Capstone Progress Report

Last updated: February 27, 2025

Instructions: Each capstone team can use this template to capture and summarize progress. This can be shared with the sponsor and mentor. When submitting the plan during the course, a PDF file is preferable.

Stakeholder Names and Roles

Stakeholder	Role
Becky Desrosiers	*Team member*
Abner Casillas-Colon	*Team member*
Naomi Ohashi	*Team member*
George Shoriz	*Team member*
Phillip Waggoner	*Mentor*
Emanuel Moss	*Sponsor*
Elizabeth Watkins	*Sponsor*
Dawn Nafus	*Sponsor*

Project Title: Bias Evaluation in Open Model Platform

Project GitHub repository: https://github.com/oatmeelsquares/BiasOMZ

Abstract

This project seeks to evaluate one model in Intel Labs' Open Model Zoo for potential bias against protected characteristic(s). The team will start by researching bias metrics, identifying useful datasets, and choosing a model that will be feasible to evaluate, with appropriate metrics. Finally, the model's inference on the datasets will be evaluated using the chosen bias metrics.

Outline of the Project

Intel's Open Model Zoo (OMZ) is an offering that allows regular people to use pretrained AI models for their own purposes. This project is important because AI models can easily have bias inherent to their training, which can impact their performance and their impact on society, depending on for what purpose they are leveraged. Since the OMZ is publicly available, the potential for applications is expansive and, in turn, so are the potential for consequences from biased training. Stakeholders could include anyone who employs the model, or anyone who could be affected by myriad applications of the model.

The scope of this project is one model and finding a metric or set of metrics that can quantify the bias in the model's training. We assume that we start with a possibly biased, pretrained model. A stretch goal of the project will be to create a pipeline that will facilitate future bias detection in other models.

Success Criteria

SC1	Summary of existing bias metrics/evaluation methods for AI models
SC2	Identified dataset(s) with ground truth that can be used with the chosen model
SC3	One model summarized in terms of chosen bias metrics
SC4	(Stretch goal) Initial development of testing pipeline to follow the lead of this project

Progress report

SC1 – we have completed our research and assembled it into organized notes, ready to be written up into the Lit Review. Though we hoped to have the Lit Review drafted for this report, we are ahead on getting our data loaded into a jupyter environment (evidenced by our EDA below).

SC2 – we have identified a model to test and a reasonable dataset

We have decided (pending sponsor review/approval) on the metrics to use for our evaluation and the model to test for bias. These decisions will be confirmed or adjusted by our sponsors in our meeting on Monday, March 3.

SC3 & SC4 will be tackled in the second half!

Model: face-detection-0200

We propose to go with a face detection model because it seems like a straightforward starting point and one that is more likely to suffer from bias based on a protected characteristic like, race, gender, and age. Once we looked at the FairFace dataset, it confirmed our choice.

Dataset: FairFace (described below in Section **Summary of the Data**)

We propose to use the FairFace dataset because it works for the model we're testing and it comes pre-labelled with age, gender, and face data. It is big enough to get generalized statistics, and a subset that is uniformly distributed across gender and race, which may be helpful to our metrics. This dataset will also be useful for future testing on facial recognition models.

Proposed Metrics:

- **Equal Parity** – ensure every group is represented equally in the dataset
- **False Positive Rate Parity** – same false positive rate between groups, important when the model is used for punitive measures
- **False Discovery Rate Parity** – same false discovery rate between groups, important when the model is used for punitive measures
- **False Negative Rate Parity** – same false negative rate between groups, important when the model is used for benefitting people
- **False Omission Rate Parity** – same false omission rate between groups, important when the model is used for benefitting people
- **AUC** – same model performance between groups, useful for overall fairness
- **Qualitative analysis***

We consider an 80% default disparity intolerance for all qualitative, to be discussed with the sponsor. This should vary based on the metric applied and based on the implementation of the model.

*Time permitting, we may use curated datasets and/or heatmapping of pixels to investigate how the model makes its decisions, which may give insight into the "psychology" of the model, and if it holds biases.

Data Assumptions and Limitations

For any project, there may be assumptions [A] *and limitations* [L] *on the data and the modeling approach. These can be documented here. Example:* [L] *Ideally, the dataset would include variable X, but we did not have access to this data, which was a limitation.*

Identifier	Description
[A] FairFace	The FairFace Dataset is supposed to be balanced in terms of race, gender, and age. We found that a partial subset of the data is balanced by race and gender, but there is a nonuniform distribution of ages.

Summary of the Data

We will be using data that is publicly available online. There will be no need to store it long-term.

