Can sampling based mitigation techniques lower gender bias within the MORPH-II dataset?

1st Kaan Gögcay
Master of Applied Intelligence
Hogeschool van Amsterdam
Amsterdam, Netherlands

***Abstract*—This study investigates whether sampling-based bias mitigation techniques can reduce gender bias in facial age prediction models trained on the MORPH-II dataset. MORPH-II contains 50,015 images of individuals aged 16 to 77, but suffers from a severe gender imbalance, with 86.1% male and 13.9% female subjects. Three experimental workflows were designed: a baseline using the original dataset, a randomly undersampled version with balanced gender distribution, and an oversampled version that included additional female images from the UTKFace dataset. All models were trained using both classification and regression approaches, with evaluation based on Mean Absolute Error (MAE) and the absolute difference in MAE between genders as a measure of bias. Results show that random undersampling reduced the gender bias from 1.29 to 0.68, while oversampling increased it to 1.62, likely due to inconsistencies between the MORPH-II and UTKFace datasets. These findings suggest that simple undersampling can effectively improve fairness in age prediction models, whereas naive oversampling using external data may worsen bias when dataset characteristics are not well aligned.**

I. Introduction

In the Netherlands, research into automated age prediction systems have slowly become more popular over the past ten years [1] [2]. Over time, more supermarkets have begun integrating such systems into their checkout processes to automatically verify a customer's age for restricted purchases. Motivated by this growing trend, we, a group of Master's students in Applied Artificial Intelligence, aim to explore and improve the performance and fairness of age prediction models.

During our initial exploration, we examined several publicly available facial age datasets, including UTKFace, FG-NET, MORPH-II, and IMDB-WIKI. We observed that each dataset contains unique characteristics and varying levels of metadata quality. Among these, the MORPH-II dataset stood out due to its significant gender imbalance, with a substantial overrepresentation of male subjects. Such imbalances can lead to systematic differences in model performance across genders, resulting in gender bias.

Although this imbalance initially discouraged us from using MORPH-II for our main project, the dataset's quality and scale make it an excellent case for studying bias mitigation. Therefore, this research investigates whether sampling-based mitigation techniques can reduce gender bias in age prediction models trained on the MORPH-II dataset.

The central research question that guides this study is as follows: *Can sampling based mitigation techniques lower gender bias within the MORPH-II dataset?*

II. Methodology

This study consists of three main experimental workflows: using the original MORPH-II dataset, an oversampled dataset, and an undersampled dataset. Each workflow follows the same sequence of steps: exploratory data analysis (EDA), model training, evaluation, and measurement of gender bias. The models will be compared to assess the effectiveness of each sampling technique in reducing gender bias.

A. Dataset Analysis

The MORPH-II dataset contains 50,015 images with metadata including age and gender, with ages ranging from 16 to 77. A major limitation is the gender imbalance, with 86.1% male and 13.9% female subjects. Unlike UTKFace, MORPH-II does not provide ethnicity information. Additionally, since MORPH-II contains multiple images per subject, it could introduce a form of data leakage or overfitting if similar faces appear across training, validation, and test sets.

B. Model Training

Models were trained using a classification approach with 62 age classes, covering ages 16 to 77. Both classification and regression outputs were evaluated, partially following the workflow described in the DEX paper [3]. The analysis focuses on regression-based predictions, as they consistently achieved lower MAE scores compared to classification outputs.

Experiments were conducted using different learning rates and numbers of epochs. A learning rate of 0.001 was too high, preventing effective learning. Reducing it to 0.0001 allowed the model to learn, but overfitting occurred after approximately six epochs. Epoch selection was used to choose the best performing model for each workflow (see Figures 4, 5, 7, 8, 10, 11). Only two learning rates were tested, but future experiments could explore even lower rates to potentially improve generalization. Early stopping or additional regularization techniques could also help reduce overfitting.

C. Sampling-Based Bias Mitigation

Two sampling strategies were applied to address gender imbalance:

- **Random Undersampling:** The majority class (male) was reduced to achieve a 50/50 gender distribution [4].
- **Traditional Oversampling:** Female images from the UTKFace dataset (ages 16–77) were added to the original dataset, changing the gender distribution to 70.5% male and 29.5% female [4] [5].

