Grenze International Journal of Engineering and Technology, January Issue

Ethical Implications of Biases in AI and Machine Learning Algorithms

Shreya[1] and Praveen Ailawalia[2]
[1-2]Department of Mathematics, University Institute of Sciences, Chandigarh University, Gharuan- Mohali, Punjab, India
Email: rajshreya8271@gmail.com, Praveen_2117@rediffmail.com

***Abstract*— Due to data-driven decision-making, artificial intelligence (AI) systems are able to create "shadow profiles," which allow them to infer sensitive personally identifiable information, like gender, unnecessarily from indirect profiling sources without user consent. Even when those predictions can seem to be accurate, this experience raises serious ethical challenges and problems when an AI model's predictions become biased against certain groups of individuals seeking to use a variety of platform purposes. In this work, we have investigated the shadow profile creation in ethical terms with the UCI Adult Dataset by building a model that accurately predicts gender. A Random Forest classifier was built to make gender predictions based on only the non-sensitive attributes in the dataset. With the use of SHAP (SHapley Additive exPlanations), we examined the features that influenced the model and potentially led to unequal performance across groups based on race. With Fairlearn implemented Exponentiated Gradient algorithm with demographic parity, racial bias deterioration was possible, but at the expense of decreased overall predictive accuracy of the model. This study adds to the argument for fairness-aware modeling and fairness-friendly explainability to help develop ethical AI systems.**

***Keywords*— Explainable AI (XAI), Machine Learning (ML), Shadow Profiles, Algorithmic Bias, Gender Inference, Fairness in AI.**

I. INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) systems play an ever-increasing role in real-world decision-making, from job hiring and lending to targeted advertising and profiling users. An area related to AI and ML that continues to develop and, in some cases, demonstrates irresponsible social practices is the inference derived from data about people. This data can be a combination of indirect, publicly available data sources that do not require the user's explicit consent to profile them, thus creating shadow profiles [1]. That being said, these profiles implicitly infer values of sensitive characteristics (e.g., gender, age, income, and ethnicity), which implicitly introduces ethical and privacy issues.

In particular, the models that infer gender provide a serious concern. Although the inference models could help meet a user's needs or optimize services, they perpetuate existing societal bias [2]. For example, if the model learns from biased or entering biased historical data, it may learn one gender over the other when associating features such as income, education, or job type. The bias in the model impacts not only inferred predictions but perpetuates societal stereotypes and discrimination.

Grenze ID: 01.GIJET.12.1.338

This research explores the ethics and practicality of shadow profiling through a case study of gender inference. In this study, we utilize the UCI Adult dataset and retroactively construct a shadow profile to predict gender with a model strictly from non-sensitive attributes. We examine the resulting biases by utilizing SHAP (SHapley Additive exPlanations) to understand feature importance, and using Fairlearn's Exponentiated Gradient algorithm to mitigate unfairness by race. Overall, our study highlights how AI models can mobilize unintended harm towards certain groups, as well as methods we can implement to construct fairer systems.

II. METHODOLOGY

A. Dataset

The research utilized the UCI Adult dataset, which has 48,842 records with demographic and socio-economic factors including age, education, occupation, capital gains, marital status, race, and hours worked per week.

- The dataset is a well-established dataset for fairness and bias analysis, as sensitive attributes like gender, race, and income are well represented in the data.
- To ultimately create a shadow profiling situation, the explicit gender category was unavailable during training. The model infers gender indirectly through correlated non-sensitive attributes instead.

The UCI Adult dataset was used for this study. It contains demographic and economic information of individuals. The main attributes are summarized in Table 1.

TABLE I: KEY ATTRIBUTES OF THE UCI ADULT DATASET USED IN THIS STUDY

Dataset Overview		
Attribute	Description	Example Values
Age	Age of the individual	39, 28, 52
Education	Highest level of education attained	HS-grad, Bachelors, Masters
Occupation	Type of work	Sales, Tech-support, Craft
Hours-per-week	Number of working hours per week	40, 50, 60
Capital-gain	Income from investments	0, 7688, 99999
Race	Race of the individual	White, Black, Asian-Pac-Islander
Income	Income class	<=50K, >50K

B. Model Development

- A Random Forest Classifier was chosen for its capabilities of robustness and interpretability.
- The data was split into training and testing subsets in an 80-20 split.
- Data preprocessing 'done' involved label encoding categorical variables, and normalizing continuous features.
- The model's goal was to predict the gender of an individual with social economic features, performing a shadow profile inference task.

