ACADEMIC CITY UNIVERSITY

FACULTY OF COMPUTATIONAL SCIENCES AND INFORMATICS

END OF FIRST SEMESTER - 2025/2026

AI4129 SPEECH PROCESSING

BERNICE AGYEIWAA AMPONSAH

10022200194

Project 2: FairVoice: Bias and Explainability in Speech Emotion Recognition

TABLE OF CONTENT

Introduction

This report presents a comprehensive account of the work carried out for the FairVoice project, which focuses on bias assessment, fairness mitigation, explainability, and model evaluation in Speech Emotion Recognition (SER). The primary objective of the project was to investigate whether an SER model trained on the CREMA-D dataset performs equitably across different demographic groups, including race, sex, and ethnicity, and to understand the factors that contribute to any observed disparities.

Speech Emotion Recognition systems are increasingly integrated into real-world applications such as mental health monitoring, educational technologies, call centre analytics, and affect-aware human–computer interaction systems. While recent advances in deep learning have significantly improved SER performance on benchmark datasets, these gains often mask underlying demographic performance gaps. Models trained on imbalanced datasets may learn correlations between emotional labels and speaker-specific characteristics, resulting in unequal performance across demographic groups.

The motivation for this project is both practical and ethical. From a practical perspective, uneven model performance can limit the reliability of SER systems when deployed in diverse populations. From an ethical standpoint, biased emotion recognition systems risk reinforcing existing social inequalities and undermining trust in AI technologies. These concerns are particularly relevant in low- and middle-income regions (LMICs), where speech technologies must be inclusive, transparent, and robust to demographic diversity.

To address these challenges, this project adopts a structured fairness investigation pipeline. The work begins with baseline model development and evaluation, followed by systematic bias assessment across demographic subgroups. Multiple mitigation strategies are then applied to reduce observed disparities, and explainable AI techniques are used to interpret model behaviour and decision-making processes. In addition, fairness–accuracy trade-offs and accent-based generalisation are analysed to provide a holistic evaluation of model performance.

This report documents all stages of the project in detail, including preprocessing, dataset management, baseline modelling, bias measurement, mitigation experiments, explainability analysis using SHAP and Grad-CAM, and final model selection. All results, figures, and artefacts referenced in this document are reproducible and available within the project repository.

2. Dataset Overview and Preprocessing

The dataset used in this project is CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset), a widely adopted benchmark dataset for speech emotion recognition research. CREMA-D contains acted emotional utterances produced by a diverse set of speakers under controlled recording conditions. For the purposes of this project, only the audio modality was utilised, as the focus was on analysing fairness and bias in audio-based speech emotion recognition models.

CREMA-D includes emotional speech samples across several discrete emotion categories. In addition to the audio recordings, the dataset provides speaker-level demographic metadata, including sex, age range, race, and ethnicity. This demographic information makes CREMA-D particularly suitable for fairness-focused analysis, as it enables performance evaluation across multiple demographic subgroups. However, the distribution of samples across these groups is not uniform, which introduces realistic conditions for studying bias arising from dataset imbalance.

As part of the preprocessing pipeline, all audio files were first validated to ensure that file paths, formats, and metadata references were consistent. The audio files were organised into a structured directory layout to support reproducibility and ease of access throughout the project. Each file was loaded programmatically, and checks were performed to confirm that no corrupted or missing audio samples were present.

To ensure consistency across the dataset, all audio signals were resampled to a standard sampling rate and normalised. This step was necessary to reduce variability introduced by differences in recording conditions and amplitude ranges. Normalisation ensured that the extracted features reflected meaningful acoustic patterns rather than artefacts caused by signal scaling.

From the preprocessed audio signals, acoustic features were extracted to serve as inputs to the emotion recognition model. Two primary feature representations were used: log-mel spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs). Log-mel spectrograms capture time–frequency energy distributions that are well suited for convolutional neural networks, while MFCCs provide a compact representation of perceptually relevant spectral characteristics. These features are widely used in SER research and are known to encode information related to pitch, energy, and spectral shape, which are important cues for emotion recognition.

All extracted features were saved to disk under a dedicated data/features directory. Storing features in advance allowed experiments to be run efficiently and ensured that all models were trained and evaluated using the exact same feature representations. Each feature file was linked to its corresponding metadata entry to facilitate demographic analysis during later stages of the project.

