pages. https://doi.org/10.1145/3551390
8. Yikuan Li, Hanyin Wang, Yuan Luo. Improving Fairness in the Prediction of Heart Failure Length of Stay and Mortality by Integrating Social Determinants of Health, Journal Article, 2022. https://doi.org/10.1161/CIRCHEARTFAILURE.122.009473
9. Park Y, Hu J, Singh M, et al. Comparison of Methods to Reduce Bias From Clinical Prediction Models of Postpartum Depression. JAMA Netw Open. 2021;4(4):e213909. doi:10.1001/jamanetworkopen.2021.3909
10. Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., & Mark, R. (2024). MIMIC-IV (version 3.1). PhysioNet. https://doi.org/10.13026/kpb9-mt58.
11. Yuqing Wang, Malvika Pillai, Yun Zhao, Catherine Curtin, and Tina Hernandez-Boussard.2024. FairEHR-CLP: Towards Fairness-Aware Clinical Predictions with Contrastive Learning in Multimodal Electronic Health Records. https://arxiv.org/abs/2402.00955