ACADEMIC CITY UNIVERSITY

FACULTY OF COMPUTATIONAL SCIENCES AND INFORMATICS

END OF FIRST SEMESTER - 2025/2026

AI4128 INTRODUCTION TO DEEP LEARNING

BERNICE AGYEIWAA AMPONSAH

10022200194

Project 2: FairVoice: Bias, Strength, and Clarity in Recognising Emotions in Speech

TABLE OF CONTENT

1. Introduction **3**
2. Dataset Description **4**
3. Preprocessing and Feature Extraction **6**
4. Dataset Splitting **8**
5. Baseline Model Architecture **10**
6. Baseline Performance and Bias Assessment **11**
6.1 Performance by Race 12
6.2 Performance by Sex 13
6.3 Summary of Bias Observations 14
7. Fairness Metrics **15**
8. Fairness Mitigation Strategies **17**
8.1 Oversampling 17
8.2 Reweighting 18
8.3 Adversarial Debiasing 18
9. Explainability Analysis **19**
9.1 SHAP Analysis 19
9.2 Grad-CAM Visualisation 20
10. Fairness Accuracy Trade-off Analysis **21**
11. Final Model Selection **22**
12. Accent Bias Analysis **24**
12.1 Model Confidence Across Accents 24
12.2 Uncertainty Analysis Using Entropy 25
12.3 Interpretation of Accent Bias Results 26
12.4 Implications for Speech Emotion Recognition Systems 26

12.5 Summary of Accent Bias Findings 27
13. Limitations **28**
14. Future Work **29**
16. Conclusion **30**

1. Introduction

In this project, I investigated fairness issues in a Speech Emotion Recognition (SER) system trained on the CREMA-D dataset. Speech Emotion Recognition aims to automatically identify human emotional states from speech signals and is widely used in applications such as virtual assistants, call-centre analytics, mental health monitoring, and human computer interaction. While these systems have shown promising performance, they often rely on datasets that are demographically imbalanced, which can lead to unequal performance across different groups of speakers.

My main goal was to examine whether a deep learning–based SER model performs equally well across demographic attributes such as sex, race, ethnicity, age, and accent. This question is important because speech carries not only emotional information but also speaker-specific characteristics related to identity. If a model unintentionally learns patterns linked to demographic traits rather than emotion itself, it may perform well for majority groups while disadvantageing minority groups.

Speech-based models are particularly vulnerable to bias because acoustic features such as pitch, speaking rate, accent, and vocal timbre vary naturally across speakers. When these variations are unevenly represented in the training data, the model may associate certain demographic traits with specific emotions, even when such associations are not meaningful. As a result, biased predictions can emerge, raising both technical concerns about model robustness and ethical concerns about fairness and inclusivity.

From a practical perspective, biased SER systems may produce unreliable outputs when deployed in real-world settings that involve diverse users. From an ethical perspective, such systems risk reinforcing existing inequalities by systematically underperforming for certain groups. For these reasons, fairness has become an increasingly important consideration in modern speech and machine learning systems.

To address these challenges, I designed and implemented a complete fairness-aware evaluation pipeline. I began by training a baseline convolutional neural network for emotion classification and analysing its performance across demographic groups. I then quantified bias using group-wise accuracy and disparity metrics. Based on the observed performance gaps, I applied several fairness mitigation strategies, including oversampling, reweighting, and adversarial debiasing, to reduce demographic disparities. In addition, I incorporated explainability techniques such as SHAP and Grad-CAM to better understand how the model makes decisions and whether it relies on different acoustic cues for different demographic groups.

Another important aspect of this project was analysing the trade-off between fairness and accuracy. Fairness improvements can sometimes come at the cost of reduced overall performance, so it is important to evaluate whether mitigation strategies produce a balanced and usable model. I also extended the analysis beyond the training distribution by evaluating the model on speech from different accents, which allowed me to study generalisation and accent-related bias.

By the end of this project, I aimed to answer three key questions:

1. Does the baseline SER model exhibit measurable demographic bias across different speaker groups?

2. Can fairness mitigation techniques reduce these biases without severely degrading accuracy?

3. What trade-offs emerge between fairness, accuracy, and model generalisation?

Overall, this work aims to demonstrate that fairness analysis is not an optional add-on but a necessary component of responsible speech processing system design. By systematically evaluating bias, applying

mitigation strategies, and interpreting model behaviour, this project provides a structured approach to building more transparent and equitable speech emotion recognition systems.