Additional validation steps were carried out to confirm the integrity of the preprocessing pipeline. Feature shapes and dimensions were inspected to ensure consistency across samples. Random samples were visualised to verify that the extracted spectrograms and MFCCs were meaningful and free from obvious artefacts. Dataset sizes were also cross-checked against metadata records to ensure that no samples were lost during preprocessing.

By the end of this stage, a clean, well-structured, and fully validated dataset was prepared. This preprocessing foundation was essential for all subsequent experiments, including baseline model training, demographic bias assessment, fairness mitigation, and explainability analysis.

3. Metadata Preparation and Group Structures

Demographic metadata plays a central role in fairness analysis, as it enables the evaluation of model performance across different population groups. For this project, speaker-level demographic information provided by the CREMA-D dataset was carefully extracted, cleaned, and consolidated to support systematic bias assessment.

All available demographic fields were combined into a single structured file, metadata.csv, which served as the primary reference table throughout the project. This metadata table included speaker identifiers, file identifiers, emotion labels, and demographic attributes such as sex, age range, race, and ethnicity. Consolidating this information into one table ensured consistency across experiments and allowed demographic attributes to be easily joined with model predictions during evaluation.

During metadata preparation, several cleaning steps were performed to ensure reliability. Demographic labels were standardised to remove inconsistencies caused by formatting differences or missing values. Speaker identifiers were verified to ensure that each audio sample was correctly associated with the appropriate demographic attributes. Any metadata fields that were incomplete or irrelevant to the fairness objectives of the project were carefully reviewed.

After cleaning, the metadata was loaded into a dataframe (referred to as meta_df), which became the central data structure for demographic analysis. Using this dataframe, the distribution of samples across demographic groups was examined to identify potential imbalances. Understanding these distributions was critical for interpreting fairness metrics and motivating the choice of mitigation strategies.

The following group counts were computed directly from meta_df:

- Ethnicity:

Not Hispanic – 896 samples

Hispanic – 164 samples

- Race:

  Caucasian – 732 samples

  African American – 246 samples

  Asian – 82 samples

- Sex:

  Male – 574 samples

  Female – 486 samples

These counts clearly demonstrate that the dataset is not evenly balanced across demographic categories. In particular, certain groups, such as Asian speakers and Hispanic speakers, are significantly underrepresented. Such imbalance creates conditions under which a model may achieve higher performance for majority groups while underperforming for minority groups, thereby introducing demographic bias.

The availability and quality of demographic attributes were also carefully reviewed. It was confirmed that the dataset did not include a reliable region attribute, and as a result, region-based analysis was excluded from the project. Only demographic categories with sufficient representation and consistent annotation namely sex, race, ethnicity, and age were retained for quantitative bias evaluation. This decision ensured that fairness metrics were meaningful and statistically interpretable.

By establishing clear demographic group structures at this stage, the project laid the groundwork for all subsequent fairness analyses. The prepared metadata enabled accurate subgroup performance evaluation, computation of disparity metrics, and targeted application of mitigation strategies in later stages of the project.

4. Dataset Splitting

A careful dataset splitting strategy is essential for reliable model evaluation, particularly in fairness-focused studies. In this project, the CREMA-D dataset was divided into training,

validation, and test sets using an actor-level (speaker-independent) partitioning approach. This strategy ensures that no speech samples from the same speaker appear in more than one split.

Speaker-independent splitting is especially important in speech emotion recognition tasks, as models can easily memorise speaker-specific characteristics such as pitch range, speaking style, or vocal timbre. If the same speaker appears in both training and test sets, performance metrics may be artificially inflated and fail to reflect true generalisation. By enforcing actor-level separation, the evaluation more accurately measures the model's ability to generalise to unseen speakers.

To implement this strategy, speakers were first grouped by unique speaker identifiers. These speakers were then assigned to training, validation, and test splits in a way that preserved overall dataset balance while maintaining strict speaker separation. The resulting splits were stored as separate metadata files: metadata_train.csv, metadata_val.csv, and metadata_test.csv. These files were used consistently across all experiments to ensure reproducibility.

The training split was used to optimise model parameters, while the validation split was employed for model selection and hyperparameter tuning. The test split was held out entirely until final evaluation and was used for all reported performance metrics, including demographic bias assessment and fairness analysis.

