Casual Model for Algorithmic Bias Detection and Mitigation in Convolutional Neural Networks

Min Sik Byun
Singapore Institution of Technology
Singapore
2102150@sit.singaporetech.edu.sg

line 1: 2nd Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

line 1: 3rd Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

This study explores algorithmic bias in Convolutional Neural Networks (CNNs) for emotion classification, focusing on gender as a protected attribute. Using the FairFace dataset, complemented with emotion labels generated by the DeepFace pre-trained model, a custom CNN achieved a baseline accuracy of 65%. Structural Equation Modeling (SEM) and a causal framework were employed to identify gender-related biases in the model's predictions, revealing disparities in classification accuracy for certain emotions. To address these biases, a SoftMax-based adjustment method was proposed, which effectively reduced gender disparities in predictions. This work highlights the importance of fairness-aware AI design and provides a framework for detecting and mitigating bias in machine learning systems. Future work includes improving labelling accuracy, refining CNN architectures, and exploring alternative input representations to enhance fairness and reliability in emotion classification.

Keywords: *AI Fairness, Casual Modelling, Bias Detection, Bias Mitigation, Convolutional Neural Network*

I. Introduction

Artificial Intelligence (AI) has permeated various sectors, transforming decision-making processes in domains such as healthcare, finance, and law enforcement. While AI systems promise efficiency and scalability, a growing concern revolves around their fairness and ethical implications. Central to this concern is the question: Is AI truly fair? Biases in AI systems can originate from multiple sources, primarily input data and algorithms. For instance, training datasets often reflect historical inequalities, which are inadvertently encoded into AI systems. Moreover, algorithms themselves may exhibit tendencies that disadvantage specific demographic groups, raising critical questions about fairness in social and demographically protected attributes such as gender, race, and religion [1][2].

The fairness of AI systems depends heavily on how biases are identified and mitigated. While some efforts focus on curating unbiased input data, this approach poses significant challenges. Ensuring equal representation across various demographic categories—such as age, race, gender, and religion—during each iteration of model training is not only labour-intensive but also difficult to standardize [3]. Moreover, even with balanced datasets, algorithms might inherently favour certain groups due to their design or optimization objectives. This issue highlights the importance of understanding and mitigating algorithmic biases, which often prove more elusive and pervasive [4].

Among machine learning models, convolutional neural networks (CNNs) have gained prominence, particularly in image recognition and classification tasks. However, studies exploring algorithmic biases have predominantly focused on traditional regression-based algorithms, leaving gaps in understanding biases within deep learning models like CNNs. CNNs, due to their complex architectures and inherent learning dynamics, may exhibit unique biases that demand further investigation. These biases not only affect the accuracy of predictions but also have profound implications for equity and ethical decision-making [5][6].

This paper aims to address this gap by examining algorithmic biases in CNNs using a causal model approach. By identifying causal pathways through which biases emerge and persist, the study seeks to provide a systematic framework for analysing and mitigating biases in neural networks. This exploration will contribute to the broader discourse on ensuring fairness and accountability in AI systems, fostering trust and equity in their applications.

II. Related Work

The challenge of reducing algorithmic bias in machine learning has been a topic of extensive research, with efforts generally categorized into three key stages: pre-processing, in-processing, and post-processing [7]. Each of these approaches tackles bias at a specific point in the machine learning pipeline, targeting its sources and manifestations. Additionally, causal models have emerged as a promising framework for addressing bias by examining its root causes. Below, we discuss each approach in detail and highlight relevant advancements.

A. Pre-Processing Approaches

Pre-processing methods aim to mitigate bias by modifying the training data before model training begins. These techniques include re-sampling, re-weighting, and altering data representations to balance demographic distributions and reduce disparities in feature correlations [8]. While traditional methods focus on ensuring fairness in data, recent advancements have introduced innovative strategies such as domain adaptation and latent space de-biasing.

Joshi and Burlina (2020) applied domain adaptation techniques to address fairness in healthcare applications, specifically for detecting age-related macular degeneration (AMD). Their method adjusted data distributions across demographic groups to reduce disparities in prediction performance, leading to more equitable outcomes [9].

Similarly, Sharma et al. (2020) proposed data augmentation methods to generate synthetic samples, addressing demographic imbalances and mitigating bias while preserving model accuracy. This approach is particularly effective in domains where underrepresented groups are a concern [10]. These methods emphasize the importance of fairness at the data preparation stage but often require significant expertise and computational resources.

B. In-Processing Approaches

In-processing approaches integrate fairness into the model training process by directly modifying the learning algorithm. These methods aim to achieve a balance between accuracy and fairness, often through the use of fairness constraints or adversarial techniques. For example, adversarial training has been employed to obscure sensitive attributes during learning, ensuring that the model cannot infer these attributes and thereby reducing bias [11] Moreover, hybrid methods that combine data balancing with fairness-aware training have shown promise. For instance, generative models have been used to address data imbalances while simultaneously minimizing dependencies between sensitive attributes and model predictions [12]. While in-processing techniques often yield robust fairness outcomes, they tend to be computationally intensive and may require significant modifications to standard training procedures.

C. Post-Processing Approaches

Post-processing methods focus on bias mitigation after the model has been trained, treating the model as a black box. These approaches adjust the outputs of the model to align with fairness metrics, such as equalized odds or equal opportunity, ensuring equitable predictive outcomes across demographic groups [13]. Extensions of post-processing techniques have been developed for multiclass classification tasks, where fairness criteria are adapted to address the complexities of multi-category outputs [14]. While post-processing is flexible and computationally efficient, it can sometimes face challenges related to interpretability and may introduce trade-offs in performance consistency.