C. Bias Detection and Explainability

- Performance Evaluation: We evaluated model accuracy overall and across subgroups (by income and race).
- Explainability Analysis: To understand the influence of features on predictions, we used SHAP (SHapley Additive Explanations), which allowed us to see which attributes (e.g., hours-per-week, education, capital-gain) most influenced our ability to predict gender.
- Fairness Evaluation: We evaluated group-wise performance (accuracy) across the racial categories, revealing discrepancies and possible biases.

D. Bias Mitigation

To minimize the likelihood of unfair outcomes, we utilized Fairlearn's Exponentiated Gradient algorithm framework with a Demographic Parity fairness constraint [3].

- Sensitive Feature: Race was selected as the sensitive feature/attribute.
- Mitigation Method: The mitigation algorithm provided reweighted samples to impose fairness to the groups.
- Post-Mitigation Evaluation: Group-wise revise the accuracy of the model to see if the differences amongst racial groups were reduced.

E. Ethical Considerations

- The work recognizes that shadow profiling raises privacy and consent problems [1] by inferring attributes without user awareness.

- Bias mitigation was also assessed for its ethical trade-offs concerning technical effectiveness [4], where fairness could be achieved only at the expense of overall accuracy.

III. Experiments and Results

The model achieved an overall accuracy of 86.7%. However, disparities were observed in subgroup accuracies:
- Accuracy by income:
    - <=50K: 80.6%
    - 50K: 95.0%

The baseline Random Forest model achieved an overall accuracy of 86.7%. However, subgroup performance varied considerably, as shown in Table 2.

TABLE II: Model Accuracy Across Income and Racial Subgroups Before Applying Mitigation

Model Performance Before Mitigation	
Group	Accuracy (%)
Overall	86.7
Income <=50K	80.6
White	84.8
Amer-Indian-Eskimo	83.9
Asian-Pac-Islander	83.1
Other	78.5
Black	76.9

Figure 1 shows accuracy of gender prediction across racial groups. The horizontal bar chart displays prediction accuracy, ranging from a low of 0.60 to a high of 0.96, for categories, such as White, Amer-Indian-Eskimo, Asian-Pac-Islander, Other, and Black. These results indicate accuracy is high for all groups; generally, between 0.78 and 0.82, with the exception of some variability among racial groups. Taken together, the data would suggest the models worked at similar abilities across race, with slight deviation in accuracies based on racial group.

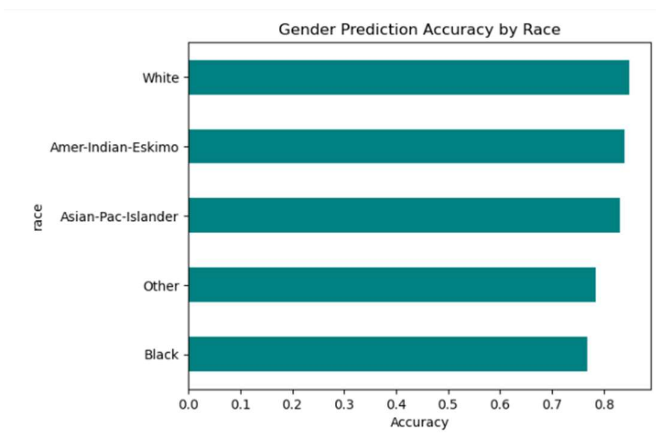

Figure 1. Gender prediction accuracy across racial groups before bias mitigation.

Accuracy by race:
- White: 84.8%
- Amer-Indian-Eskimo: 83.9%
- Asian-Pac-Islander: 83.1%
- Other: 78.5%
- Black: 76.9%

Figure 2 shows group-wise accuracy by race before and after Fairlearn mitigation.
- The blue bars show model accuracy for various racial groups (e.g., White, American Indian Eskimo, Asian Pac Islander) prior to mitigation.
- The orange bars convey the accuracy of all groups after Fairlearn mitigation was applied to reduce bias.

The takeaways from this chart are that:

- Before mitigation the accuracies for some groups (e.g.; White, American-Indian-Eskimo, Asian-Pac-Islander) were quite a bit above the others (e.g., Black, Other)
- After mitigation all groups showed a decline in accuracy but were more balanced across races reducing the performance gap.