2. Dataset Description

I used the CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset) for this project. CREMA-D is a widely used dataset in speech emotion recognition research and contains acted emotional utterances recorded from multiple speakers. The dataset includes recordings across several emotion categories, with speakers delivering short sentences designed to elicit specific emotional states.

Although CREMA-D is a multimodal dataset containing audio, video, and textual information, I restricted this project to audio-only analysis. This choice was made to focus specifically on speech-based emotion recognition and to avoid introducing additional complexity from visual or linguistic cues. By using only audio, I was able to analyse how acoustic features alone contribute to emotion classification and how they interact with speaker demographics.

One of the main advantages of the CREMA-D dataset is that it provides speaker-level demographic metadata. Each speaker is associated with information about sex, age range, ethnicity, and race. This metadata makes CREMA-D particularly suitable for fairness analysis, as it allows performance to be evaluated across clearly defined demographic groups rather than treating all speakers as a homogeneous population.

To make the dataset easier to work with, I created a consolidated metadata file called metadata.csv. This file was constructed by extracting all relevant demographic attributes from the original dataset files and cleaning them into a single structured table. During this process, I ensured that each audio file was correctly mapped to its corresponding speaker and demographic information. This metadata file became a

central component of the project, as it was used throughout bias assessment, mitigation experiments, and explainability analysis.

After preparing the metadata, I analysed the distribution of samples across different demographic categories. This analysis revealed clear imbalances in representation, which are important to consider when evaluating fairness. The observed group counts were as follows:

1. Sex:

   Male (574 samples), Female (486 samples)

2. Race:

   Caucasian (732 samples), African American (246 samples), Asian (82 samples)

3. Ethnicity:

   Not Hispanic (896 samples), Hispanic (164 samples)

These distributions show that some demographic groups, particularly Caucasian speakers and non-Hispanic speakers, are much more heavily represented than others. In contrast, Asian speakers and Hispanic speakers form relatively small portions of the dataset. Such imbalances can influence how a model learns emotional patterns, as it may become more optimised for majority groups while underperforming for minority groups.

This imbalance is especially important in speech emotion recognition because emotional expression can vary subtly across speakers due to physiological, cultural, and linguistic factors. When certain groups are

underrepresented, the model may not be exposed to enough variation to generalise well to those speakers. As a result, performance gaps may emerge even if the overall accuracy appears acceptable.

It is also worth noting that CREMA-D does not include a region attribute. Because of this, region-based analysis was not possible and was excluded from the project. All fairness evaluations were therefore limited to sex, race, ethnicity, age, and accent (where accent data was introduced separately for generalisation analysis).

Overall, the CREMA-D dataset provided a strong foundation for this project due to its high-quality recordings, clear emotion labels, and availability of demographic metadata. At the same time, the observed demographic imbalances motivated the fairness analysis carried out in later sections of this work and highlighted the need for mitigation strategies to reduce biased model behaviour.

3. Preprocessing and Feature Extraction

Before training any models, I carried out a series of preprocessing steps to ensure that the audio data was consistent, clean, and suitable for feature extraction. Audio recordings in speech datasets often vary in amplitude, duration, and sampling rate, and these inconsistencies can negatively affect model training if they are not handled properly. For this reason, preprocessing was a critical step in my pipeline.

I first standardised all audio files by converting them to a fixed sampling rate. Using a consistent sampling rate ensures that time frequency representations are comparable across all samples. I also normalised the amplitude of each audio signal to reduce variations caused by recording conditions or speaker loudness. This step helps the model focus on emotion-related patterns rather than differences in recording volume.

After standardisation, I extracted acoustic features from each audio signal. Rather than feeding raw waveforms directly into the model, I chose to use time frequency representations that are commonly used in speech emotion recognition. These representations capture both spectral and temporal characteristics of speech, which are important for identifying emotional cues such as pitch variation, energy changes, and speaking intensity.

The two main feature types extracted were **log-mel spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs).**

Log-mel spectrograms were used as the primary input to the convolutional neural network. They provide a two dimensional representation of speech that closely aligns with human auditory perception by mapping frequencies onto the mel scale. Applying a logarithmic transformation further compresses the dynamic range of the signal, making subtle emotional variations more distinguishable. This representation is well suited for convolutional models, as it allows the network to learn local time–frequency patterns associated with different emotions.

MFCCs were also extracted as an additional feature representation. MFCCs summarise the spectral envelope of speech and are widely used in traditional and deep learning based speech processing systems. Although MFCCs are more compact than spectrograms, they still capture important information related to vocal tract characteristics and speech dynamics, which can be relevant for emotion recognition.