D. One versus All Technique

The One-versus-All (OvA) technique is a widely used approach in multiclass classification problems, where a series of binary classifiers are trained, each dedicated to distinguishing a single class from all others. Rifkin and Klautau (2004) critically evaluated the effectiveness of this method, emphasizing its simplicity and scalability when implemented with robust classifiers like Support Vector Machines (SVMs) [15]. In this paper, this technique was used in apply casual model to multiclass emotion classifier.

E. Casual Models in Fairness

Causal models provide a structured framework for identifying and addressing bias by examining the relationships between features, sensitive attributes, and predictions. Unlike traditional fairness methods, which often rely on statistical correlations, causal models delve into the root causes of bias, offering deeper insights and more targeted interventions.

Madras et al. [16] demonstrated the potential of causal modelling to improve prediction accuracy in the presence of confounding factors using health datasets. Similarly, Khademi et al. [17] used causal models to identify and quantify biases in both synthetic and real-world datasets. These studies underscore the potential of causal models for bias detection and analysis.

Hui, W. and W. K. Lau (2023) [18] extends the application of causal models to both bias detection and correction, with a focus on regression algorithms. Their method integrates causal reasoning into post-processing workflows, leveraging causal pathways to identify and address disparities in outputs effectively. However, the scope of their work is limited to regression tasks and does not address neural networks or complex architectures like Convolutional Neural Networks (CNNs).

Building on this foundation, our work advances the use of causal models by applying them to bias mitigation in CNNs. By leveraging causal reasoning, we aim to address biases in neural network predictions, particularly in post-processing, to ensure fairness without sacrificing performance. This novel application extends the reach of causal frameworks to deep learning models, where bias mitigation remains an ongoing challenge.

III. Methodology

This section outlines the methodology adopted to detect and mitigate biases in Convolutional Neural Networks (CNNs) for an emotion classification task.

A. Dataset

The FairFace dataset [19] was selected due to its balanced representation of race, gender, and age attributes. FairFace contains over 108,000 images, each labeled with demographic attributes such as race, gender, and age. By minimizing input bias concerning race, it allows for a focused evaluation of gender-related bias in CNN models. Its equitable representation of diverse demographic groups makes it an ideal choice for fairness studies in machine learning.

B. Data Labelling for Ground Truth

Since the FairFace dataset does not include emotion as a predefined attribute, we labelled the images with probabilities for seven emotion categories using the DeepFace pre-trained model. The emotion categories are: angry, happy, sad, fear, neutral, surprise, and disgust.

This automated labelling process was applied to the entire dataset, ensuring consistent emotion annotations across training, validation, and test sets. Fig. 1 illustrates sample visualizations from the labelled dataset.

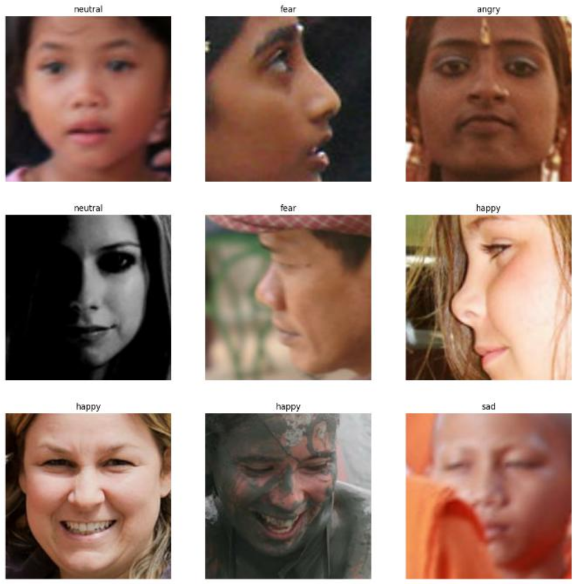


Fig. 1. A batch from of Emotion Lablled Dataset

C. Data Splitting

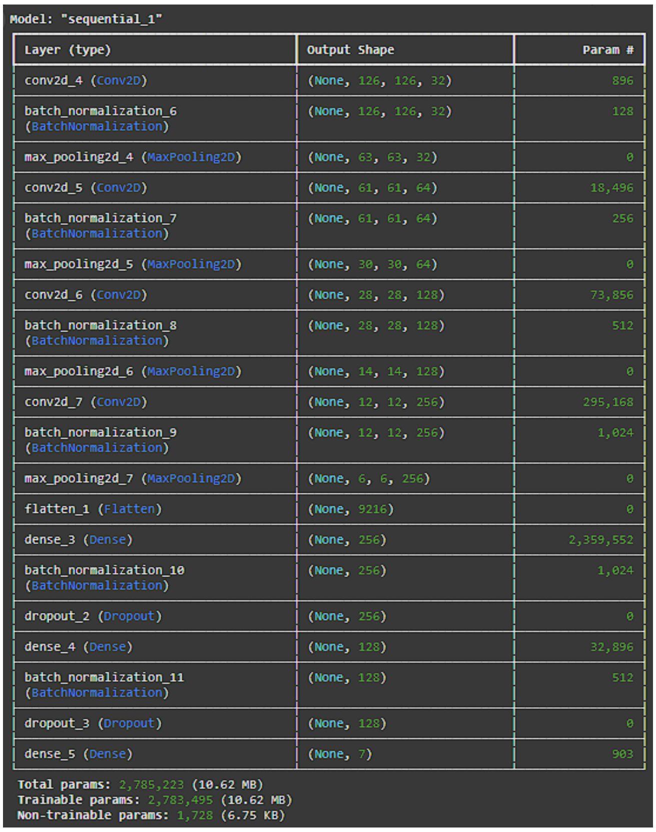
The labelled dataset was split into training, validation, and test sets to facilitate model evaluation. The splitting was performed in a stratified manner to maintain the proportion of gender and emotion categories in each subset. TABLE I. details the dataset split. The overall gender distribution in the combined dataset is: Male: 53%, Female: 47%.