We recognize there are three main biases around ethical concerns in facial recognition technology. We plan to focus on the race bias (target variable) in the first round, followed by potentially age and gender.

1. Racial bias towards certain races
2. Gender bias towards certain genders
3. Age bias towards certain age groups

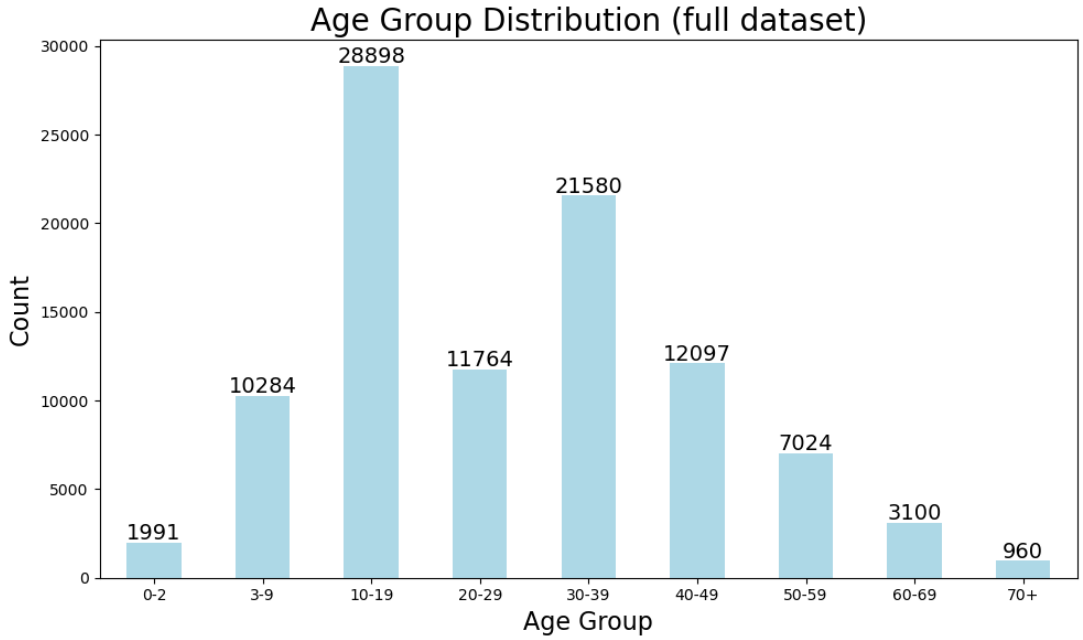
We found FairFace, a novel face image dataset including 97,698 collected from the YFCC-100M Flickr dataset and labeled with race, gender, and age groups. Each image has 5 labels:

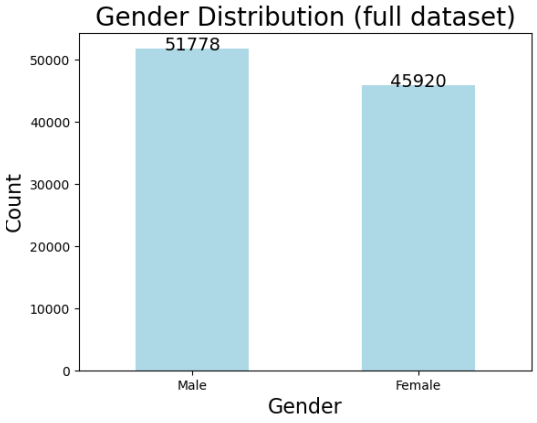
1. file
2. age
3. gender
4. race
5. service_test

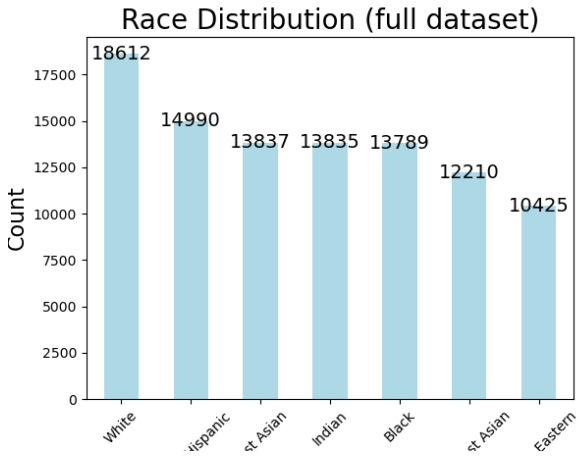
Feature categories:

age	gender	race	service_test
0-2	Male	East Asian	True
3-9	Female	Indian	False
10-19		Black	
20-29		White	
30-39		Middle Eastern	
40-49		Latino_Hispanic	
50-59		Southeast Asian	
60-69			
70+			

