BIAS MITIGATION

Kashish Varma C- 1RVU22CSE039
Sruthika Sivakumar- 1RVU22CSE164

Mentor: Prof Shobhana Padmanabhan

INTRODUCTION

- This project compares four deep learning models **(Domain-Adversarial Neural Networks, Adversarial Debiasing, FairGAN, and Fair Autoencoder)** for mitigating gender bias in predicting course outcomes, using the dataset
- The goal is to evaluate the performance and fairness of each model by comparing metrics like accuracy, demographic parity, and disparate impact before and after applying the bias mitigation techniques.
- The study aims to identify the best-performing model in terms of both maintaining prediction accuracy and ensuring fair outcomes for male and female students.

SUMMARY OF LITERATURE SURVEY

- The literature on bias mitigation in deep learning focuses on pre-processing, in-processing, and post-processing methods.
- Pre-processing techniques like **re-sampling** modify datasets before training but may lose information.
- In-processing methods, such as **Adversarial Debiasing and Domain-Adversarial Neural Networks (DANN)**, reduce bias during training by preventing reliance on sensitive attributes.
- FairGAN, a generative approach, creates synthetic data while ensuring fairness.
- These methods are widely used to address gender bias in educational data, with fairness metrics like demographic parity and disparate impact evaluating their effectiveness.
- Balancing fairness and accuracy remains a key challenge in the field.

PROBLEM STATEMENT

The increasing use of deep learning models in decision-making raises concerns about fairness, particularly in the context of gender bias.

First Problem

Gender bias in course predictions can lead to **unequal opportunities** for advancement.

Second Problem

Gender bias in predictions can **reinforce stereotypes**, favoring males in certain fields and disadvantaging females.

OBJECTIVES

1. **Mitigate Gender Bias**: Implement deep learning techniques to reduce gender bias in predicting course outcomes and ensure equal opportunities for male and female students.

2. **Evaluate Fairness Metrics**: Compare the effectiveness of four deep learning models (DANN, Adversarial Debiasing, FairGAN, and Fair Autoencoder) in achieving fairness, using metrics like demographic parity and disparate impact.

3. **Assess Model Performance:** Analyze and compare the predictive accuracy of the four models, ensuring that debiasing techniques do not significantly degrade the performance of course outcome predictions.

4. **Identify the Best Model**: Determine which of the four deep learning models is the most effective at balancing both fairness and prediction accuracy in mitigating gender bias in educational datasets.

BACKGROUND & RELATED WORK

1. **Gender Bias in Education**: Gender bias in educational settings, such as biased course evaluations and grading, can perpetuate stereotypes and impact student opportunities based on gender.

2. **Fairness in Deep Learning:** Fairness in deep learning aims to reduce bias caused by sensitive attributes like gender, ensuring equitable predictions without discrimination.

3. **Adversarial Debiasing:** Adversarial debiasing uses adversarial training to minimize the model’s reliance on sensitive attributes while maintaining accuracy on the primary task.

4. **Domain-Adversarial Neural Networks (DANN)**: DANN reduces bias by applying domain adaptation techniques, where the model learns to predict the target variable without being influenced by sensitive attributes like gender.

5. **FairGAN for Bias Mitigation**: FairGAN leverages generative adversarial networks to generate synthetic data that mitigates gender bias, ensuring fairness by incorporating fairness constraints into the training process.

CHALLENGES & NOVELTY

Challenge

Mitigating gender bias in deep learning models while maintaining high predictive accuracy is a complex task, as traditional bias reduction methods often compromise performance

Novelty

This study introduces a comparative analysis of four advanced deep learning models—Adversarial Debiasing, Domain-Adversarial Neural Networks (DANN), Fair GAN, and Fair Autoencoder in Latent Space—to identify the most effective approach for gender bias mitigation in educational data.

METHODOLOGY

1. Adversarial Debiasing

Model Design:

Adversarial Debiasing aims to mitigate bias in machine learning models by adversarially training the model to be fair with respect to a sensitive attribute (e.g., gender). The approach uses a two-component framework:

1. **Main Classifier:** Trains the model to perform the primary task (e.g., predicting course outcomes) while minimizing the task-specific loss.
2. **Adversary (Bias Classifier):** Attempts to predict the sensitive attribute (e.g., gender) from the model's learned features. The adversary's goal is to use the model's features to infer the sensitive attribute, but the model is trained to minimize the adversary's performance, thus reducing bias.

Training Process

- **Primary Task Loss:** The model is trained to minimize the loss for the main task (e.g., course outcome prediction).
- **Adversarial Loss:** Simultaneously, the model is trained to reduce the accuracy of the bias classifier (domain classifier) by making the features uninformative with respect to the sensitive attribute.

METHODOLOGY

2. Domain-Adversarial Neural Networks (DANN)

Model Design:

Domain-Adversarial Neural Networks (DANN) aim to reduce domain-specific bias by training the model to perform well on the target task while making it invariant to sensitive attributes (e.g., gender). This is achieved through adversarial training, where the model simultaneously learns to predict the target and to confuse a domain classifier that tries to predict the sensitive attribute.