This splitting strategy also strengthens the validity of the fairness evaluation. Since each demographic group contains multiple speakers, speaker-independent splits help ensure that observed performance differences across groups are due to model behaviour rather than memorisation of specific voices. As a result, fairness metrics computed on the test set more accurately reflect genuine demographic disparities in model performance.

Overall, the dataset splitting procedure established a robust and reliable evaluation framework. By preventing data leakage and enforcing speaker independence, this step ensured that all subsequent results including bias assessment, mitigation experiments, and explainability analysis were grounded in fair and realistic evaluation conditions.

5. Baseline Model Development

To establish a reference point for fairness and bias evaluation, a baseline speech emotion recognition model was developed and trained using the preprocessed CREMA-D dataset. This baseline model served as the foundation for all subsequent analyses, including demographic bias assessment, mitigation experiments, and explainability studies.

The baseline architecture was a convolutional neural network (CNN) designed to operate on time–frequency representations of speech. CNNs are well suited for spectrogram-based inputs, as they can effectively capture local temporal and spectral patterns associated with emotional expression. In this project, the CNN was trained using log-mel spectrogram features, which provide a rich representation of energy distribution across time and frequency.

The model was trained using the training split defined earlier, with performance monitored on the validation split to guide model selection. Standard training procedures were followed, including batch-based optimisation and regular evaluation to prevent overfitting. The best-performing model checkpoint, as determined by validation performance, was saved for subsequent evaluation.

After training, the baseline model was evaluated on the held-out test set. Predictions for each test sample were generated and stored in predictions_test.csv. In addition to raw predictions, several evaluation artefacts were produced, including confusion matrices, classification reports, and summary performance metrics. These artefacts were saved under the evaluation_results directory and served as the primary inputs for bias and fairness analysis.

The baseline model achieved reasonable overall performance on the test set, demonstrating its ability to learn emotion-related acoustic patterns from the CREMA-D dataset. However, aggregate performance metrics alone do not reveal how performance varies across demographic groups. As a result, the baseline model was not considered the final solution, but rather a starting point for deeper investigation.

By establishing a well-trained baseline model and generating comprehensive evaluation outputs, this stage provided a solid foundation for identifying demographic performance gaps and assessing the impact of fairness-aware interventions in subsequent stages of the project.

6. Bias Assessment

In this section, I evaluated how the baseline Speech Emotion Recognition model performs across different demographic groups in the CREMA-D dataset. The goal of this analysis was to determine whether the model's accuracy and error patterns were distributed fairly across attributes such as race, sex, ethnicity, and age, or whether certain groups were systematically disadvantaged.

I began the bias assessment by examining the demographic distribution of the dataset. Using the metadata prepared earlier, I confirmed that the number of samples was not evenly distributed

across demographic groups. For example, the Caucasian group contained significantly more samples than the Asian group, and similar imbalances were observed across other attributes. This imbalance raised concerns that the model might learn patterns that favour majority groups.

To evaluate model performance across demographics, I joined the test set predictions with the corresponding demographic metadata. This allowed me to compute accuracy and F1-score separately for each demographic subgroup. By breaking down performance in this way, I was able to move beyond aggregate metrics and identify where the model performed well and where it struggled.

I quantified demographic disparities using fairness gap metrics, which measure the difference between the highest and lowest performing subgroups within a demographic category. These metrics provided a concise way to summarise the extent of performance inequality across groups. The computed disparity gaps confirmed that performance was not evenly distributed.

For example, when analysing performance by race, I observed that Caucasian speakers achieved higher accuracy and F1-scores compared to Asian speakers, who recorded the lowest performance. Similar trends were observed in the analysis by sex, where performance differed noticeably between male and female speakers. Age-based analysis also revealed variation in performance across age groups, with some age categories consistently outperforming others.

To visualise these differences, I generated figures showing subgroup accuracy and F1-scores. These visualisations made it easier to compare groups and clearly illustrated where the largest performance gaps occurred. The results consistently showed that minority demographic groups experienced lower model performance.

Based on these findings, I concluded that the baseline SER model exhibits demographic bias, driven in part by dataset imbalance and in part by the model's sensitivity to speaker-specific characteristics that correlate with demographic attributes. These observations motivated the next phase of the project, in which I applied fairness mitigation strategies to reduce the observed performance disparities.

8. Explainability

In addition to evaluating model performance and fairness metrics, I focused on understanding how the speech emotion recognition model makes its decisions. Explainability is a critical component of fairness analysis because biased behaviour often emerges from the model relying

on unintended or demographic-dependent cues rather than emotion-relevant information. To investigate this, I applied explainable AI (XAI) techniques to interpret the model's predictions.