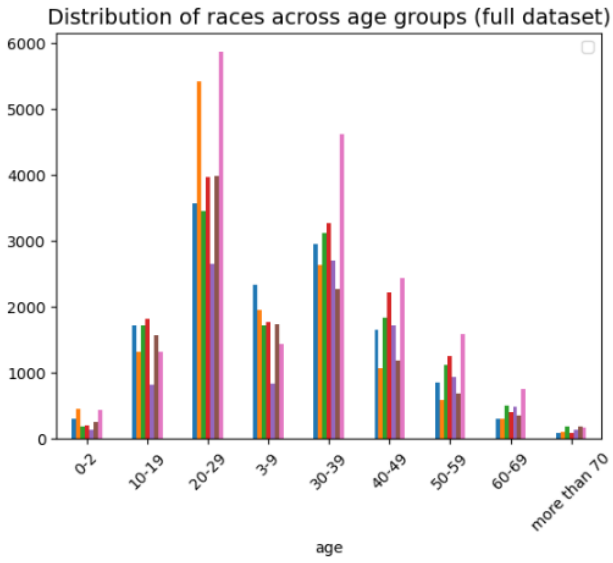
`file` is the filename of the associated image. `age`, `gender`, and `race` represent the demographic categories of each image, and their respective distributions in the dataset are shown below:

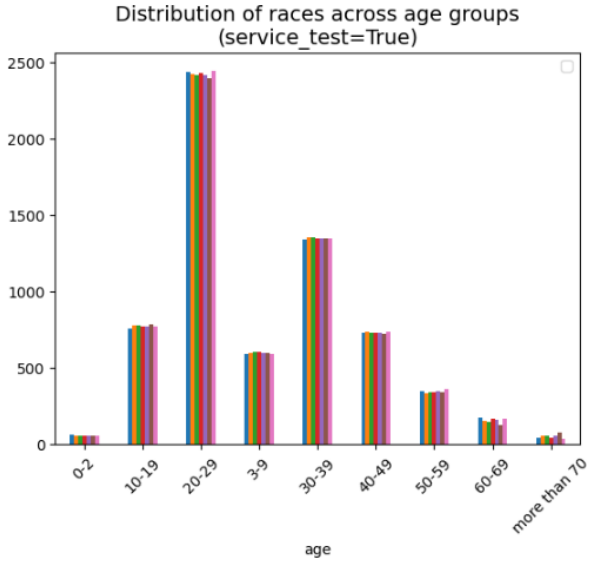






The meaning of the boolean `service_test` column is somewhat ambiguous, but based on an issue[1] on the FairFace GitHub page, it is an indicator of a subset of 40,252 rows where the dataset is balanced by gender and race within each age group as well as across all age groups, as can be seen in comparing the figures below:



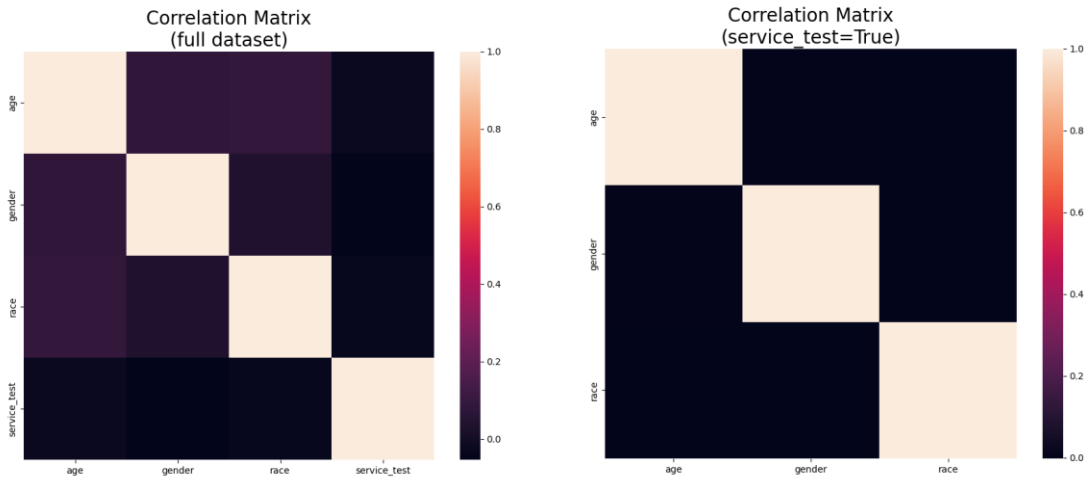


We find the same alignment in gender across age ranges when applying the service_test=True filter. This also results in a uniform distribution across gender and race for the sub-dataset overall, while the age distribution remains about the same. We will experiment with using both the full dataset and the subset where service_test=True.