TABLE I. DATASET DESCRIPTION

Dataset Component	Size (rows)	Description
Training Dataset	86,744	Used to train the emotion classifier
Validation Dataset	5477	Used for validation
Combined Dataset	97,698	Unified dataset used for emotion labelling
Test Dataset	5477	Used to evaluate and predict emotion classification

D. Model Architecture

A custom CNN model was developed for emotion classification. The architecture consists of four convolutional layers, progressively increasing filters (32, 64, 128, 256) for hierarchical feature extraction, each followed by batch normalization and max-pooling for stabilization and dimensionality reduction. The extracted 3D feature maps are flattened into a 1D vector, passed through two dense layers (256 and 128 units) with dropout for regularization, and finally outputted through a dense layer with 7 units for classification. It has 2.78M trainable parameters, efficiently balancing feature learning and prediction. Fig. 2 provides a detailed architecture diagram.



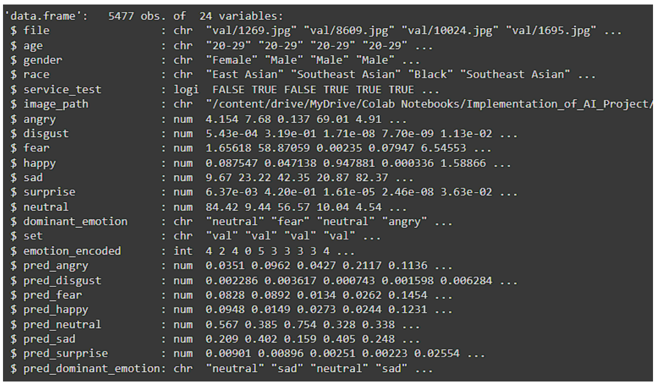
```
Model: "sequential_1"
┌──────────────────────────────────┬────────────────────────┬──────────────┐
│ Layer (type)                     │ Output Shape           │      Param # │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ conv2d_4 (Conv2D)                │ (None, 126, 126, 32)   │          896 │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ batch_normalization_6            │ (None, 126, 126, 32)   │          128 │
│ (BatchNormalization)             │                        │              │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ max_pooling2d_4 (MaxPooling2D)   │ (None, 63, 63, 32)     │            0 │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ conv2d_5 (Conv2D)                │ (None, 61, 61, 64)     │       18,496 │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ batch_normalization_7            │ (None, 61, 61, 64)     │          256 │
│ (BatchNormalization)             │                        │              │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ max_pooling2d_5 (MaxPooling2D)   │ (None, 30, 30, 64)     │            0 │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ conv2d_6 (Conv2D)                │ (None, 28, 28, 128)    │       73,856 │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ batch_normalization_8            │ (None, 28, 28, 128)    │          512 │
│ (BatchNormalization)             │                        │              │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ max_pooling2d_6 (MaxPooling2D)   │ (None, 14, 14, 128)    │            0 │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ conv2d_7 (Conv2D)                │ (None, 12, 12, 256)    │      295,168 │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ batch_normalization_9            │ (None, 12, 12, 256)    │        1,024 │
│ (BatchNormalization)             │                        │              │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ max_pooling2d_7 (MaxPooling2D)   │ (None, 6, 6, 256)      │            0 │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ flatten_1 (Flatten)              │ (None, 9216)           │            0 │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ dense_3 (Dense)                  │ (None, 256)            │    2,359,552 │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ batch_normalization_10           │ (None, 256)            │        1,024 │
│ (BatchNormalization)             │                        │              │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ dropout_2 (Dropout)              │ (None, 256)            │            0 │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ dense_4 (Dense)                  │ (None, 128)            │       32,896 │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ batch_normalization_11           │ (None, 128)            │          512 │
│ (BatchNormalization)             │                        │              │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ dropout_3 (Dropout)              │ (None, 128)            │            0 │
├──────────────────────────────────┼────────────────────────┼──────────────┤
│ dense_5 (Dense)                  │ (None, 7)              │          903 │
└──────────────────────────────────┴────────────────────────┴──────────────┘
 Total params: 2,785,223 (10.62 MB)
 Trainable params: 2,783,495 (10.62 MB)
 Non-trainable params: 1,728 (6.75 KB)
```

Fig. 2. Custom CNN Model Architecture

E. Training and Testing

The model was trained using the Adam optimizer with a learning rate of 0.0001 and a categorical cross-entropy loss function. An early stopping mechanism was implemented, monitoring the validation loss with a patience of five epochs to prevent overfitting. The model was trained for a maximum of 10 epochs, balancing computational efficiency and performance.

After training, the model's performance was evaluated on the test set. For each test sample, the model produced probabilities for the seven emotion categories, and the emotion with the highest probability was selected as the predicted label. Fig. 3 shows sample outputs, including predicted probabilities, predicted emotions, and ground truth labels.