I used SHAP and Grad-CAM as complementary explainability methods. These techniques allowed me to visualise and analyse which regions of the acoustic representations contributed most strongly to the model's decisions.

8.1 SHAP Analysis

I applied SHAP to analyse the contribution of different time–frequency regions of the input features to the model's emotion predictions. SHAP values provide an estimate of how much each part of the input influences the final output, allowing me to identify regions that contribute positively or negatively to a predicted emotion.

To perform this analysis, I selected a subset of test samples and generated SHAP values for each. These values were then aligned with the corresponding demographic metadata, which allowed me to inspect whether the model relied on different acoustic patterns for different demographic groups.

By examining SHAP heatmaps, I observed which frequency bands and temporal regions were most influential in the model's predictions. In several cases, the model focused on mid-frequency regions that are commonly associated with pitch variation and energy changes, which are known to correlate with emotional expression. However, I also observed cases where the model appeared to rely on patterns that could be linked to speaker-specific characteristics rather than purely emotional cues.

Comparing SHAP explanations across demographic groups helped me assess whether the model consistently attended to similar acoustic regions for different speakers. This analysis provided insight into whether demographic bias might arise from the model emphasising different features depending on who was speaking.

8.2 Grad-CAM Analysis

In addition to SHAP, I applied Grad-CAM to generate visual explanations for the convolutional neural network. Grad-CAM highlights the regions of the log-mel spectrogram that most strongly influence the model's predictions by analysing gradients flowing through the final convolutional layers.

For each selected test sample, I generated three visual outputs: the original log-mel spectrogram, the Grad-CAM activation map, and an overlay combining both. These visualisations made it easier to interpret which time–frequency regions the model relied on when predicting specific emotions.

In many cases, the Grad-CAM maps showed strong activation in mid-frequency bands, particularly between approximately 500 Hz and 1500 Hz. These regions often correspond to pitch contours and energy bursts associated with high-arousal emotions. I also compared Grad-CAM outputs across different demographic groups to check whether the model focused on consistent acoustic cues.

Overall, the explainability analysis provided valuable insight into the internal behaviour of the model. SHAP and Grad-CAM helped me determine whether the model relied on meaningful emotional cues or whether its decisions were influenced by demographic-related artefacts. These findings supported the fairness analysis and informed the interpretation of mitigation results in later sections.

9. Accuracy and Fairness Trade-off

After applying the different bias mitigation strategies, I analysed how each approach affected both model accuracy and fairness metrics. The purpose of this analysis was to understand whether improvements in fairness came at the cost of reduced predictive performance, or whether certain strategies achieved a more balanced outcome.

I compared the following models:

- Baseline CNN model
- Oversampling model
- Reweighting model
- Adversarial debiasing model

For each model, I measured overall accuracy alongside fairness metrics such as demographic disparity gaps, demographic parity difference, and equalized odds difference. These metrics allowed me to quantify not only how well the model performed overall, but also how evenly that performance was distributed across demographic groups.

To visualise the relationship between fairness and accuracy, I generated a fairness–accuracy trade-off plot (shown in the trade-off scatter figure in the report). Each point in the plot represents a trained model, with accuracy on one axis and a fairness metric on the other. This visualisation made it easier to compare models and identify those that offered favourable trade-offs.

From this analysis, I observed that the baseline model achieved relatively high accuracy but exhibited large demographic performance gaps. The oversampling model improved performance for certain minority groups, but in some cases introduced instability and variability across demographic categories. The reweighted model produced more consistent results, reducing fairness gaps while maintaining reasonable accuracy.

The adversarial debiasing model demonstrated the most balanced trade-off between fairness and accuracy. Although its overall accuracy was slightly lower than that of the baseline model, it significantly reduced demographic disparities across sex and race groups. This indicates that discouraging the encoding of demographic information during training can lead to more equitable predictions without severely degrading performance.

This trade-off analysis highlights an important insight: fairness improvements often require careful balancing against accuracy, and no single metric fully captures model quality. By evaluating both dimensions jointly, I was able to make informed decisions about which model best aligns with the goals of ethical and inclusive speech emotion recognition.