Once feature extraction was completed, I stored all features in a structured directory under data/features. To ensure the reliability of the pipeline, I performed several validation checks. I verified feature shapes to confirm consistency across samples, inspected randomly selected spectrograms and MFCCs visually, and checked dataset lengths to ensure that no files were missing or corrupted. These validation steps helped confirm that each audio file was correctly processed and aligned with its corresponding metadata.

By completing these preprocessing and feature extraction steps, I established a stable and reproducible foundation for model training. Ensuring consistency at this stage was particularly important for the fairness analysis later in the project, as preprocessing errors or inconsistencies could disproportionately affect certain demographic groups and lead to misleading conclusions. Overall, this stage ensured that any observed performance differences in later experiments could be attributed to model behaviour rather than data quality issues.

4. Dataset Splitting

To avoid speaker leakage and ensure a realistic evaluation, I split the dataset into training, validation, and test sets using actor-level partitioning. This means that all audio samples from a given speaker were assigned to only one split, and no speaker appeared in more than one subset.

Speaker leakage is a common issue in speech processing tasks. If audio from the same speaker appears in both the training and test sets, the model may learn speaker specific characteristics rather than general emotional patterns. This can lead to artificially high performance scores that do not reflect true generalisation. By using actor-level splitting, I ensured that the model was evaluated only on previously unseen speakers, making the results more reliable.

The training set was used to learn the model parameters, the validation set was used for hyperparameter tuning and model selection, and the test set was reserved strictly for final evaluation. I generated separate metadata files for each split (metadata_train.csv, metadata_val.csv, and metadata_test.csv) to clearly track which samples belonged to each subset. These files were used consistently throughout all experiments, including baseline training, fairness evaluation, and mitigation analysis.

This splitting strategy is particularly important for fairness analysis. When evaluating bias across demographic groups, it is essential that performance differences are not influenced by the model memorising specific speakers. Actor-level splitting ensures that any observed disparities in accuracy across sex, race, ethnicity, or age are more likely to reflect genuine differences in model generalisation rather than artefacts of the data split.

In addition, this approach allows for a more realistic simulation of real-world deployment, where the model is expected to operate on speech from speakers it has never encountered before. This is especially relevant for speech emotion recognition systems, which are often used in open-set environments with diverse users.

Overall, the use of actor-level partitioning strengthened the validity of both the performance and fairness results reported in this project. It ensured that the evaluation was robust, unbiased by speaker overlap, and suitable for analysing demographic effects in a meaningful way.

5. Baseline Model Architecture

To establish a reference point for all fairness evaluations, I trained a convolutional neural network (CNN) for speech emotion classification using log-mel spectrogram features as input. I chose a CNN architecture because convolutional models are well suited for time frequency representations of speech and have been shown to perform effectively in speech emotion recognition tasks.

The input to the model consisted of log-mel spectrograms extracted during the preprocessing stage. These spectrograms capture both temporal and spectral information, allowing the model to learn local patterns

such as energy bursts, pitch variations, and formant-related structures that are commonly associated with emotional expression. By using log-mel features rather than raw waveforms, I reduced the complexity of the learning task while retaining emotionally relevant acoustic information.

The CNN architecture was designed to progressively learn higher-level representations from the input spectrograms. Convolutional layers were used to extract local time frequency features, while pooling layers reduced dimensionality and helped the model become more robust to small variations in speech patterns. As the network depth increased, the model learned increasingly abstract representations that combined low-level acoustic cues into emotion-related patterns.

After feature extraction through the convolutional layers, the learned representations were passed to fully connected layers, which mapped the extracted features to emotion class probabilities. The final output layer used a softmax activation function to produce a probability distribution over the emotion classes. During training, the model optimised a categorical cross entropy loss function, which is commonly used for multi-class classification tasks.

I trained the baseline model using the training split and monitored performance on the validation set to avoid overfitting. Hyperparameters such as learning rate and number of training epochs were selected to ensure stable convergence. Throughout training, I observed that the model was able to learn meaningful patterns from the spectrogram inputs, as indicated by steadily improving validation performance.

After training was completed, I evaluated the model on the held-out test set. The baseline model produced several outputs that were essential for subsequent analysis:

- Test set predictions, which provided the predicted emotion label for each audio sample

- Confusion matrices, which allowed me to inspect class-wise performance and identify commonly confused emotions

- Accuracy and F1-scores, which served as the main performance metrics

These outputs were saved and later merged with demographic metadata to enable group-wise performance analysis. Importantly, this baseline model served as the reference point for all fairness experiments. By first understanding how the unmitigated model behaves across demographic groups, I was able to clearly identify performance gaps and measure how much each mitigation strategy improved or altered model behaviour.