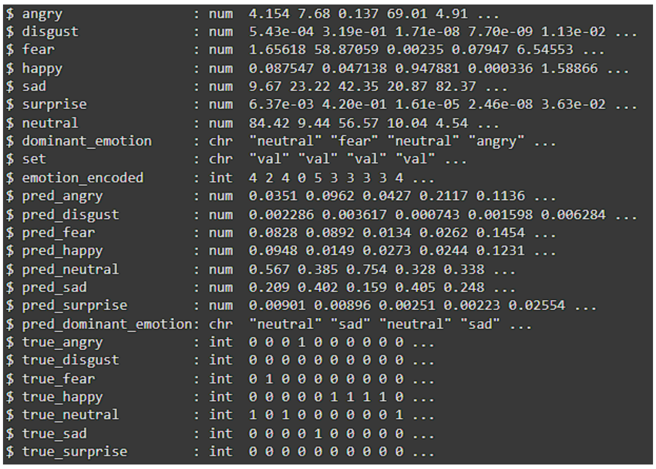


```
'data.frame':   5477 obs. of  24 variables:
 $ file                 : chr  "val/1269.jpg" "val/8609.jpg" "val/10024.jpg" "val/1695.jpg" ...
 $ age                  : chr  "20-29" "20-29" "20-29" "20-29" ...
 $ gender               : chr  "Female" "Male" "Male" "Male" ...
 $ race                 : chr  "East Asian" "Southeast Asian" "Black" "Southeast Asian" ...
 $ service_test         : logi  FALSE TRUE FALSE TRUE TRUE TRUE ...
 $ image_path           : chr  "/content/drive/MyDrive/Colab Notebooks/Implementation_of_AI_Project/
 $ angry                : num  4.154 7.68 0.137 69.01 4.91 ...
 $ disgust              : num  5.43e-04 3.19e-01 1.71e-08 7.70e-09 1.13e-02 ...
 $ fear                 : num  1.65618 58.87059 0.00235 0.07947 6.54553 ...
 $ happy                : num  0.087547 0.047138 0.947881 0.000336 1.58866 ...
 $ sad                  : num  9.67 23.22 42.35 20.87 82.37 ...
 $ surprise             : num  6.37e-03 4.20e-01 1.61e-05 2.46e-08 3.63e-02 ...
 $ neutral              : num  84.42 9.44 56.57 10.04 4.54 ...
 $ dominant_emotion     : chr  "neutral" "fear" "neutral" "angry" ...
 $ set                  : chr  "val" "val" "val" "val" ...
 $ emotion_encoded      : int  4 2 4 0 5 3 3 3 3 4 ...
 $ pred_angry           : num  0.0351 0.0962 0.0427 0.2117 0.1136 ...
 $ pred_disgust         : num  0.002286 0.003617 0.000743 0.001598 0.006284 ...
 $ pred_fear            : num  0.0828 0.0892 0.0134 0.0262 0.1454 ...
 $ pred_happy           : num  0.0948 0.0149 0.0273 0.0244 0.1231 ...
 $ pred_neutral         : num  0.567 0.385 0.754 0.328 0.338 ...
 $ pred_sad             : num  0.209 0.402 0.159 0.405 0.248 ...
 $ pred_surprise        : num  0.00901 0.00896 0.00251 0.00223 0.02554 ...
 $ pred_dominant_emotion: chr  "neutral" "sad" "neutral" "sad" ...
```

Fig. 3. Dataframe of Predicted Test Dataset

F. Computing Accuracy for Emotions and Gender

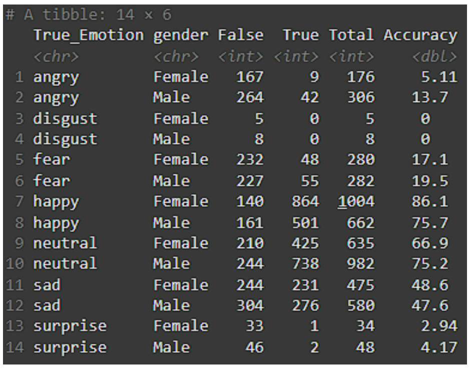
To compute accuracy metrics for each emotion, the true emotion classes underwent a one-hot encoding process. In this encoding, each sample's "dominant emotion" was assigned a binary value of 1 for the correct emotion and 0 for others. This transformation facilitated comparison between the true labels and the predicted labels. An example of the encoded structure is shown in Fig. 4.



```
$ angry                : num  4.154 7.68 0.137 69.01 4.91 ...
$ disgust              : num  5.43e-04 3.19e-01 1.71e-08 7.70e-09 1.13e-02 ...
$ fear                 : num  1.65618 58.87059 0.00235 0.07947 6.54553 ...
$ happy                : num  0.087547 0.047138 0.947881 0.000336 1.58866 ...
$ sad                  : num  9.67 23.22 42.35 20.87 82.37 ...
$ surprise             : num  6.37e-03 4.20e-01 1.61e-05 2.46e-08 3.63e-02 ...
$ neutral              : num  84.42 9.44 56.57 10.04 4.54 ...
$ dominant_emotion     : chr  "neutral" "fear" "neutral" "angry" ...
$ set                  : chr  "val" "val" "val" "val" ...
$ emotion_encoded      : int  4 2 4 0 5 3 3 3 3 4 ...
$ pred_angry           : num  0.0351 0.0962 0.0427 0.2117 0.1136 ...
$ pred_disgust         : num  0.002286 0.003617 0.000743 0.001598 0.006284 ...
$ pred_fear            : num  0.0828 0.0892 0.0134 0.0262 0.1454 ...
$ pred_happy           : num  0.0948 0.0149 0.0273 0.0244 0.1231 ...
$ pred_neutral         : num  0.567 0.385 0.754 0.328 0.338 ...
$ pred_sad             : num  0.209 0.402 0.159 0.405 0.248 ...
$ pred_surprise        : num  0.00901 0.00896 0.00251 0.00223 0.02554 ...
$ pred_dominant_emotion: chr  "neutral" "sad" "neutral" "sad" ...
$ true_angry           : int  0 0 0 1 0 0 0 0 0 0 ...
$ true_disgust         : int  0 0 0 0 0 0 0 0 0 0 ...
$ true_fear            : int  0 1 0 0 0 0 0 0 0 0 ...
$ true_happy           : int  0 0 0 0 0 1 1 1 1 0 ...
$ true_neutral         : int  1 0 1 0 0 0 0 0 0 1 ...
$ true_sad             : int  0 0 0 0 1 0 0 0 0 0 ...
$ true_surprise        : int  0 0 0 0 0 0 0 0 0 0 ...
```