After applying Fairlearn's mitigation techniques, group-wise accuracy became more equitable:

- White: 72.7%
- Amer-Indian-Eskimo: 69.8%
- Asian-Pac-Islander: 69.8%
- Other: 68.4%
- Black: 63.4%

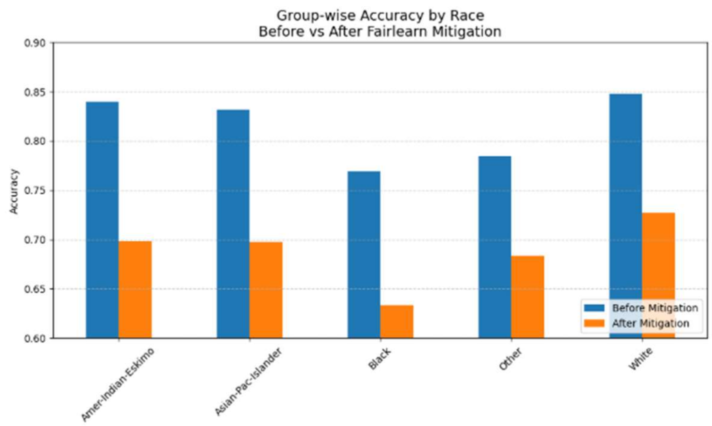

Figure 2. Gender prediction accuracy across racial groups after

SHAP analysis showed that features like relationship, hours-per-week, and education had high influence on predictions.
The strongest features related to gender inference are shown in Figure 3, with relationship status, occupation, and hours worked per week on the top three. This appears to reinforce gender-based stereotypes, for example, higher income and longer hours per week is associated with males.

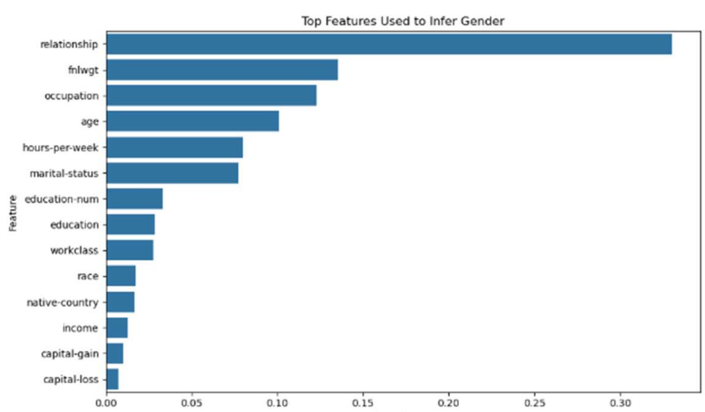

Figure 3. Comparison of subgroup accuracies by race before and after bias mitigation.

Figure 4 demonstrates the variance in gender prediction accuracy across income categories. It is apparent the model generator's accuracy is much more accurate for the income category above $50K (95%) as compared to the income category below or equal to $50K (80.6%). The variance demonstrates that factors about income, have an intangible bias in difference of model's ability to predict gender for those in higher income brackets. The palpable bias described here, is not only an example of some bias in the data set, but is perhaps more troubling, an ethical issue. Predictive systems are better at performing for some socio-economic groups and as result will likely perpetuate existing socio-economic disparities.

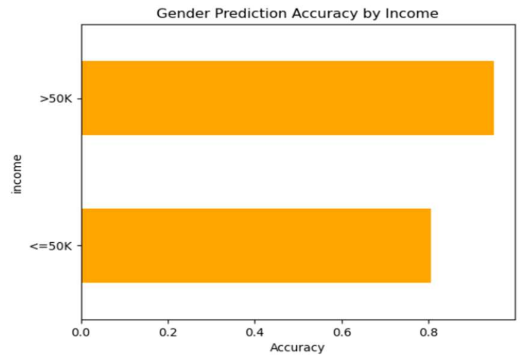

Figure 4. Gender prediction accuracy by income category (<=50K vs. >50K).

IV. DISCUSSION

Model accuracy varies markedly between denominations, which means that inherent bias exists in either the data representation or the model behaviour. Therefore, without modelling for fairness, models can continue social inequities [5].

We improved fairness using a form of bias mitigation based on Demographic Parity; nevertheless, this resulted in a trade-off with overall accuracy—a classic dilemma in responsible AI development that often involves sacrificing predictive effectiveness for fairness.

Finally, the interpretability offered by SHAP and other features improves transparency and trust, providing stakeholder understanding of model decisions and challenging unfair decisions.