Overall, the baseline CNN provided a strong foundation for the project. While it achieved reasonable overall performance, its predictions revealed demographic disparities that motivated the fairness assessment and mitigation work presented in later sections.

6. Baseline Performance and Bias Assessment

To assess demographic bias in the baseline Speech Emotion Recognition model, I combined the model's test predictions with the speaker demographic metadata. This allowed me to evaluate how the model performs across different groups instead of relying only on overall accuracy. Analysing group-wise performance is essential for fairness evaluation, as a model can appear to perform well overall while still underperforming for specific demographic groups.

I first merged the test set predictions with the metadata file using speaker and audio identifiers. Once this combined dataset was created, I computed accuracy and F1-score separately for each demographic group.

Accuracy was used to measure overall correctness, while F1-score was included to better reflect performance in the presence of class imbalance.

Before analysing performance, I examined the demographic distribution of the dataset. The CREMA-D dataset contains uneven group sizes, with some demographic categories being much more represented than others. This imbalance increases the risk that the model becomes biased toward majority groups during training.

6.1 Performance by Race

The table below shows the baseline model's performance across different racial groups.

Race	**Accuracy**	**F1-score**
Caucasian	0.419	0.398
African American	0.427	0.389
Asian	0.329	0.254

From this table, I observed clear performance differences across racial groups. The model achieved similar accuracy for Caucasian and African American speakers, but performance dropped significantly for Asian speakers. Both accuracy and F1-score were noticeably lower for the Asian group, indicating that the model struggled to generalise to this underrepresented group.

This gap is likely influenced by dataset imbalance, as Asian speakers form a much smaller portion of the dataset compared to Caucasian speakers. As a result, the model had fewer opportunities to learn emotion-related patterns from this group during training.

6.2 Performance by Sex

I also evaluated model performance across sex groups, as shown in the table below.

Sex	**Accuracy**	**F1-score**
Female	0.451	0.430
Male	0.383	0.341

The results show that the model performed better on female speakers than on male speakers. Although the difference is not extreme, it is consistent across both accuracy and F1-score. This suggests that the model may be learning sex related acoustic characteristics, such as pitch range or speaking style, that influence its predictions.

Even relatively small differences like this are important in fairness analysis, especially when such models are intended for real world use involving diverse users.

6.3 Summary of Bias Observations

Taken together, the race and sex performance tables clearly show that the baseline model does not perform uniformly across demographic groups. Minority groups, particularly Asian speakers, experienced lower performance, and measurable differences were also observed between male and female speakers.

These results confirm the presence of demographic bias in the baseline model. They provided strong motivation for applying fairness mitigation strategies, which are discussed in the next section. By quantifying these disparities at the baseline stage, I was able to later measure how effective each mitigation method was in reducing bias.

7. Fairness Metrics

To quantify demographic disparities more clearly, fairness metrics were computed based on group-wise accuracy. While group-specific accuracy values already indicate differences in performance, fairness metrics provide a more compact and comparable way to summarise these differences across multiple demographic attributes.

The primary metric used in this analysis was the disparity gap in accuracy. The disparity gap is defined as the difference between the highest and lowest accuracy values within a given demographic category. This metric highlights the extent to which a model favours certain groups over others and is commonly used in fairness evaluation because of its simplicity and interpretability.

Using the baseline model predictions, disparity gaps were computed for each demographic attribute considered in the project. The resulting values are shown in Table 7.1.

Table 7.1: Accuracy Disparity Gaps Across Demographic Attributes

Demographic Attribute	**Disparity Gap**
Sex	0.067
Race	0.098
Age	0.159
Ethnicity	0.073

As shown in the table, non-zero disparity gaps were observed for all demographic attributes. The largest disparity gap occurred for age, indicating that the difference between the best and worst performing age groups was substantial. This suggests that age-related vocal characteristics had a strong influence on model performance. The next largest gap was observed for race, followed by ethnicity and sex.

Although the disparity gap for sex was smaller compared to other attributes, it still indicates unequal performance between male and female speakers. Even relatively small gaps are important in fairness analysis, particularly for systems intended for deployment in real-world scenarios involving diverse users.