Fig. 4. Test Dataset after One-hot Encoding

Accuracy was calculated by comparing the predicted labels to the true labels, grouped by gender (male and female). This allowed us to assess the model's performance across different demographic groups. Fig. 5 and Fig. 6 visualize the classification accuracies for each emotion category and gender.



```
# A tibble: 14 × 6
  True_Emotion gender False  True Total Accuracy
  <chr>        <chr>  <int> <int> <int>    <dbl>
1 angry        Female   167     9   176     5.11
2 angry        Male     264    42   306    13.7
3 disgust      Female     5     0     5     0
4 disgust      Male       8     0     8     0
5 fear         Female   232    48   280    17.1
6 fear         Male     227    55   282    19.5
7 happy        Female   140   864  1004    86.1
8 happy        Male     161   501   662    75.7
9 neutral      Female   210   425   635    66.9
10 neutral     Male     244   738   982    75.2
11 sad         Female   244   231   475    48.6
12 sad         Male     304   276   580    47.6
13 surprise    Female    33     1    34     2.94
14 surprise    Male      46     2    48     4.17
```

Fig. 5. Accuracy by Gender and Emotion

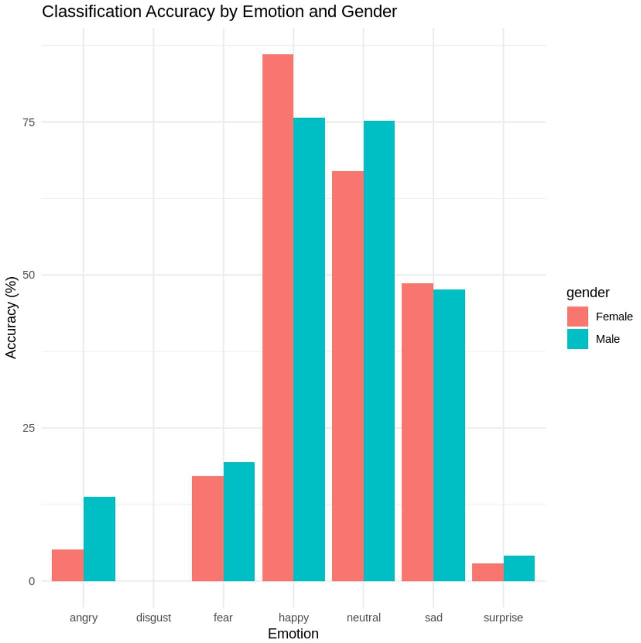
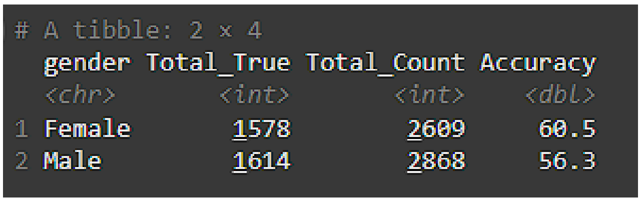


Fig. 6. Classification accuracy by emotion and gender

The overall classification accuracy by gender is presented in Fig. 7.



```
# A tibble: 2 × 4
  gender Total_True Total_Count Accuracy
  <chr>       <int>       <int>    <dbl>
1 Female       1578        2609     60.5
2 Male         1614        2868     56.3
```

Fig. 7. Accuracy by gender

G. Causal Modelling

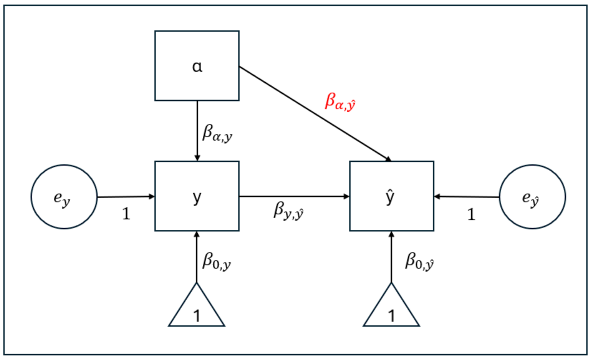


Fig. 8. Casual Model

To analyse potential bias in the model's predictions, we constructed a causal model focusing on gender as the protected attribute, denoted by α (gender binary). In the causal diagram shown in Fig. 8, observable variables, such as the protected attribute (α), the true emotion (y), and the predicted emotion ($e_{\hat{y}}$), are represented as rectangles. Latent or unobservable variables, such as the error terms (e_y and $e_{\hat{y}}$), are depicted as circles or ovals. These error terms account for noise or unmeasured factors that influence the outcomes and are assumed to have a mean of zero. The diagram can be described with following mathematical equations:

$$\hat{y} = e_{\hat{y}} + \beta_{y,\hat{y}} y + \beta_{0,\hat{y}} + \beta_{a,\hat{y}} a \tag{1}$$

$$y = e_y + \beta_{a,y}\, a + \beta_{0.y} \tag{2}$$

In the model:

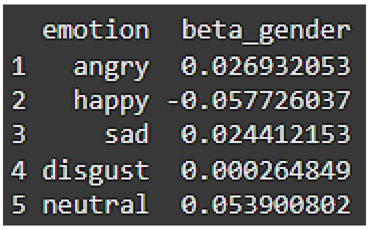
- The protected attribute (α) directly affects both the true emotion (y) and the predicted emotion ($\hat{y}$), as indicated by the arrows from α to y and α to $\hat{y}$.
- The true emotion (y) affects the predicted emotion ($\hat{y}$) through the path.
- Importantly, $\hat{y}$ is not modelled to influence α or y, maintaining the unidirectional nature of the relationships.

The path coefficients ($\beta_{\alpha,y}$, $\beta_{y,\hat{y}}$, and $\beta_{\alpha,\hat{y}}$) quantify the causal impact of gender on the true and predicted emotions. By estimating these coefficients, we can assess the extent to which gender contributes to potential bias in the model's predictions. TABLE II. summarizes the values and significance of $\beta_{\alpha,\hat{y}}$ for each emotion category.

TABLE II. PATH ANALYSIS RESULT OF EMOTIONS

Emotions	est. $\beta_{\alpha,\hat{y}}$ value	p-value
Angry	0.026932053	0.000000e+00
Happy	-0.05772604	0
Sad	0.02441215	2.643155e-09
Fear	0.0001316176	0.9561599
Disgust	2.648490e-04	1.600250e-03
Neutral	0.05390080	0.000000e+00
Surprise	-0.0007961771	1.994572e-01

Algorithmic bias can be assessed by examining the coefficient $\beta_{\alpha,\hat{y}}$, which represents the direct influence of the protected attribute (α) on the predicted emotion ($\hat{y}$) If $\beta_{\alpha,\hat{y}}$ is significantly different from 0, it indicates that the model's predictions are biased with respect to the protected attribute. In social science, p-value less than 0.05 is used as a threshold to determine the significance. Fig. 9 shows emotions which the biases have been detected.



```
  emotion  beta_gender
1   angry  0.026932053
2   happy -0.057726037
3     sad  0.024412153
4 disgust  0.000264849
5 neutral  0.053900802
```

Fig. 9. Emotions with Bias

H. Mitigation of Algorithmic Bias

To mitigate this bias, we estimate a bias-adjusted predicted value, denoted as $\tilde{y}$. This is achieved by removing the effect of the protected attribute (α) from the prediction equation. Specifically, we set $\beta_{\alpha,\hat{y}} = 0$ in the causal model, thereby recalculating $\tilde{y}$ in a counterfactual scenario where the protected attribute has no influence on the prediction.

This can be represented as following equation:

$$\tilde{y} = e_{\hat{y}} + \beta_{y,\hat{y}} y + \beta_{0,\hat{y}} \tag{3}$$

This can be represented in this format:

$$\tilde{y} = \hat{y} - \beta_{a,\hat{y}} a \tag{4}$$

After subtracting the estimated gender influence, $\tilde{y}$ may fall outside the range [0,1]. To ensure that they remain valid probability, we perform min-max scaling to rescale the adjusted values back to the original scale:

$$\hat{y}_{i,e}^{rescaled} = min(\hat{y}_e) + \left(\frac{\tilde{y}_{i,e} - min(\tilde{y}_e)}{max(\widetilde{y}_e) - min(\widetilde{y}_e)}\right)(max(\widehat{y}_e) - min(\widehat{y}_e)) \tag{5}$$

On the other hand, for emotions which the biases are not introduced, the predicted probabilities remains the same.

To ensure the adjusted de-biased probabilities and non-biased probabilities form a valid probability distribution (non-negative and summing to 1), we apply the SoftMax function:

$$\tilde{p}_{i,e} = \frac{\exp(\hat{y}_{i,e}^{rescaled})}{\sum_{k=1}^{K} \exp(\hat{y}_{i,e}^{rescaled})} \tag{6}$$

where $K$is the total number of emotions.

Then, the final de-biased predicted emotion for individual i after mitigation is determined by selecting the emotion with the highest adjusted probability:

$$PredictedEmotion_i^{adj} = \underset{e}{arg\,max}\, \tilde{p}_{i,e} \tag{7}$$

IV. Results and Analysis

This section presents the evaluation of gender bias in the CNN model before and after the application of the bias mitigation technique described in Section III.G. By comparing classification accuracies across genders for each emotion category, we assess the effectiveness of the mitigation strategy in improving the fairness of the model's outputs.