For each strategy, a separate dataset was created and the model was trained using the same hyperparameters as the baseline (Figures 4, 5, 7, 8, 10, 11). Performance metrics and gender bias were measured consistently across all workflows.

D. Evaluation Metrics

Model performance was evaluated using the mean absolute error (MAE). Gender bias was quantified as the absolute difference in MAE between male and female subjects:

$$\text{Bias}_{\text{abs}} = |\text{MAE}_{\text{male}} - \text{MAE}_{\text{female}}|$$

These metrics were used to compare the baseline, undersampled, and oversampled models.

III. Results

A. Original Dataset

On the baseline MORPH-II dataset, regression predictions achieved an MAE of 3.69 for male subjects and 4.98 for female subjects, resulting in a gender bias of 1.29 (Figure 6). Overfitting was observed after approximately six epochs (Figure 4).

B. Random Undersampling

For the undersampled dataset, MAE increased to 4.23 for male subjects while decreasing slightly to 4.91 for female subjects. This reduced the absolute gender bias to 0.68. These results suggest that random undersampling effectively reduced the performance gap between genders while slightly lowering overall model accuracy (Figure 9).

C. Traditional Oversampling

For the oversampled dataset, the MAE for male subjects remained similar to the baseline, while MAE for female subjects increased, leading to an absolute gender bias of 1.62. This counterintuitive result is likely due to differences between the UTKFace and MORPH-II datasets: UTKFace images have greater variability in pose, lighting, and background (Figure 1), whereas MORPH-II consists of standardized mugshots (Figure 2). Adding UTKFace images increased training variability, negatively impacting predictions for female subjects (Figure 12).

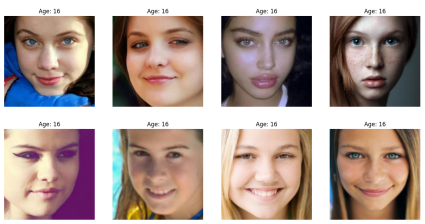

Fig. 1. Several female images samples between the age of 16 and 77 from the UTKFace dataset

Fig. 2. Several female images samples between the age of 16 and 77 from the MORPH-II dataset

D. Comparison Across Workflows

Figure 3 summarizes regression MAE for all workflows. Undersampling reduced gender bias by increasing MAE for men while keeping MAE for women nearly constant. Oversampling increased bias due to the added variability from external images.

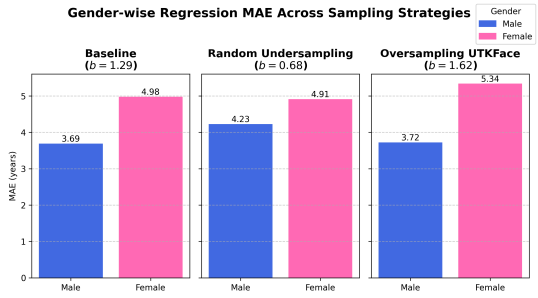

Fig. 3. MAE score comparison between all tested models.

Overall, undersampling proved more effective in reducing gender bias, while oversampling demonstrated the importance of dataset consistency when adding external images.

IV. Discussion

The results of this study provide clear insights into how sampling-based mitigation techniques affect gender bias in the MORPH-II dataset.

The dataset itself presents some challenges. Its strong gender imbalance (86.1% male, 13.9% female) causes the model to perform better on the majority class. Additionally, multiple images per subject may lead to data leakage, which

can slightly inflate performance metrics and affect bias measurements.

Among the mitigation strategies, random undersampling reduced gender bias effectively, lowering the absolute bias from 1.29 to 0.68. This is expected, as removing male samples reduced the model's reliance on the majority class, while performance for female subjects stayed mostly the same. However, this came at the cost of slightly higher MAE for male subjects, showing the trade-off between fairness and overall accuracy.