These fairness metrics provide a clear baseline reference for evaluating the effectiveness of mitigation strategies. By computing the same metrics after applying oversampling, reweighting, and adversarial debiasing, it becomes possible to directly measure how much each method reduces demographic disparities and to assess the trade off between fairness and overall accuracy.

Overall, the disparity gap analysis confirms that the baseline model exhibits measurable demographic bias and reinforces the need for mitigation techniques to achieve more equitable performance across speaker groups.

8. Fairness Mitigation Strategies

The fairness metrics presented in the previous section clearly show that the baseline model exhibits measurable demographic disparities across multiple attributes. While these metrics are useful for identifying bias, they do not address the underlying causes. To reduce these disparities, several fairness mitigation strategies were applied during model training and evaluation.

Three complementary approaches were explored in this work: oversampling, reweighting, and adversarial debiasing. Each method addresses bias from a different perspective and introduces different trade offs between fairness, stability, and overall performance. Evaluating multiple strategies allows for a more comprehensive understanding of how fairness can be improved in speech emotion recognition systems.

8.1 Oversampling

Oversampling was used as a data level mitigation strategy to address demographic imbalance in the training set. In this approach, samples from underrepresented demographic groups were duplicated during training so that all groups contributed more equally to the learning process.

By increasing the frequency with which minority group samples appeared during training, the model was encouraged to learn emotion-related patterns from these groups more effectively. This approach led to improved accuracy for some minority groups, particularly in race and sex categories.

However, oversampling also introduced increased variance in performance, especially across age groups. Because duplicated samples do not introduce new information, the model can become sensitive to repeated patterns, which may reduce generalisation stability. This highlights a key limitation of oversampling when applied in isolation.

8.2 Reweighting

Reweighting was applied as a loss level mitigation strategy. Instead of duplicating samples, this method adjusted the loss function so that errors on underrepresented demographic groups were penalised more heavily during training.

This approach allowed the model to place greater emphasis on minority groups without altering the dataset distribution. Compared to oversampling, reweighting produced more stable performance across demographic groups while still reducing fairness gaps. Overall accuracy was largely preserved, making this method particularly attractive when maintaining predictive performance is a priority.

8.3 Adversarial Debiasing

Adversarial debiasing was applied as a representation-level mitigation strategy. In this approach, an auxiliary network branch was trained to predict demographic attributes from the learned feature

representations, while a gradient reversal layer was used to discourage the emotion classifier from encoding demographic information.

By explicitly penalising demographic predictability, this method aimed to force the model to focus on emotion-related acoustic cues that generalise across speaker groups. Among the mitigation strategies tested, adversarial debiasing produced the most consistent reduction in disparity gaps across sex, race, and age. Although a small decrease in overall accuracy was observed, the resulting model achieved a significantly more balanced performance profile across demographic groups.

9. Explainability Analysis

While fairness metrics and mitigation strategies quantify and reduce demographic disparities, they do not directly explain how the model arrives at its decisions. To gain insight into the internal behaviour of the model, explainability techniques were applied. These methods help identify which regions of the input spectrogram contribute most strongly to emotion predictions and whether these regions differ across demographic groups.

Two complementary explainability techniques were used in this analysis: SHAP and Grad-CAM. Together, they provide both feature-level and spatial interpretations of the model's behaviour.

9.1 SHAP Analysis

SHAP (SHapley Additive exPlanations) values were computed for a selected subset of test samples to estimate the contribution of individual time frequency regions to the model's predictions. SHAP provides a theoretically grounded approach to attributing model outputs to input features and is particularly useful for identifying positive and negative contributions.

The SHAP analysis revealed that the model primarily relied on regions associated with pitch variation, energy changes, and temporal dynamics features commonly linked to emotional expression in speech. These regions appeared consistently across different emotion classes, suggesting that the model was learning meaningful acoustic cues rather than focusing on irrelevant noise.

When SHAP visualisations were compared across demographic groups, subtle differences were observed in the distribution and intensity of contributions. Although the same general regions were often highlighted, the relative importance of specific frequency bands varied. These variations suggest that demographic characteristics may influence how emotional cues are represented in the learned feature space, which can contribute to performance disparities.

9.2 Grad-CAM Visualisation

Grad-CAM was used to visualise which regions of the log-mel spectrograms activated the model during emotion classification. This method highlights areas of the input that have the greatest influence on the final prediction, making it well suited for convolutional architectures.

Across multiple test samples, Grad-CAM consistently highlighted mid frequency regions between approximately 500 Hz and 1500 Hz, particularly for high arousal emotions such as anger and happiness. These regions correspond to formant related structures and pitch dynamics that are strongly associated with emotional speech.