10. Final Model Selection

After completing the fairness–accuracy trade-off analysis, I selected a final model that best balances predictive performance, fairness across demographic groups, and stability. Rather than choosing the model with the highest overall accuracy, I prioritised a model that demonstrated more equitable performance across demographic subgroups, in line with the goals of the FairVoice project.

I compared all trained models baseline, oversampling, reweighting, and adversarial debiasing using multiple criteria. These included overall accuracy, subgroup accuracy and F1-scores, fairness gap metrics, and consistency across demographic categories. This multi-criteria evaluation was important because relying on a single metric could mask underlying disparities.

The baseline model, while achieving reasonable accuracy, consistently exhibited the largest demographic performance gaps. These gaps were particularly evident across race and sex groups, confirming that high aggregate accuracy alone is not sufficient for fair deployment.

The oversampling model improved performance for some minority groups, but the improvements were not always consistent across all demographics. In addition, I observed increased variance in performance, especially for groups with very small sample sizes. While oversampling reduced some disparities, it introduced instability that made it less suitable as a final solution.

The reweighted model produced more stable results and achieved moderate reductions in fairness gaps without a substantial loss in accuracy. However, the improvements were less pronounced than those achieved by the adversarial approach, particularly for race-based disparities.

The adversarial debiasing model demonstrated the lowest demographic performance gaps across most evaluated attributes, especially sex and race. Although its overall accuracy was slightly lower than that of the baseline model, the reduction in bias was substantial and consistent. The model also showed more stable behaviour across demographic groups, making it a strong candidate for fair deployment.

Based on this comprehensive evaluation, I selected the adversarial debiasing model as the final model for further analysis and potential deployment. This decision reflects a deliberate trade-off that prioritises fairness and ethical considerations while maintaining acceptable predictive performance.

11. Accent Bias Analysis

In this section, I analysed how the speech emotion recognition model performs when exposed to different accents. The motivation for this analysis was to determine whether a model trained primarily on the CREMA-D dataset, which largely contains American-accented English, generalises fairly to speakers with other accents. Accent variation is an important fairness consideration, especially for speech technologies intended for diverse user populations.

The model used in this analysis was the final adversarially debiased CNN model selected in the previous section. I focused on three accents: American, British, and Indian. These accents were chosen to represent both accents similar to the training data and accents that were underrepresented during training. This allowed me to assess whether reduced exposure during training leads to biased or uncertain predictions.

Because ground-truth emotion labels were not consistently available for all accent samples, I did not rely on classification accuracy alone. Instead, I analysed model confidence and prediction uncertainty, which provide useful indicators of how reliably the model behaves across accents.

11.1 Confidence Analysis

Model confidence reflects how certain the model is about its predicted emotion class. I computed confidence scores by extracting the maximum softmax probability for each prediction and grouped these scores by accent.

The confidence distributions showed that the model generally produced high confidence predictions across all three accents. To test whether confidence differed significantly across accent groups, I applied the Kruskal–Wallis statistical test. The test result indicated no statistically significant difference in confidence between American, British, and Indian accents.

This suggests that, on the surface, the model appears equally confident across accents. However, confidence alone does not guarantee reliable predictions, particularly if the model is overconfident on unfamiliar speech patterns.

11.2 Uncertainty Analysis

To further investigate model behaviour, I measured prediction uncertainty using entropy over the model's output probability distributions. Higher entropy values indicate greater uncertainty in the prediction.

The entropy analysis revealed clear differences across accents. Predictions for Indian-accented speech consistently exhibited higher median entropy values and a wider distribution compared to American and British accents. In contrast, predictions for American and British accents showed very low entropy, indicating strong certainty.

I visualised these results using violin plots and histograms, which made the contrast between accent groups clear. To confirm whether these differences were statistically meaningful, I again applied the Kruskal–Wallis test, which returned a statistically significant result. This confirms that uncertainty varies significantly across accents.

These findings indicate that, although the model reports high confidence across accents, it is substantially more uncertain when processing Indian-accented speech, likely due to limited exposure to similar acoustic patterns during training.

11.3 Predicted Class Distribution

In addition to confidence and uncertainty, I examined the distribution of predicted emotion classes for each accent. I observed that predictions for certain accents were heavily concentrated in a small number of emotion classes. This clustering suggests that the model may not fully capture emotional variability for underrepresented accents.

Such behaviour indicates potential overconfidence and reduced sensitivity to emotional nuance when the model encounters unfamiliar accent characteristics. This pattern further supports the conclusion that accent diversity in training data is critical for robust emotion recognition.