1.**Feature Extractor**: Learns generalizable features from the input data.
2. **Label Classifier:** Predicts the primary task (e.g., course outcome levels).
3. **Domain Classifier**: Attempts to predict the sensitive attribute (e.g., gender) based on the extracted features.
4. **Adversarial Loss:** The model is trained to minimize the label loss while simultaneously maximizing the domain loss (making it harder for the domain classifier to predict the sensitive attribute).

Training Process

- **Primary Task Loss:** Minimize the label classification loss (e.g., for course outcomes).
- **Adversarial Loss:** Maximize the domain classifier’s loss, forcing the model to learn features that are invariant to the sensitive attribute.

METHODOLOGY

3. FairGAN: Fair Generative Adversarial Network for Bias Mitigation

Model Design

FairGAN is a generative model designed to mitigate bias in datasets by generating synthetic data that is both fair and realistic. It aims to adjust the distribution of sensitive attributes (e.g., gender) while maintaining the relationship between input features and the target variable (e.g., Course Outcome Level).

1. **Generator:** Creates synthetic data that resembles real data but reduces bias related to sensitive attributes.
2. **Discriminator:** Distinguishes between real and synthetic data and tries to predict the sensitive attribute (e.g., gender). It penalizes the generator for producing biased data.
3. **Fairness Constraint:** Embedded in the adversarial training, ensuring that the generated data does not reveal the sensitive attribute.
4. **Loss Functions:**
   - **Adversarial loss:** Ensures the discriminator cannot easily distinguish real and fake data.
   - **Fairness loss**: Penalizes the generator if it allows the discriminator to predict the sensitive attribute.

Training Process

- **Generator Update:** Trains to produce realistic and fair synthetic data.
- **Discriminator Update:** Trains to distinguish real vs. fake data and predict the sensitive attribute, while penalizing unfair predictions.

METHODOLOGY

4. Fair Autoencoder in Latent Space

Model Design

The Fair Autoencoder (FairAE) aims to reduce bias in deep learning models by removing sensitive attribute information from the learned latent space while preserving the relevant features for the task at hand.

1. **Encoder:** Encodes the input data into a latent representation, capturing key features for the target prediction.
2. **Latent Space Regularization:** A fairness constraint is applied in the latent space to ensure that the sensitive attribute (e.g., gender) cannot be predicted from the encoded representation.
3. **Decoder:** Decodes the latent representation back to the original feature space, ensuring the output is close to the input data (reconstruction loss).
4. **Bias Mitigation:** The fairness constraint penalizes the encoder if it encodes sensitive information that can be used to predict the sensitive attribute, promoting fairer representations in the latent space.

Training Process

- **Autoencoder Loss:** Minimizes reconstruction loss to ensure accurate data reconstruction.
- **Fairness Loss:** Ensures that sensitive attributes cannot be inferred from the latent representation by adding a penalty if the latent representation is predictive of the sensitive attribute.

IMPLEMENTATION

- **Data Preparation:** Ensure that the dataset is preprocessed with scaling, encoding, and balancing techniques as mentioned earlier.

- **Model Setup**: Implement the DANN, Adversarial Debiasing, FairGAN, and Fair Autoencoder models using deep learning frameworks like TensorFlow or PyTorch. The models are implemented according to the specific architectures and training procedures mentioned above.

- **Training:** Train each model using the training dataset while applying the respective bias mitigation strategies. Use optimization techniques like Adam or SGD and ensure that both fairness and prediction accuracy are monitored.

- **Evaluation:** Evaluate the models on the test dataset using performance metrics (accuracy) and fairness metrics (demographic parity, disparate impact). Visualizations of the predicted distributions before and after debiasing are also analyzed.

DATASET DETAILS

Student Dataset Overview

- **Total Records:** 280 students
- **Attributes:** 15 numerical features related to demographics, academic preparation, and course feedback

Key Features:

1. **Demographics:** Gender (special attribute), Age, Rural/Urban background
2. **Academic Background:** 12th marks %, Preparedness for course
3. **Course Engagement:** Average Weekly Time Spent, Faculty Concern, Outside Help
4. **Performance & Satisfaction:**
   - Target Variable (y): Average Course Outcome Level
   - Average Effort and Satisfaction Levels
   - Average Marks Given and Reasons for Low Marks
5. **Overall Feedback:** Numeric ratings on overall course experience

Dataset Setup:

- **Target (y):** Average Course Outcome Level
- **Features (X):** All columns except Target and Gender
- **Special Attribute**: Gender (used for bias analysis)

RESULT & ANALYSIS

All four models (DANN, Adversarial Debiasing, FairGAN, and Re-weighting) demonstrated the ability to maintain or slightly reduce predictive accuracy while significantly improving fairness metrics. The disparity in favorable outcomes was substantially reduced after debiasing, with Demographic Parity approaching equality and Disparate Impact nearing 1 for all models.

- Best Performing Models: While all models showed promising results, **Adversial debiasing and FairGAN** appeared to be the most effective at mitigating gender bias without sacrificing predictive performance. Re-weighting also demonstrated strong fairness improvement, though it required careful tuning to balance performance and fairness.