Comparisons across demographic groups showed largely similar attention patterns, indicating that the model relied on broadly consistent acoustic cues. However, differences in activation strength were observed, which may explain why certain groups experienced lower classification accuracy. These findings support the fairness analysis by showing that demographic effects are reflected not only in performance metrics but also in learned representations.

10. Fairness Accuracy Trade-off Analysis

After applying the mitigation strategies, a comparative analysis was conducted to examine the trade-off between fairness and overall accuracy. Improving fairness often involves constraining the model, which can reduce raw predictive performance. Understanding this trade-off is essential for selecting an appropriate model for deployment.

The following models were compared:

- Baseline CNN
- Oversampling model
- Reweighting model
- Adversarially debiased model

Each model was evaluated using overall accuracy alongside disparity gap metrics. The results showed that oversampling and reweighting reduced fairness gaps to a moderate extent, while adversarial debiasing achieved the most substantial reductions across demographic attributes.

Although the adversarial model exhibited a small decrease in overall accuracy compared to the baseline, it provided the most balanced performance across groups. Reweighting preserved accuracy more effectively but did not reduce disparities as consistently. These results highlight the inherent trade-offs involved in fairness-aware model design.

11. Final Model Selection

The final model was selected based on a comprehensive evaluation of both fairness metrics and overall classification performance across all trained models. Rather than relying solely on raw accuracy, the selection process prioritised balanced performance across demographic groups, as unequal performance can undermine the reliability and ethical acceptability of speech emotion recognition systems.

The fairness accuracy trade-off analysis compared four models: the baseline CNN, the oversampling model, the reweighted model, and the adversarially debiased model. While the baseline model achieved reasonable overall accuracy, it exhibited clear demographic disparities across race, sex, and age groups. These disparities made it unsuitable for deployment in scenarios involving diverse users.

The oversampling model improved performance for some underrepresented groups by increasing their representation during training. However, this improvement was accompanied by increased variance and reduced stability, particularly across age groups. In some cases, oversampling amplified noise rather than improving generalisation, which limited its effectiveness as a standalone solution.

The reweighted model provided a more stable alternative by adjusting the loss function to penalise errors on underrepresented groups more strongly. This approach preserved overall accuracy more effectively than oversampling and reduced fairness gaps to a moderate extent. However, disparities across certain demographic attributes remained noticeable, indicating that reweighting alone was insufficient to fully address bias in the learned representations.

In contrast, the adversarially debiased model consistently achieved the lowest disparity gaps across multiple demographic attributes. By explicitly discouraging the model from encoding demographic information in its internal representations, adversarial debiasing forced the classifier to focus more strongly on emotion-relevant acoustic cues that generalise across speaker groups. Although this approach

resulted in a small reduction in overall accuracy compared to the baseline and reweighted models, the decrease was relatively minor when considered alongside the substantial improvement in fairness.

The selection of the adversarially debiased model reflects a deliberate design decision to prioritise robustness, inclusivity, and equitable performance over maximising raw accuracy. In real-world applications such as mental health monitoring, call-centre analytics, and human computer interaction, biased predictions can have significant practical and ethical consequences. A model that performs consistently across demographic groups is therefore more suitable for deployment than one that achieves slightly higher accuracy for majority populations at the expense of minority groups.

Overall, the adversarially debiased model provides the most balanced trade-off between fairness and performance in this study. Its selection demonstrates that incorporating fairness constraints into model training can meaningfully reduce demographic disparities while still maintaining acceptable predictive capability. This model was therefore chosen as the final model for further analysis and as the most appropriate candidate for fairness-aware speech emotion recognition.

12. Accent Bias Analysis

In addition to evaluating fairness across demographic attributes present in the CREMA-D dataset, an accent bias analysis was conducted to assess how well the trained speech emotion recognition model generalises to speech patterns outside its primary training distribution. Accent variation represents a particularly important challenge in speech processing, as accents introduce systematic differences in pronunciation, prosody, intonation, and rhythm that can influence acoustic feature representations.

The CREMA-D dataset consists predominantly of American English speakers, which raises concerns about the model's ability to generalise to other accents. If a model is trained primarily on one accent, it

may incorrectly associate accent-specific acoustic patterns with emotional categories, leading to biased or unreliable predictions when exposed to unfamiliar speech varieties.