[1] https://github.com/joojs/fairface/issues/9

Correlation

We found no/negligible correlation between the labels (max magnitude 0.088) in the full dataset, and even less in the age-balanced data subset, as shown in the heatmaps:



Future Work Plan

We are on track with our original timeline:

Week 7: (Feb 21-27)
Finalize writeup of Report 1

Feb 27: Report 1 DUE

Week 8: (Feb 28 - Mar 6)
Get data into (jupyter) environment

Week 9-11: (Mar 7-27)
Spring break
Implement tests, evaluate model
DETAILED documentation of findings

Week 12: (Mar 28 - Apr 3)
Flex week - extra time for implementing tests/evaluating model if things go wrong, or starting to establish the bias eval pipeline
Begin writeup of report 2: what we have accomplished so far and plan for pipeline
May begin work on the pipeline in this week

Week 13-14: (Apr 4-16)
Write/finalize report 2
Start on pipeline in earnest

DETAILED documentation of pipeline

Apr 17: Report 2 DUE

Week 15: (Apr 17-23)
Finalize pipeline
DETAILED documentation of pipeline
Start final report writeup

Week 16: (Apr 24-30)
No more technical work (it is what it is)
Final report
Final presentation

May 1: Final Deliverables DUE

Potential Concerns [C] and Blockers [B]

Identifier	Description
C	Keep in mind the nuance of metrics, that rarely can things be encompassed in a single number
C	We anticipate slow data processing without a GPU. Plan to ask for a Rivanna (HPC at UVA) project

Resources:

[1]
T. P. Pagano et al., "Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," BDCC, vol. 7, no. 1, p. 15, Jan. 2023, doi: 10.3390/bdcc7010015.

[2]
S. Simpson, J. Nukpezah, K. Brooks, and R. Pandya, "Parity benchmark for measuring bias in LLMs," AI Ethics, Dec. 2024, doi: 10.1007/s43681-024-00613-4.

[3]
Y. Yang et al., "Enhancing fairness in face detection in computer vision systems by demographic bias mitigation," in Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, Jul. 2022, pp. 813–822, doi: 10.1145/3514094.3534153.

[4]
P. Saleiro et al., "Aequitas: A Bias and Fairness Audit Toolkit," arXiv, 2018, doi: 10.48550/arxiv.1811.05577.

[5]
T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," arXiv, 2016, doi: 10.48550/arxiv.1607.06520.

[6]
S. Dehdashtian et al., “Fairness and Bias Mitigation in Computer Vision: A Survey,” arXiv, 2024, doi: 10.48550/arxiv.2408.02464.
[7]
B. Wilson, J. Hoffman, and J. Morgenstern, “Predictive inequity in object detection,” arXiv, 2019, doi: 10.48550/arxiv.1902.11097.
[8]
W. Samek, T. Wiegand, and K.-R. Müller, “Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models,” arXiv, 2017, doi: 10.48550/arxiv.1708.08296.
[9]
A. Chinchure et al., “TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models,” arXiv, 2023, doi: 10.48550/arxiv.2312.01261.
[10]
S. T. Erukude, A. Joshi, and L. Shamir, “Identifying bias in deep neural networks using image transforms,” Computers, vol. 13, no. 12, p. 341, Dec. 2024, doi: 10.3390/computers13120341.
[11]
A. A. Kuriakose, “Algomox Blog | Bias in Generative AI: Detection, Mitigation, and Management through MLOps,” Bias in Generative AI: Detection, Mitigation, and Management through MLOps, Apr. 16, 2024. https://www.algomox.com/resources/blog/bias_generative_ai_detection_mitigation_mlops/ (accessed Feb. 07, 2025).
[12]
J. Himmelreich, A. Hsu, K. Lum, and E. Veomett, “The Intersectionality Problem for Algorithmic Fairness,” arXiv, 2024, doi: 10.48550/arxiv.2411.02569.
[13]
B. Moses and A. Dhinakaran, “ML Observability Overview | Machine Learning Observability Resources,” Beyond Monitoring: The Rise of ML Observability, May 19, 2021. https://www.montecarlodata.com/blog-beyond-monitoring-the-rise-of-observability/ (accessed Feb. 07, 2025).
[14]
G. Barcelos, “Understanding Bias in Machine Learning Models - Arize AI,” Understanding Bias in Machine Learning Models, Mar. 15, 2022. https://arize.com/blog/understanding-bias-in-ml-models/ (accessed Feb. 04, 2025).
[15]
M. Shah and N. Sureja, “A comprehensive review of bias in deep learning models: methods, impacts, and future directions,” Arch. Computat. Methods Eng., vol. 32, no. 1, pp. 255–267, Jan. 2025, doi: 10.1007/s11831-024-10134-2.
[16]
Y. Wan, W. Wang, P. He, J. Gu, H. Bai, and M. R. Lyu, “Biasasker: measuring the bias in conversational AI system,” in Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, New York, NY, USA, Nov. 2023, pp. 515–527, doi: 10.1145/3611643.3616310.
[17]