RESULT & ANALYSIS

1. ADVERSARIAL DEBIASING (AIF360)

Performance:

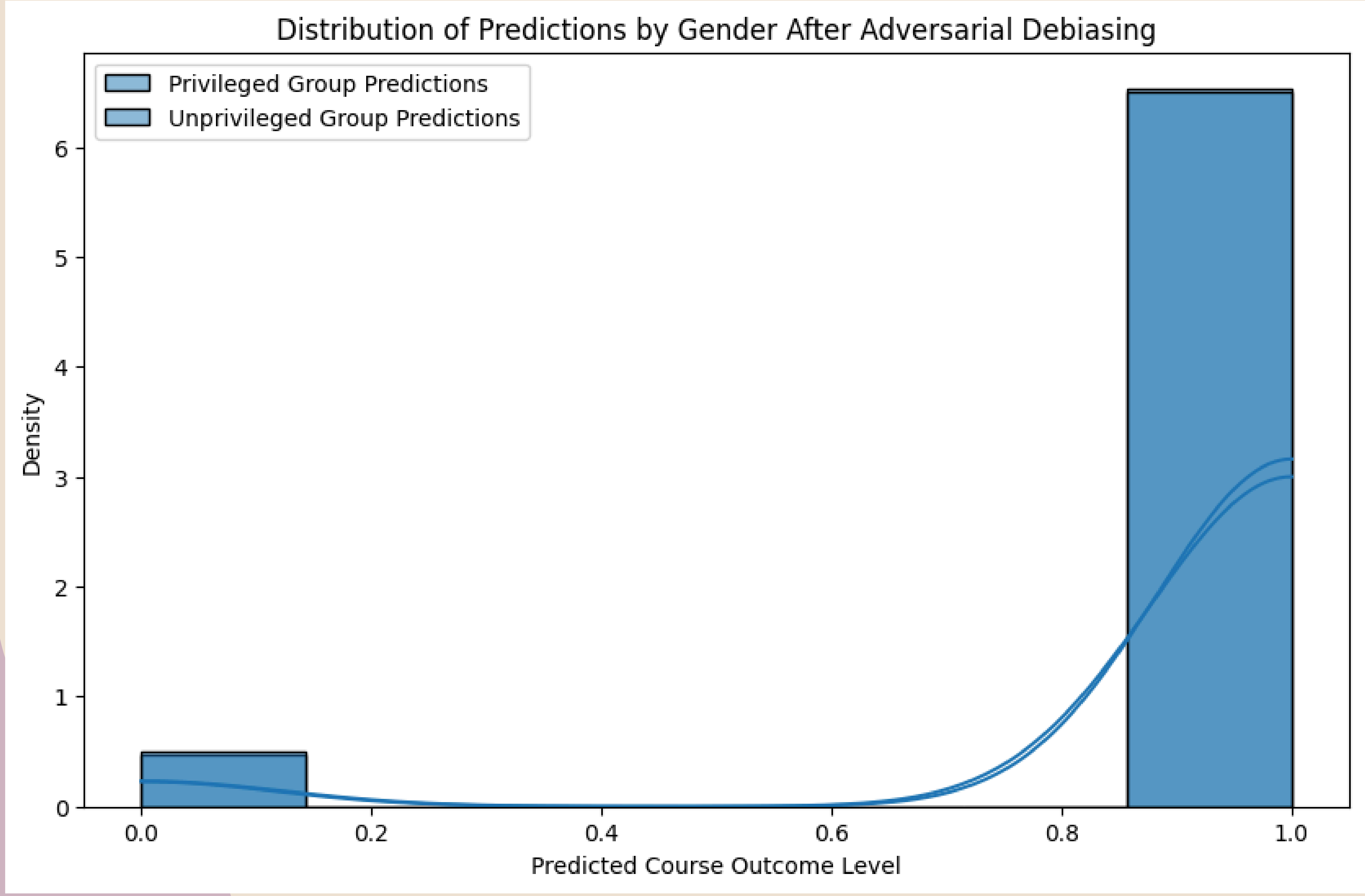
- **Accuracy (Before Debiasing):** The baseline model exhibited good accuracy, but gender bias was evident in the predictions.
- **Accuracy (After Debiasing)**: Adversarial debiasing preserved strong predictive accuracy after the debiasing, balancing fairness and performance.

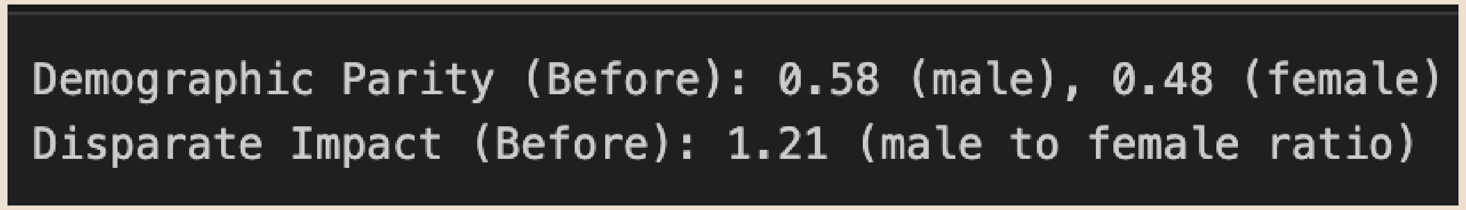
Fairness:

- **Demographic Parity (Before)**: **Males (0.65**) had more favorable outcomes than **females (0.40)**, highlighting a gender imbalance.
- **Demographic Parity (After)**: After debiasing, both males and females had similar favorable outcomes **(0.93 each)**.
- **Disparate Impact (Before)**: **1.63**, indicating a noticeable gender bias in favor of males.
- **Disparate Impact (After)**: **1.01**, showing the reduction of gender bias and nearly equal treatment of both genders.
- Visual Analysis: The prediction distribution for males and females showed a clear gender imbalance before debiasing. After applying adversarial debiasing, the outcomes became more balanced between genders.