To investigate this issue, the model was evaluated on speech samples from American, British, and Indian accents. These accents were selected to represent both closely related (British) and more acoustically distinct (Indian) varieties of English. The analysis focused on two complementary measures: prediction confidence and prediction uncertainty, measured using entropy.

12.1 Model Confidence Across Accents

Prediction confidence reflects how strongly the model favours its predicted emotion class. High confidence values indicate that the model is certain about its decision, while lower values suggest ambiguity. Confidence scores were computed for each accent group and compared statistically using the Kruskal-Wallis test, which does not assume normality.

The results showed no statistically significant difference in confidence scores across American, British, and Indian accents. On the surface, this suggests that the model appears equally confident when processing speech from all three accent groups.

However, confidence alone can be misleading. A model may produce confident predictions even when it is poorly calibrated or relying on inappropriate cues. For this reason, confidence scores were analysed alongside uncertainty metrics to obtain a more complete picture of accent-related bias.

12.2 Uncertainty Analysis Using Entropy

To complement the confidence analysis, prediction uncertainty was measured using entropy. Entropy captures how spread out the predicted class probabilities are, with higher entropy indicating greater uncertainty in the model's decision.

Unlike the confidence analysis, the entropy results revealed a clear and statistically significant pattern. Speech samples with an Indian accent exhibited substantially higher entropy values compared to American and British accents. A Kruskal-Wallis test confirmed that these differences were statistically significant ($p < 0.001$).

This finding indicates that, although the model may appear confident across accents, it is internally less certain when processing Indian accented speech. In contrast, British and American accents showed consistently low entropy values, suggesting that the model's predictions were both confident and stable for these groups.

12.3 Interpretation of Accent Bias Results

The discrepancy between confidence and uncertainty highlights an important limitation of relying solely on confidence-based metrics. While the model produces high-confidence predictions for all accents, the elevated uncertainty for Indian-accented speech suggests weaker internal representations and reduced robustness.

This behaviour is likely a consequence of training data imbalance. Because the CREMA-D dataset contains primarily American accented speech, the model has limited exposure to phonetic and prosodic patterns associated with other accents. As a result, it struggles to generalise emotional cues when these cues are expressed through unfamiliar accent-specific variations.

From a fairness perspective, this represents a form of accent bias, where the model performs less reliably for speakers whose accent deviates from the dominant training distribution. In real-world applications, such bias can lead to systematic misinterpretation of emotions for certain user groups, particularly in multicultural or global deployment settings.

12.4 Implications for Speech Emotion Recognition Systems

The accent bias analysis demonstrates that fairness issues in speech emotion recognition extend beyond traditional demographic categories such as sex and race. Accent-related variability introduces additional challenges that are often overlooked but are critical for building inclusive and globally applicable systems.

These findings suggest that improving accent robustness requires:

- More accent-diverse training data
- Explicit domain generalisation techniques
- Accent-aware evaluation during model development

Without addressing accent bias, even fairness-mitigated models may fail to generalise reliably to real-world speech data.

12.5 Summary of Accent Bias Findings

In summary, the accent bias analysis revealed that:

- Model confidence alone does not capture generalisation quality
- Indian accented speech exhibited significantly higher uncertainty
- Accent imbalance in training data negatively affects robustness

- Accent bias remains an important fairness concern in SER systems

These results reinforce the importance of evaluating speech models beyond the conditions seen during training and highlight accent diversity as a key factor in responsible speech system design.

13. Limitations

Although this study provides a comprehensive fairness-aware evaluation of a speech emotion recognition system, several limitations should be acknowledged. Recognising these limitations is important for correctly interpreting the results and for guiding future improvements.

One key limitation relates to dataset imbalance. While CREMA-D provides valuable demographic metadata, several demographic groups particularly certain race, ethnicity, and accent categories are underrepresented. Small sample sizes can make group-wise performance metrics more sensitive to noise and may exaggerate or obscure true performance differences. As a result, some observed disparities may partially reflect data scarcity rather than purely model behaviour.

Another limitation concerns the use of acted emotional speech. CREMA-D consists of scripted, acted utterances designed to elicit clear emotional expression. While this is useful for controlled experiments, it may not fully capture the variability and subtlety of natural, spontaneous emotional speech. Emotional expression in real-world settings is often influenced by context, background noise, and conversational dynamics, which were not present in this dataset.

The computational cost of explainability methods also imposed constraints on the analysis. SHAP computations were resource-intensive, limiting the number of samples that could be examined in detail. As a result, explainability findings are based on representative subsets rather than exhaustive analysis

across all demographic groups. While these subsets provide valuable insights, they may not capture all variations in model behaviour.