J. Finocchiaro et al., “Bridging Machine Learning and Mechanism Design towards Algorithmic Fairness,” arXiv, 2020, doi: 10.48550/arxiv.2010.05434.

[18]

S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, “A comparative study of fairness-enhancing interventions in machine learning,” arXiv, 2018, doi: 10.48550/arxiv.1802.04422.

[19]

H. Suresh and J. V. Guttag, “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle,” arXiv, 2019, doi: 10.48550/arxiv.1901.10002.

[20]

G. Montavon, W. Samek, and K.-R. Müller, “Methods for Interpreting and Understanding Deep Neural Networks,” arXiv, 2017, doi: 10.48550/arxiv.1706.07979.

[21]

D. Paperno, M. Marelli, K. Tentori, and M. Baroni, “Corpus-based estimates of word association predict biases in judgment of word co-occurrence likelihood.,” Cogn. Psychol., vol. 74, pp. 66–83, Nov. 2014, doi: 10.1016/j.cogpsych.2014.07.001.

[22]

D. DeAlcala, I. Serna, A. Morales, J. Fierrez, and J. Ortega-Garcia, “Measuring Bias in AI Models: An Statistical Approach Introducing N-Sigma,” in 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), Jun. 2023, pp. 1167–1172, doi: 10.1109/COMPSAC57700.2023.00176.

[23]

“[2111.09983] Towards Measuring Fairness in Speech Recognition: Casual Conversations Dataset Transcriptions.” https://doi.org/10.48550/arXiv.2111.09983 (accessed Jan. 30, 2025).

[24]

R. K. E. Bellamy et al., “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias,” arXiv, 2018, doi: 10.48550/arxiv.1810.01943.

[25]

Y. Zhou, M. Kantarcioglu, and C. Clifton, “Improving Fairness of AI Systems with Lossless De-biasing,” arXiv, 2021, doi: 10.48550/arxiv.2105.04534.

[26]

M. Gray et al., “Measurement and mitigation of bias in artificial intelligence: A narrative literature review for regulatory science.,” Clin. Pharmacol. Ther., vol. 115, no. 4, pp. 687–697, Apr. 2024, doi: 10.1002/cpt.3117.

[27]

X. Ferrer, T. van Nuenen, J. M. Such, M. Cote, and N. Criado, “Bias and Discrimination in AI: A Cross-Disciplinary Perspective,” IEEE Technol. Soc. Mag., vol. 40, no. 2, pp. 72–80, Jun. 2021, doi: 10.1109/MTS.2021.3056293.

[28]

S. Ritter, D. G. T. Barrett, A. Santoro, and M. M. Botvinick, “Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study,” arXiv, 2017, doi: 10.48550/arxiv.1706.08606.

[29]

P. Stock and M. Cisse, “ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases,” in Computer vision – ECCV 2018: 15th european conference, munich, germany, september 8–14, 2018, proceedings, part VI, vol. 11210, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 504–519.

[30]

E. A. Watkins, M. McKenna, and J. Chen, “The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness,” arXiv, 2022, doi: 10.48550/arxiv.2202.09519.

[31]

O. Aka, K. Burke, A. Bäuerle, C. Greer, and M. Mitchell, “Measuring Model Biases in the Absence of Ground Truth,” arXiv, 2021, doi: 10.48550/arxiv.2103.03417.

[32]

T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, “A survey on datasets for fairness-aware machine learning,” WIREs Data Min & Knowl, vol. 12, no. 3, May 2022, doi: 10.1002/widm.1452.

[33]

D. Roselli, J. Matthews, and N. Talagala, “Managing bias in AI,” in Companion Proceedings of The 2019 World Wide Web Conference on - WWW ’19, New York, New York, USA, May 2019, pp. 539–544, doi: 10.1145/3308560.3317590.

[34]

Karkkainen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1548-1558).