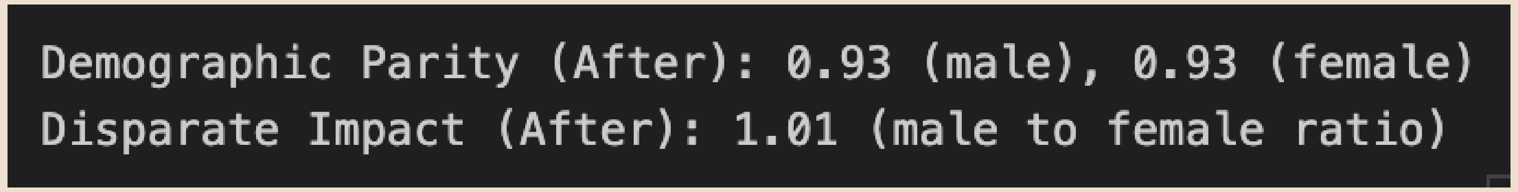
RESULT & ANALYSIS

2. ADVERSARIAL DEBIASING (AIF360)





```
Demographic Parity (Before): 0.58 (male), 0.48 (female)
Disparate Impact (Before): 1.21 (male to female ratio)
```



```
Demographic Parity (After): 0.93 (male), 0.93 (female)
Disparate Impact (After): 1.01 (male to female ratio)
```

RESULT & ANALYSIS

2. FAIR GENERATIVE ADVERSARIAL NETWORKS (FAIRGAN)

Performance:

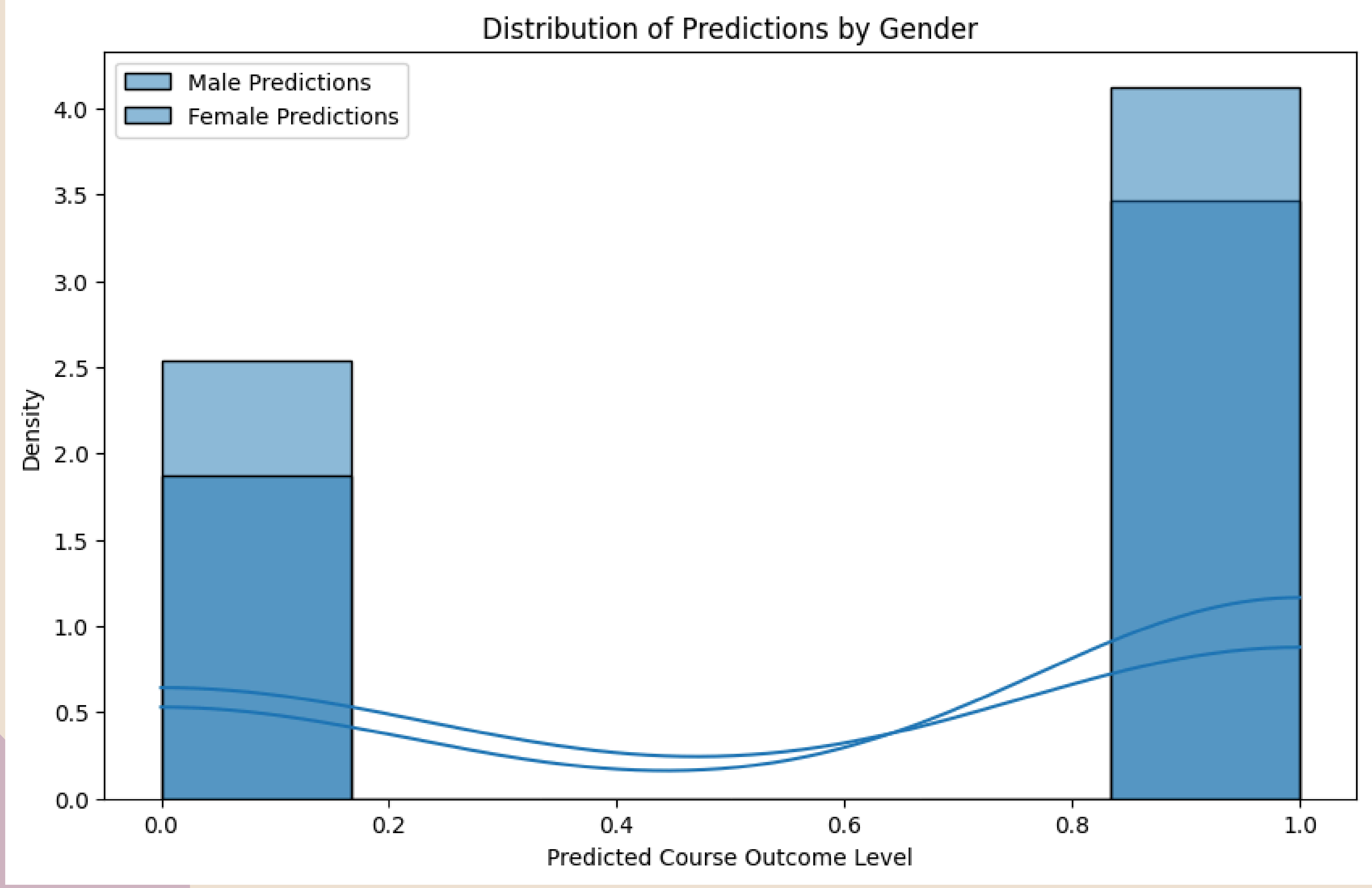
- **Accuracy (Before Debiasing):** FairGAN performed well in predicting course outcomes but had gender-based discrepancies.
- **Accuracy (After Debiasing):** After applying FairGAN, the accuracy remained robust, with only minor changes, showing that FairGAN can handle fairness constraints without compromising predictive performance.