On the other hand, traditional oversampling using UTKFace images increased gender bias from 1.29 to 1.62. This is likely due to differences between the two datasets: MORPH-II contains standardized mugshots, while UTKFace images vary in pose, lighting, and background. Adding these images increased variability in the training data, which negatively affected predictions for female subjects.

Finally, the model showed signs of overfitting across all datasets. Although epoch selection was used to choose the best model, applying early stopping or additional regularization could improve generalization. Only two learning rates were explored (0.001 and 0.0001), and since the model overfitted quickly, it could also be useful to try an even lower learning rate (e.g., 0.00005 or 0.00001) to observe how the model's training dynamics would change.

In conclusion, random undersampling is a simple and effective method for reducing gender bias in this context, while naive oversampling with an external dataset can have the opposite effect. These findings highlight the importance of dataset characteristics and domain consistency when applying sampling-based mitigation techniques.

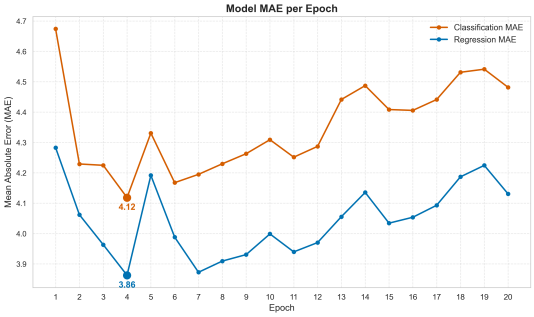

Fig. 5. Validation MAE per epoch for classification and regression on regular dataset.

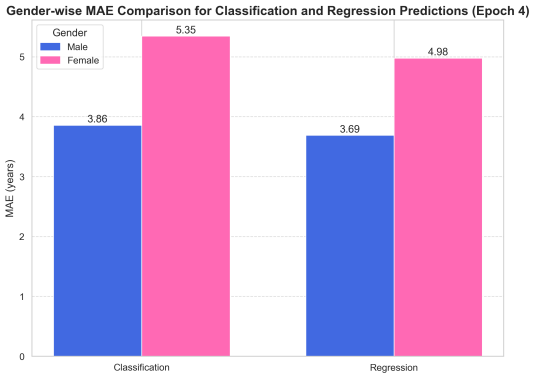

Fig. 6. MAE score comparison for men and women using classification and regression on the test set.

Appendix A
Additional Figures

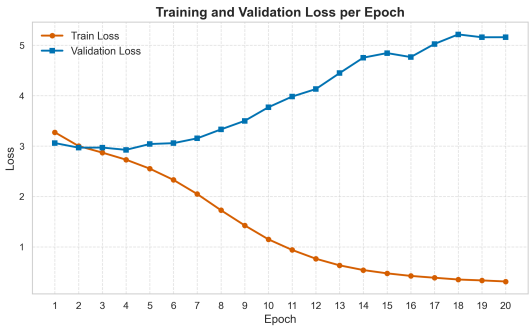

Fig. 4. Training and Validation Loss per Epoch for the model trained on the regular dataset.

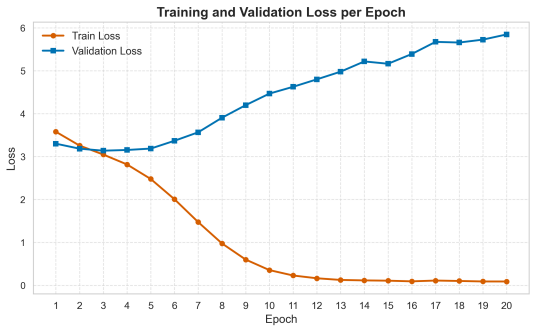

Fig. 7. Training and validation loss per epoch for the model trained on the undersampled dataset.