A. Evaluation Metrics

To quantify the model's performance and fairness, we computed the classification accuracy for each emotion category separately for male and female groups. Accuracy is defined as the ratio of correctly predicted instances to the total number of instances for each gender within each emotion category. Additionally, we calculated the accuracy gap between genders, which is the absolute difference in accuracies between male and female groups for each emotion.

B. Comparison of Pre-Mitigation and Post Mitigation

TABLE III. summarizes the change in classification accuracies for males and females before and after applying the bias mitigation method.

TABLE III. Accuracy Analysis By Emotions and Gender

Emotions	Gender	Acc Before (%)	Acc After (%)	Diff Before	Diff After	Improvement
Angry	F	5.11	6.82	8.59	7.28	T
	M	13.7	14.1			
Disgust	F	0	0	0	0	-
	M	0	0			
Fear	F	17.1	14.3	2.4	4.8	F
	M	19.5	19.1			
Happy	F	86.1	84.9	10.4	6.5	T
	M	75.7	78.4			
Neutral	F	66.9	70.4	8.3	2.9	T
	M	75.2	73.3			
Sad	F	48.6	48.2	1	0.4	T
	M	47.6	47.8			
Surprise	F	2.94	2.94	1.23	1.23	-
	M	4.17	4.17			

The results demonstrates a general reduction in accuracy gaps between genders. However, for the Fear emotion, the accuracy gap increased slightly after mitigation. This indicates that while the mitigation strategy improved fairness for most emotions, it did not have the desired effect for all categories. Fig. 10 illustrates improvement in the overall accuracy.

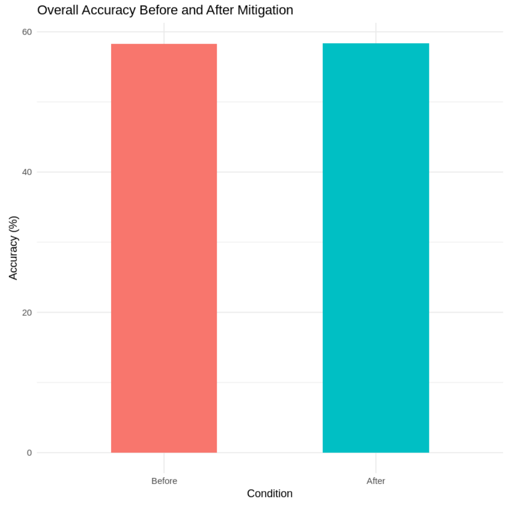


Fig. 10. Emotions with Bias

V. CONCLUSION

This study investigated algorithmic bias in Convolutional Neural Networks (CNNs) for emotion classification, focusing on gender as the protected attribute. By leveraging the FairFace dataset—known for its balanced representation of race, gender, and age—we minimized input bias related to race, allowing for a concentrated analysis of gender-related biases. Since FairFace does not include emotion labels, we utilized the DeepFace pre-trained model to annotate the dataset with probabilities for seven emotion categories: angry, happy, sad, fear, neutral, surprise, and disgust. This augmentation provided a consistent and comprehensive dataset for training and evaluation.

We developed a custom CNN model tailored for emotion classification, consisting of four convolutional blocks followed by fully connected layers and dropout layers to mitigate overfitting. The model was trained using the Adam optimizer with a learning rate of 0.0001 and employed an early stopping mechanism based on validation loss. Achieving a baseline accuracy of 65%, the model demonstrated effectiveness for the task while highlighting opportunities for further improvement.

To detect and quantify biases in the model's predictions, we employed Structural Equation Modelling (SEM) and causal modelling techniques. By constructing a causal diagram that included the protected attribute (gender), true emotions, and predicted emotions, we identified the pathways through which gender could influence the model's outputs. The path coefficients obtained from the SEM analysis revealed that gender bias significantly affected the predicted emotions in certain cases, underscoring the need for mitigation strategies.

We introduced a bias mitigation method based on probability adjustment to address these biases. Specifically, we modified the causal model by removing the direct effect of the protected attribute on the predicted emotion, effectively eliminating gender influence from the predictions. The adjusted probabilities were then rescaled using min-max scaling and normalized with the SoftMax function to ensure they formed a valid probability distribution. This approach recalibrated the predicted probabilities, reducing disparities across gender groups.

The results demonstrated that the bias mitigation technique effectively reduced the accuracy gaps between male and female groups for most emotion categories, thereby improving the fairness of the CNN model's outputs. Notably, the overall classification accuracy of the model improved slightly after mitigation, increasing from 58.28% to 58.35%, suggesting that removing bias can also enhance model performance. The causal modelling framework introduced in this study offers a principled approach for understanding and addressing the origins of bias in machine learning systems. This contribution extends the broader conversation on fairness in artificial intelligence, particularly in emotion recognition tasks where biases can have significant social implications. Future work could extend this approach by considering additional protected attributes, exploring more complex models, and applying the methodology to other domains where fairness is critical.

VI. FUTURE WORK

While the study offers valuable insights into detecting and mitigating algorithmic bias in CNN models for emotion classification, it identifies several areas for further improvement and exploration:

A. *Enhancing Ground Truth Labeling Accuracy*

The study used the DeepFace pre-trained model to label emotions in the dataset, which may not provide near-perfect accuracy, leading to mislabelled images (e.g., a happy face labelled as sad). These inaccuracies can affect the reliability of the modelling process. Future research should focus on improving emotion label accuracy by:

- **Implementing Image Preprocessing Techniques:** Applying data augmentation methods like rotation, scaling, and flipping to increase dataset diversity and improve labeling robustness.
- **Utilizing Advanced Emotion Detection Models:** Leveraging more sophisticated or fine-tuned models explicitly trained for high-accuracy emotion labeling to reduce mislabeling.