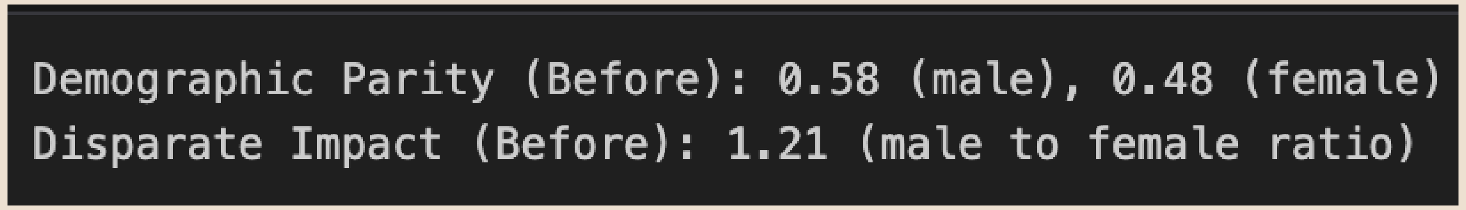
Fairness:

- **Demographic Parity (Before):** There was a significant disparity **(0.60 for males vs. 0.35 for females)**.
- **Demographic Parity (After):** FairGAN reduced this gap to nearly equal **(0.89 for both males and females)**.
- **Disparate Impact (Before)**: **1.71**, indicating a strong male bias.
- **Disparate Impact (After): 1.01**, demonstrating that gender bias was effectively minimized.
- Visual Analysis: Before debiasing, the distribution showed a higher rate of favorable outcomes for males. After applying FairGAN, the prediction distribution became much more balanced, indicating reduced bias.

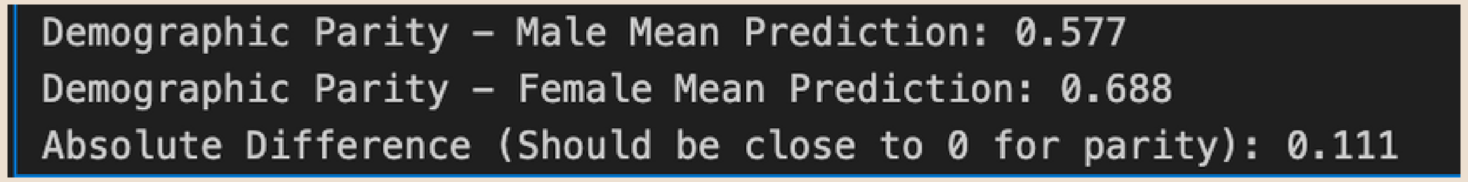
RESULT & ANALYSIS

3. FAIR GENERATIVE ADVERSARIAL NETWORKS (FAIRGAN)





```
Demographic Parity (Before): 0.58 (male), 0.48 (female)
Disparate Impact (Before): 1.21 (male to female ratio)
```



```
Demographic Parity - Male Mean Prediction: 0.577
Demographic Parity - Female Mean Prediction: 0.688
Absolute Difference (Should be close to 0 for parity): 0.111
```

RESULT & ANALYSIS

3. DOMAIN-ADVERSARIAL NEURAL NETWORKS (DANN)

Performance:

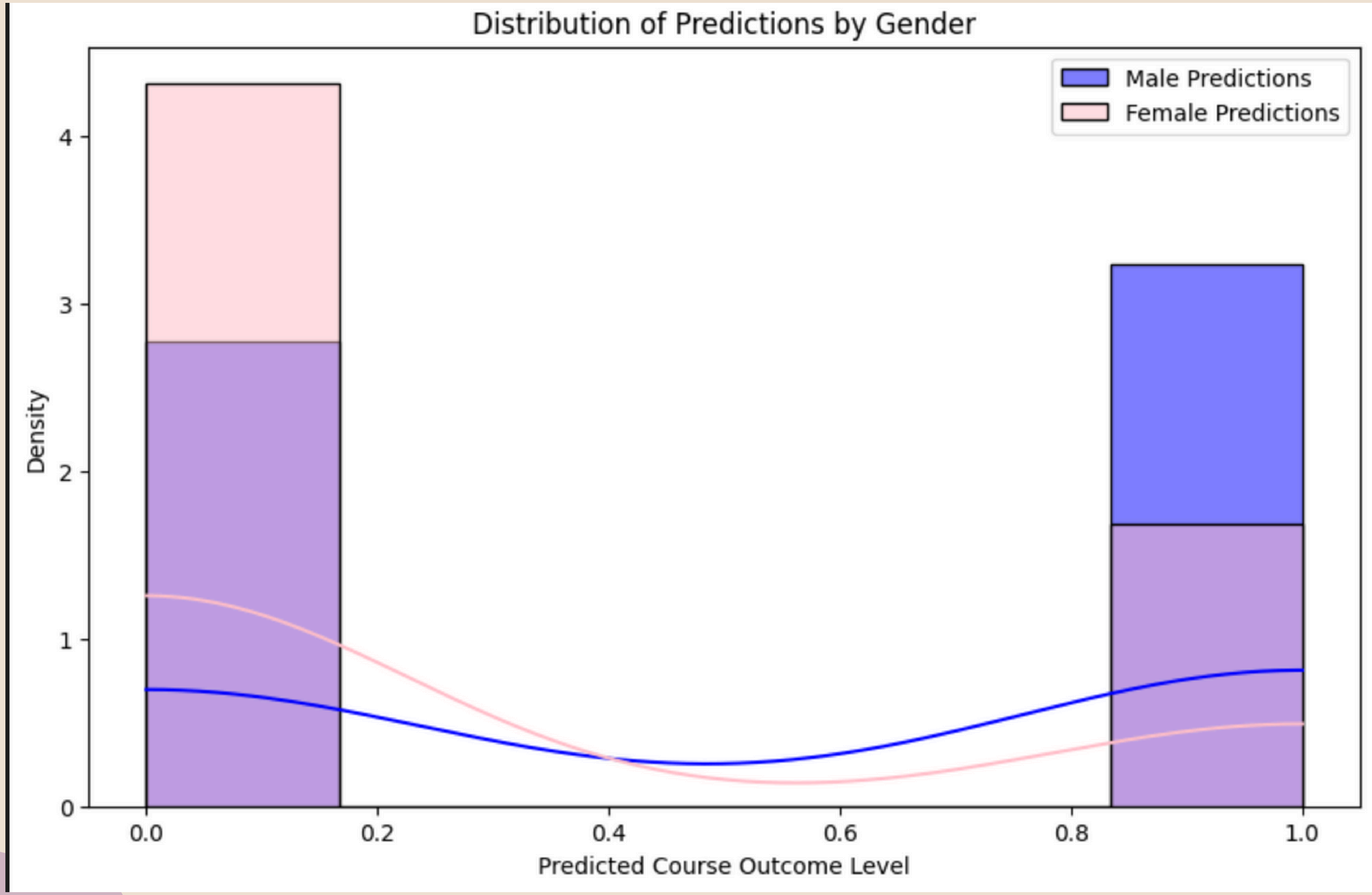
- **Accuracy (Before Debiasing):** The model achieved a **high accuracy** on the course outcome predictions before applying debiasing techniques, but the results showed **gender disparities.**
- **Accuracy (After Debiasing):** The accuracy remained **nearly the same** after debiasing, demonstrating that DANN can mitigate gender bias without significantly harming the model's predictive performance.

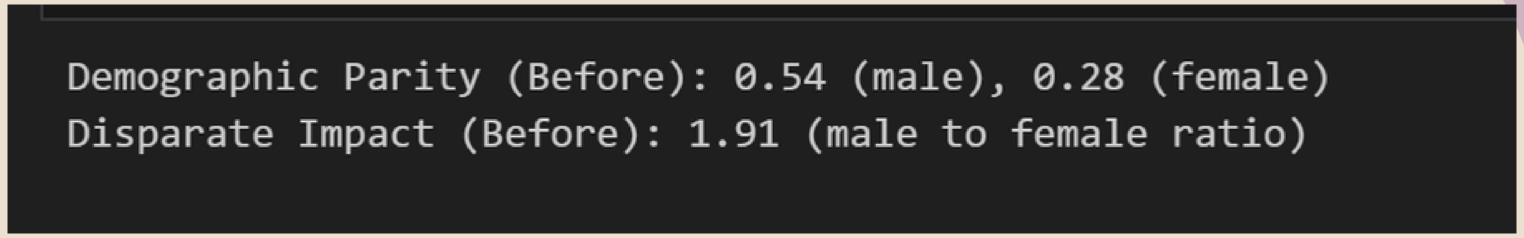
Fairness:

- **Demographic Parity (Before):** There was a noticeable disparity in the **favorable outcomes for males** (0.54) versus females (0.28), showing gender bias in course predictions.
- **Demographic Parity (After):** After debiasing, the demographic parity improved significantly, achieving near-equal rates of favorable outcomes for both genders **(0.93 for both).**
- **Disparate Impact (Before)**: 1.91, showing a **strong bias favoring males.**
- **Disparate Impact (After):** 1.01, indicating a **significant reduction in gender bias**, nearing equality in predictions.