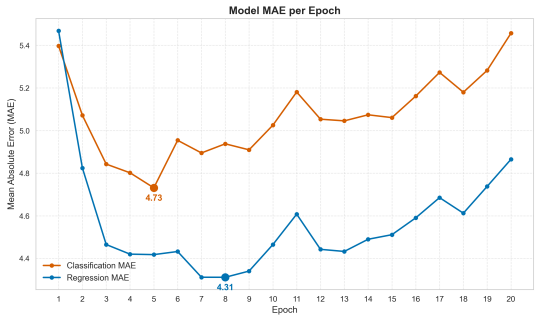

Fig. 8. Validation MAE per epoch for classification and regression on undersampled dataset.

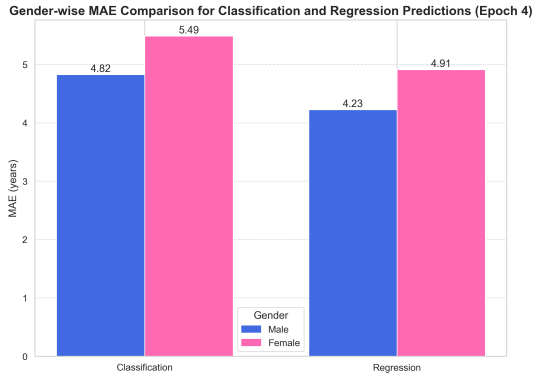

Fig. 9. MAE score comparison for men and women using classification and regression on the test set.

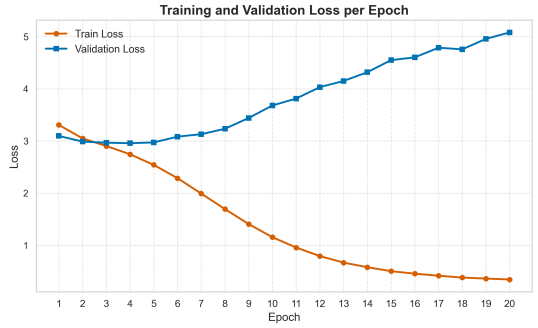

Fig. 10. Training and validation loss per epoch for the model trained on the oversampled dataset.

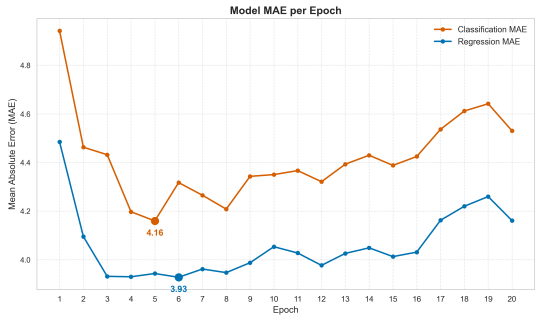

Fig. 11. Validation MAE per epoch for classification and regression on oversampled dataset

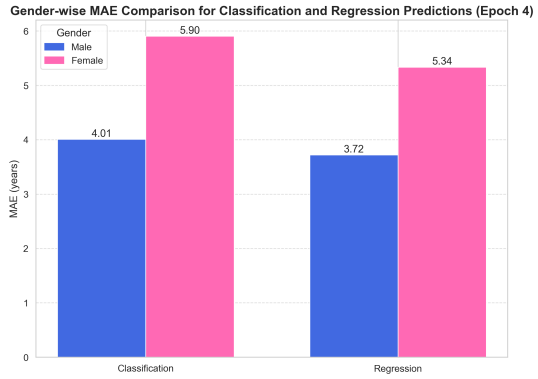

Fig. 12. MAE score comparison for men and women using classification and regression on the test set.

REFERENCES

[1] J. J. Van Hoof, "The effectiveness of id readers and remote age verification in enhancing compliance with the legal age limit for alcohol," *European Journal of Public Health*, vol. 27, pp. 357–359, 10 2016.

[2] J. {van Hoof} and B. {van Velthoven}, "Remote age verification to prevent underage alcohol sales. first results from dutch liquor stores and the economic viability of national adoption," *International journal of drug policy*, vol. 26, pp. 364–370, Apr. 2015.

[3] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 252–257, 2015.

[4] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409, 05 2017.

[5] S. Rančić, S. Radovanovic, and B. Delibašić, *Investigating Oversampling Techniques for Fair Machine Learning Models*, pp. 110–123. 05 2021.