B. *Improving CNN Model Performance*

The custom CNN model achieved approximately 56% accuracy, demonstrating feasibility but potentially limiting the reliability of bias detection and mitigation strategies. Enhancing the model's accuracy through more advanced architectures or optimization techniques would strengthen the study's findings.

C. *Investigating Feature Extraction and Attribute Reduction Techniques*

Exploring alternative feature extraction methods and considering the impact of specific image attributes on algorithmic bias can help reduce bias while maintaining or improving accuracy. Potential strategies include:

- **High-Level Feature Extraction:** Using techniques like edge detection or facial landmark identification to extract features such as eyebrow positions and mouth shapes, reducing dependence on raw pixel data.
- **Attribute Reduction:** Removing certain attributes, such as converting images to grayscale to eliminate colour information that might inadvertently introduce bias.
- **Assessing Impact on Bias and Accuracy:** Conducting experiments to evaluate how these techniques affect model performance and fairness, identifying effective bias mitigation strategies.

By addressing these areas, future research can build upon the study's findings to develop more accurate and fair emotion classification models. Such advancements would contribute significantly to the ethical deployment of machine learning systems in applications where emotion recognition plays a critical role.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Professor Wendy Hui for her invaluable guidance, support, and encouragement throughout the course of this project. Her expertise and insights were instrumental in shaping the direction and outcome of this research. I am deeply appreciative of the time and effort she invested in mentoring me.

REFERENCES

[1] N. Mehrabi et al., “A survey on bias and fairness in machine learning,” ACM Computing Surveys, https://dl.acm.org/doi/10.1145/3457607 (accessed Dec. 1, 2024).

[2] E. Ferrara, “Fairness and bias in Artificial Intelligence: A brief survey of sources, impacts, and mitigation strategies,” arXiv.org, https://arxiv.org/abs/2304.07683 (accessed Dec. 1, 2024).

[3] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in ...,” Proceedings of Machine Learning Research, https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf (accessed Dec. 1, 2024).

[4] T. Hagendorff, “The ethics of AI Ethics: An evaluation of guidelines - minds and machines,” SpringerLink, https://link.springer.com/article/10.1007/s11023-020-09517-8 (accessed Dec. 1, 2024).

[5] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, A benchmark for interpretability methods in deep neural ..., https://proceedings.neurips.cc/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf (accessed Dec. 1, 2024).

[6] A. Krizhevsky, I. Sutskever, and G. E. Geoffrey, ImageNet Classification with Deep Convolutional Neural Networks, https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf (accessed Dec. 1, 2024).

[7] N. Joshi and P. Burlina, “Ai fairness via domain adaptation,” arXiv.org, https://arxiv.org/abs/2104.01109 (accessed Dec. 1, 2024).

[8] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” arXiv.org, https://arxiv.org/abs/1908.09635 (accessed Dec. 1, 2024).

[9] N. Joshi and P. Burlina, “Ai fairness via domain adaptation,” arXiv.org, https://arxiv.org/abs/2104.01109 (accessed Dec. 1, 2024).

[10] S. Sharma et al., “Data Augmentation for Discrimination Prevention and Bias Disambiguation,” ACM Digital Library, https://dl.acm.org/doi/10.1145/3375627.3375865 (accessed Dec. 1, 2024).

[11] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. In Proceedings of Workshop on Fairness, Accountability, and Transparency in Machine Learning, Halifax, Canada, August 2017. https://arxiv.org/pdf/1707.00075.pdf

[12] William Paul, Armin Hadzic, Neil Joshi, Fady Alajaji, and Philippe Burlina. 2022. TARA: Training and representation alteration for AI fairness and domain generalization. Neural Computation, 34 (3), 716-753. https://doi.org/10.1162/neco_a_01468

[13] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. Advances in neural information processing systems, 29, pp. 3315–3323.

[14] Preston Putzel and Scott Lee. 2022. Blackbox post-processing for multiclass fairness. Retrieved October 4, 2023, from https://arxiv.org/abs/2201.04461

[15] R. Rifkin and A. Klautau, “In defense of one-vs-all classification,” Journal of Machine Learning Research 5, https://www.jmlr.org/papers/volume5/rifkin04a/rifkin04a.pdf (accessed Dec. 1, 2024).

[16] D. Madras, E. Creager, T. Pitassi, and R. Zemel “Fairness through causal awareness: Learning causal latent-variable models for biased data,” Proceedings of the Conference on Fairness, Accountability, and Transparency, January 2019, pp. 349-358. Available: https://doi.org/10.1145/3287560.3287564

[17] A. Khademi, S. Lee, D. Foley, and V. Honavar. “Fairness in algorithmic decision making: An excursion through the lens of causality,” The World Wide Web Conference, May 2019, pp. 2907-2914. Available: https://doi.org/10.1145/3308558.3313559

[18] W. Hui and W. K. Lau, “Detecting and mitigating algorithmic bias in binary classification using causal modeling,” arXiv.org, https://arxiv.org/abs/2310.12421 (accessed Dec. 1, 2024).

[19] 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).

GitHub Link: https://github.com/Roy-Byun/IDC2808-Casual-Model-for-Bias-Detection-and-Mitigation-for-CNN