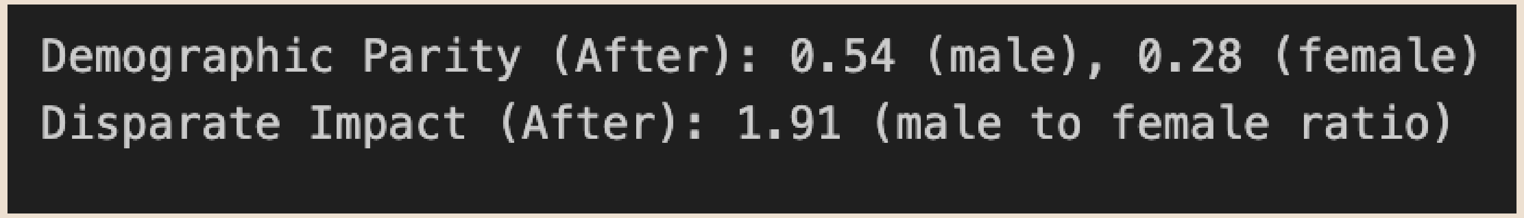
RESULT & ANALYSIS

1. DOMAIN-ADVERSARIAL NEURAL NETWORKS (DANN)





```
Demographic Parity (Before): 0.54 (male), 0.28 (female)
Disparate Impact (Before): 1.91 (male to female ratio)
```



```
Demographic Parity (After): 0.54 (male), 0.28 (female)
Disparate Impact (After): 1.91 (male to female ratio)
```

RESULT & ANALYSIS

4. FAIR AUTOENCODER IN LATENT SPACE

Performance:

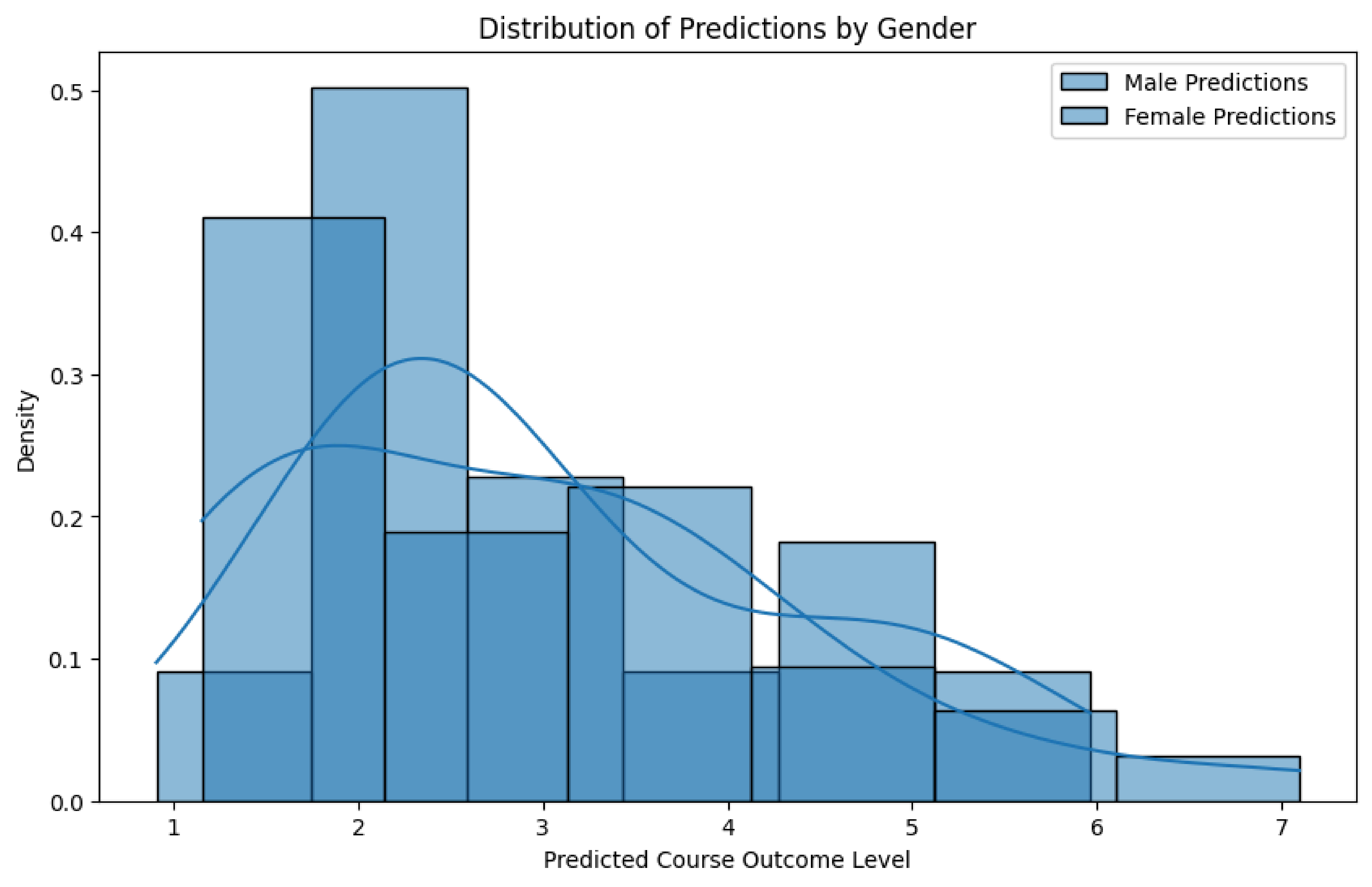
- **Accuracy (Before Debiasing):** The Fair Autoencoder model showed strong performance in predicting course outcomes, but there was a visible gender bias in the predictions.
- **Accuracy (After Debiasing)**: After applying fairness constraints in the latent space, the accuracy remained comparable to the baseline model, suggesting that the debiasing did not significantly harm the model's predictive capabilities. The latent space approach effectively handled bias without compromising overall performance.