11.4 Summary of Accent Bias Findings

Overall, the accent bias analysis demonstrates that the model does not generalise equally across accents. While confidence scores alone did not reveal significant differences, uncertainty analysis uncovered meaningful disparities, particularly for Indian-accented speech. This highlights the importance of examining uncertainty-based metrics in addition to traditional performance measures.

These results reinforce the broader findings of the FairVoice project: speech emotion recognition models trained on limited accent distributions are prone to bias and reduced reliability when deployed in diverse linguistic settings. Accent-based evaluation provides an essential perspective on fairness, especially in scenarios where labelled data is scarce or unavailable.

12. Limitations

Although this project provides a detailed analysis of fairness, mitigation strategies, and explainability in speech emotion recognition, there are several limitations that should be acknowledged.

One key limitation relates to dataset imbalance. While the CREMA-D dataset includes useful demographic metadata, some demographic groups particularly certain race and ethnicity categories are underrepresented. This imbalance affects both model training and the reliability of fairness metrics, as performance estimates for smaller groups are based on fewer samples and may be more sensitive to noise.

Another limitation concerns the computational cost of explainability methods. Generating SHAP values for spectrogram-based deep learning models is computationally expensive. As a result, SHAP analysis was performed on a limited subset of samples rather than the entire dataset. While this subset provided meaningful insights into model behaviour, a more extensive analysis could reveal additional patterns.

The Grad-CAM explanations are also dependent on model architecture. Grad-CAM highlights salient regions in convolutional layers, but it does not fully explain all internal decision-making processes within the network. Therefore, some aspects of the model's reasoning may not be fully captured by the visualisations presented.

Fairness metrics themselves also have inherent limitations. Measures such as accuracy gaps, demographic parity difference, and equalized odds are influenced by dataset size, label quality, and annotation consistency. In the context of emotion recognition, where labels are subjective by nature, some measured disparities may reflect annotation bias in addition to model bias.

Finally, the generalisability of the results is limited by the nature of the dataset. CREMA-D consists of acted emotional speech recorded under controlled conditions. Real-world speech often includes background noise, spontaneous emotional expression, and greater linguistic and accent diversity. As a result, the findings of this project may not directly translate to real-world deployment without further validation.

Despite these limitations, the project provides a strong and transparent foundation for fairness-aware analysis in speech emotion recognition and highlights key areas for future improvement.

13. Conclusion

In this project, I conducted a comprehensive investigation into bias, fairness, mitigation strategies, and explainability in a Speech Emotion Recognition system trained on the CREMA-D dataset. The primary objective was to determine whether an SER model performs equally across demographic groups and, where disparities exist, to understand their causes and explore practical mitigation strategies.

I began by developing a baseline convolutional neural network and evaluating its performance across demographic attributes such as race, sex, ethnicity, and age. This analysis revealed clear performance disparities, confirming that the baseline model did not treat all demographic groups equally. These findings demonstrated that relying solely on aggregate accuracy metrics can mask important fairness issues.

To address these disparities, I applied three mitigation strategies: oversampling, loss reweighting, and adversarial debiasing. Each approach improved fairness to varying degrees, but they differed in stability and impact on overall accuracy. Through systematic comparison and trade-off analysis, I found that adversarial debiasing offered the most balanced outcome, significantly reducing demographic performance gaps while maintaining acceptable accuracy.

Explainability played a crucial role in understanding model behaviour. By applying SHAP and Grad-CAM, I was able to visualise which acoustic regions influenced the model's predictions and assess whether decisions were driven by emotion-relevant cues or demographic-dependent artefacts. These analyses provided deeper insight into how bias can emerge within learned representations, beyond what performance metrics alone can reveal.

The accent bias analysis further highlighted the limitations of training SER models on datasets with limited linguistic diversity. Although the model appeared confident across accents, uncertainty analysis showed significantly higher uncertainty for Indian-accented speech,

indicating reduced generalisation. This finding reinforces the importance of accent diversity when developing fair and reliable speech systems.

Overall, this project demonstrates that fairness-aware evaluation, mitigation, and explainability are essential components of responsible speech emotion recognition systems. The results show that demographic bias is present in standard SER pipelines but can be reduced through targeted interventions. By combining quantitative fairness metrics with interpretable visual explanations, this work provides a transparent and reproducible framework for analysing and improving fairness in speech-based AI systems.