V. LIMITATIONS

While this study provides valuable insights, it is not without limitations. First, the analysis was conducted on the UCI Adult dataset, which is widely used in fairness research but is relatively old and U.S.-centric. This means the findings may not fully capture today's global or culturally diverse contexts.

Second, the study focused on a single bias mitigation technique — Fairlearn's Demographic Parity. Although this method helped reduce disparities, it represents only one approach. Exploring other techniques such as Equalized Odds or Adversarial Debiasing could provide a more complete picture.

Third, the gender variable in the dataset is binary (male/female), which oversimplifies the reality of gender identity. Non-binary and gender-diverse groups were not represented, which limits the inclusivity of the study.

Lastly, the model chosen was a Random Forest classifier. While it is robust and interpretable, results may differ if deep learning or other algorithms were applied. Therefore, the conclusions should be viewed as specific to this modelling choice rather than universally applicable.

VI. FUTURE WORK

This research opens up several directions for future exploration. One important step would be to experiment with other fairness techniques such as Equalized Odds or Adversarial Debiasing. These methods might strike a better balance between fairness and accuracy, and comparing them could reveal which approach is most suitable for different contexts.

Another avenue is to look beyond gender inference. Shadow profiling often extends to other sensitive traits such as political orientation, religious beliefs, or even health status. Studying these attributes would provide a deeper understanding of how AI systems can unintentionally—or intentionally—reconstruct aspects of a person's identity without consent.

In addition, applying the analysis to more modern datasets, such as social media interactions or financial application data, would help validate the findings in today's digital environment. The UCI Adult dataset offers a useful baseline, but real-world data from current platforms could reveal new forms of bias that are more relevant to contemporary society.

Finally, future research should also consider privacy-preserving methods like differential privacy or federated learning. These approaches could allow AI systems to make predictions while minimizing the risk of intrusive shadow profiling, ensuring that individuals' sensitive information remains secure and under their control.

VII. CONCLUSION

This study set out to examine how machine learning models can create shadow profiles and, in doing so, reinforce hidden biases. The findings showed that even when gender information is not explicitly provided, models can still infer sensitive traits using correlated features such as income, education, and working hours. This ability highlights both the power and the risk of modern AI systems.

Our results also revealed that bias does not affect all groups equally. Accuracy was consistently higher for certain subgroups, particularly White and higher-income individuals, while minority groups and lower-income individuals experienced lower predictive performance. This imbalance demonstrates how AI systems, if left unchecked, can unintentionally reinforce existing social inequalities.

By applying Fairlearn's mitigation techniques, we were able to reduce these disparities, showing that fairness interventions do work. However, this improvement came at the cost of lower overall accuracy, pointing to an important trade-off between performance and equity that practitioners and policymakers must grapple with.

Ultimately, the study underscores a broader ethical concern: AI systems should not be allowed to infer personal traits without user knowledge or consent. Doing so risks not only privacy violations but also the amplification of harmful stereotypes. The path forward lies in building responsible AI practices—where transparency, fairness, and respect for individual autonomy guide the design of algorithms.

REFERENCES

[1] Garcia, D. — Leaking privacy and shadow profiles in online social networks. This study demonstrates how platforms can infer personal traits from users — even non-users — using social data, introducing the concept of "shadow profiles." It lays important groundwork for understanding the ethics of inferred data.

[2] Leavy, S. et al. — Mitigating Gender Bias in Machine Learning Data Sets Offers a thoughtful exploration of identifying and removing gender bias in training data, drawing from sociolinguistics and gender theory. This is highly relevant to your discussion on biased training signals and stereotype reinforcement.

[3] Feldman, T., Peake, A. — End-to-End Bias Mitigation: Removing Gender Bias in Deep Learning Proposes a comprehensive bias mitigation framework combining pre-, in-, and post-processing techniques, specifically tackling gender bias in deep learning. This aligns well with your methodology and offers a broader context on mitigation strategies

[4] Alvarez, J. M. et al. — Policy advice and best practices on bias and fairness in AI Provides a comprehensive overview of AI fairness methods and policy frameworks, especially relevant to European regulations. This helps broaden the ethical context of your study.

[5] Sunny Shrestha, Sanchari Das — Exploring gender biases in ML and AI academic research through systematic literature review A systematized review of gender bias research in AI, highlighting methodological gaps and the need for user-centered studies. It offers disciplinary grounding and highlights the novelty of your shadow profiling focus.