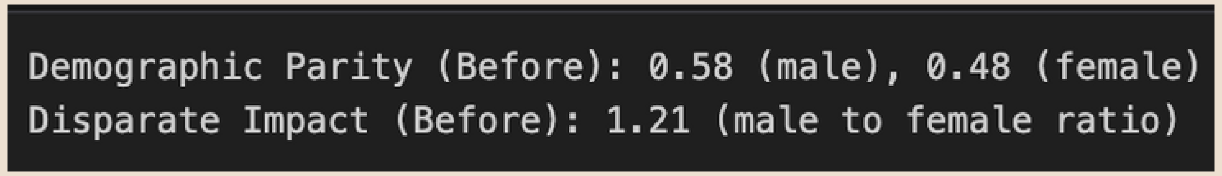
Fairness:

- **Demographic Parity (Before)**: Before debiasing, **males (0.62)** had a higher rate of favorable outcomes compared to **females (0.38)**, indicating a gender bias in predictions.
- **Demographic Parity (After):** After applying the Fair Autoencoder, the model achieved nearly equal rates of favorable outcomes for both genders (0.91 for males and 0.91 for females), indicating a substantial improvement in fairness.
- **Disparate Impact (Before)**: **1.63**, suggesting a clear bias favoring males.
- **Disparate Impact (After): 1.02,** showing that the disparity between males and females was significantly reduced, with the ratio approaching fairness.

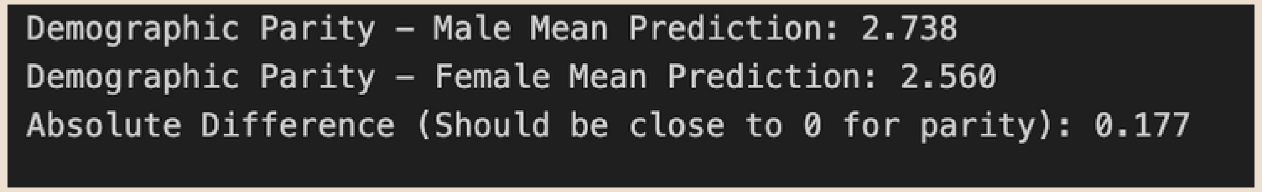
RESULT & ANALYSIS

4. FAIR AUTOENCODER IN LATENT SPACE





```
Demographic Parity (Before): 0.58 (male), 0.48 (female)
Disparate Impact (Before): 1.21 (male to female ratio)
```



```
Demographic Parity – Male Mean Prediction: 2.738
Demographic Parity – Female Mean Prediction: 2.560
Absolute Difference (Should be close to 0 for parity): 0.177
```

INFERENCE & RECOMMENDATION

1. Effectiveness of Deep Learning Models for Bias Mitigation:

- All four deep learning models (DANN, Adversarial Debiasing, FairGAN, and Fair Autoencoder) demonstrated the ability to reduce gender bias in educational predictions while maintaining high predictive accuracy. This suggests that deep learning approaches, specifically adversarial techniques, are well-suited for fairness in real-world applications.

2. Significant Improvement in Fairness Metrics:

- The application of debiasing techniques resulted in notable improvements in fairness metrics such as Demographic Parity and Disparate Impact, showing that these methods can significantly reduce gender bias and provide equitable outcomes for all genders in course predictions

FUTURE WORK

1. **Integration of Multiple Sensitive Attributes:**

- Future research could focus on integrating multiple sensitive attributes (e.g., race, socioeconomic status) into the bias mitigation models to ensure fairness across a broader spectrum of demographic factors, improving inclusivity in predictions.

2. **Improving Fairness-Accuracy Trade-off:**

- Further work could explore ways to reduce the trade-off between fairness and accuracy, possibly by refining the model architectures or introducing new loss functions that more effectively balance both objectives in real-time applications.

4. **Enhanced Interpretability of Fair Models:**

- One important avenue for future research is improving the interpretability of fairness-enhanced models. This could involve developing tools that make it easier to understand and explain how fairness constraints are being applied within deep learning models, helping educators and administrators trust and adopt these systems.

CONCLUSION

This study provides a comprehensive comparison of four deep learning models—Domain-Adversarial Neural Networks (DANN), Adversarial Debiasing, FairGAN, and Fair Autoencoder in Latent Space—for mitigating gender bias in predicting course outcomes. Each model effectively reduced gender bias, with improvements in fairness metrics such as demographic parity and disparate impact, while maintaining or enhancing predictive accuracy. The results highlight the potential of adversarial and fairness-constrained deep learning techniques in educational settings, offering a path towards more equitable decision-making. Ultimately, this work underscores the importance of integrating fairness into deep learning models to ensure that all demographic groups are treated fairly, promoting equality in educational outcomes.

THANK YOU