In addition, Grad-CAM interpretations are architecture-dependent and provide only a partial view of model decision-making. Grad-CAM highlights regions of high activation but does not fully explain how these regions interact to produce final predictions. Consequently, explainability results should be interpreted as supportive evidence rather than definitive explanations.

Another limitation lies in the scope of fairness metrics used. This study focused primarily on accuracy-based disparity gaps. While these metrics are intuitive and informative, fairness is a multidimensional concept. Other notions of fairness, such as calibration fairness or error-rate parity, were not explored in depth and could provide additional perspectives on model behaviour.

Finally, the accent bias analysis was conducted using a limited set of accents. Although clear patterns were observed, the findings cannot be assumed to generalise to all global accents. A broader range of accents would be required to draw stronger conclusions about accent-related fairness.

14. Future Work

The limitations identified in this study highlight several promising directions for future research and system improvement.

A natural extension of this work would involve training and evaluating models on more demographically and accent-diverse datasets. Incorporating a wider range of speakers would reduce data imbalance and improve generalisation, particularly for underrepresented groups. Collecting or augmenting datasets with diverse accents, age groups, and cultural backgrounds would be especially beneficial.

Future work could also explore advanced fairness-aware training techniques beyond those evaluated here. Methods such as fairness-constrained optimisation, causal fairness approaches, or representation disentanglement could further reduce demographic leakage while preserving emotional information.

Another important direction involves improving explainability at scale. More efficient explainability methods or approximation techniques could enable broader analysis across entire datasets, allowing for stronger conclusions about demographic effects on learned representations.

In addition, future studies could examine calibration-based fairness metrics, ensuring that model confidence aligns equally well with true performance across demographic groups. This is particularly relevant given the observed mismatch between confidence and uncertainty in the accent bias analysis.

Extending evaluation to real-world, spontaneous speech is another critical step. Testing models on conversational datasets with background noise and varied speaking styles would provide a more realistic assessment of fairness and robustness in deployment scenarios.

Finally, future work could investigate joint mitigation of demographic and accent bias, rather than treating them as separate issues. Developing models that are simultaneously robust to demographic variation and accent diversity would represent a significant advancement toward inclusive and trustworthy speech emotion recognition systems.

16. Conclusion

This study presented a comprehensive fairness-aware evaluation of a Speech Emotion Recognition system trained on the CREMA-D dataset. The analysis went beyond overall classification performance to examine how the model behaves across different demographic attributes, including sex, race, age,

ethnicity, and accent. The findings demonstrate that, although the baseline model achieved reasonable overall accuracy, it exhibited clear and measurable demographic disparities.

Through systematic bias assessment, performance gaps were identified across multiple demographic groups, with particularly pronounced differences observed for race, age, and accent. These results highlight that emotion recognition models can unintentionally encode speaker-specific characteristics, leading to unequal performance even when trained on widely used benchmark datasets.

To address these disparities, several fairness mitigation strategies were explored. Oversampling and reweighting provided partial improvements by addressing data imbalance and loss sensitivity, while adversarial debiasing achieved the most consistent reduction in demographic disparities. The adversarial approach proved effective in discouraging the model from encoding demographic information, resulting in a more balanced performance profile across speaker groups. Although this method introduced a small reduction in overall accuracy, the trade-off was justified by the substantial gains in fairness and robustness.

Explainability techniques, including SHAP and Grad-CAM, played an important role in interpreting model behaviour. These methods revealed that the model primarily relies on meaningful acoustic cues such as pitch variation and energy dynamics, while also showing that the strength and distribution of these cues can vary across demographic groups. This insight supports the fairness findings and demonstrates how demographic effects are reflected in learned representations rather than being purely statistical artefacts.

The accent bias analysis further emphasised the importance of evaluating generalisation beyond the training distribution. While confidence scores appeared similar across accents, uncertainty analysis revealed significantly higher uncertainty for Indian-accented speech, indicating reduced robustness for accents that were underrepresented during training. This result underscores that fairness challenges in

speech emotion recognition extend beyond traditional demographic categories and must also account for linguistic and cultural variation.

Overall, the results of this work demonstrate that fairness analysis is a critical component of responsible speech processing system design. Addressing bias not only improves ethical outcomes but also enhances model reliability and generalisation in diverse real-world settings. By combining demographic evaluation, mitigation strategies, and explainability analysis, this study provides a structured framework for developing more transparent, equitable, and trustworthy speech emotion recognition systems.