WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Gender Bias Detection and Mitigation in Audio Deepfake Detection Using Machine Learning and Deep Learning

Shruti Bandekar*[*1], *Trupti Balekundri*[*2], *Tulsi Marennavar*[*3], *Vaishnavi Birje*[*4], *Priyanka Sheelavantar*[*5]

[*1,2,3,4]Department Of Computer Science and Engineering, Angadi Institute Of Technology and Management, Belagavi, Karnataka, India.
[*5]Assistant Professor, Department Of Computer Science and Engineering, Angadi Institute Of Technology and Management, Belagavi, Karnataka, India.

ABSTRACT

The rapid adoption of AI-Generated speech as introduced serious risks to digital trust, enabling realist voice cloning that deceive individuals , organizations, and automated verification system. As audio deepfake become more convincing, the challenge is not only detect them accurately, but also to ensure that detection system treat all speakers fairly. Many existing audio deepfake detection models overlook gender- based variations in vocal characteristic which can lead to uneven performance often identifying manipulated female voices more reliably than male voices or vice-versa. This imbalance raises ethical concerns and limits the reliability of real world applications.

Keywords: Audio deepfake , Gender bias, Machine Learning ,deep learning, MFCC.

1. INTRODUCTION

Artificial Intelligence has enabled the creation of audio deepfakes that can closely imitate real human voices, making them difficult to distinguish from genuine speech. While this technology supports creative and beneficial applications, it is increasingly being misused for fraud, impersonation, and misinformation .Therefore, detecting deepfake audio has become a critical requirement to protect digital security and public trust. However, many existing detection system overlook gender -based variation speech ,leading to unfair performance differences between male and female voices. Such bias reduces the reliability and ethical credibility of AI systems deployed in real- world environments. This project aims to develop a gender-aware audio deepfake detection model using balanced datasets and advanced ML and DL techniques to improve fairness and accuracy.

2. METHODOLOGY

1.Dataset preparation :

A Gender -Balanced dataset was created with equal real and fake audio samples from male and female speaker. All files were manually labeled and organized into training, validation, testing sets for fair evaluation.

2.Preprocessing:

Audio recording were clean, resampled, segmented into smaller clip for efficient processing Background noise removal and normalization ensured consistency across sample.

3.Feature Extraction:

Both acoustic numerical features (MFCC, spectral features) and visuals mel spectrograms features were extracted. these feature represent voice characteristics crucial for distinguishing real vs fake audio.

4.Model Training:

Machine learning (XGboost)was trained using numerical features while deep learning (CNN) learn patterns from spectrograms images. Separate models were trained on male -only female-only and mixed dataset to analyze performance difference only.

5.Model Evaluation:

Accuracy, F1-Score, Equal Error Rate (EER), metrics were calculated to measure detection performance. Results were compared across gender groups to identify any existing performance bias .

6.Bias Analysis and Mitigation:

Differences in accuracy between male and female and evaluated to detect gender bias. Gender-Balanced training and optimized parameter were applied to reduce bias and improve fairness.

3. MODELING AND ANALYSIS

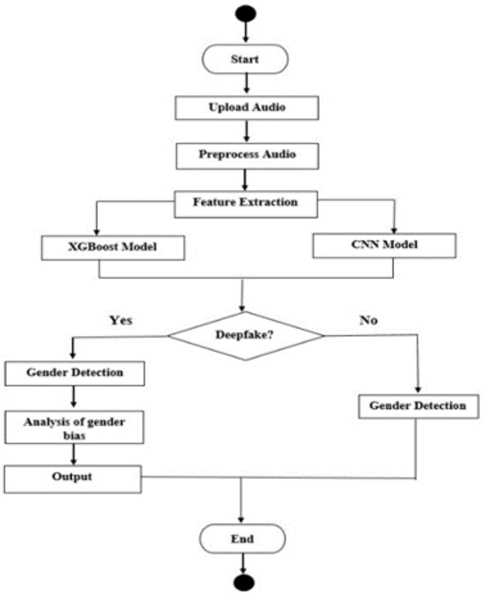


Figure1:- System Workflow

- The system implements two models: an XGBoost classifier using extracted audio features and a CNN deep learning model using Mel-Spectrogram images.
- Each model is trained separately on male-only, female-only, and mixed-gender datasets to understand performance variations between genders.
- Feature engineering and normalization techniques are applied to improve learning efficiency and reduce noise impact.

4. RESULTS AND DISCUSSION

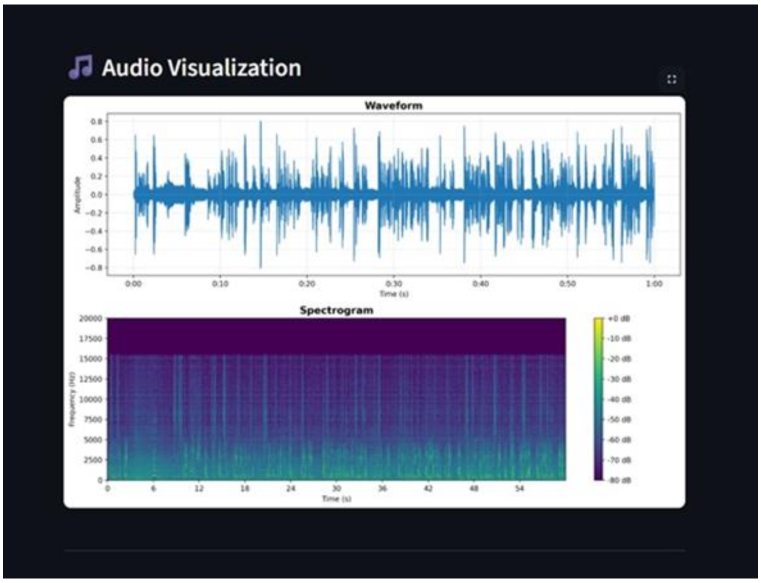
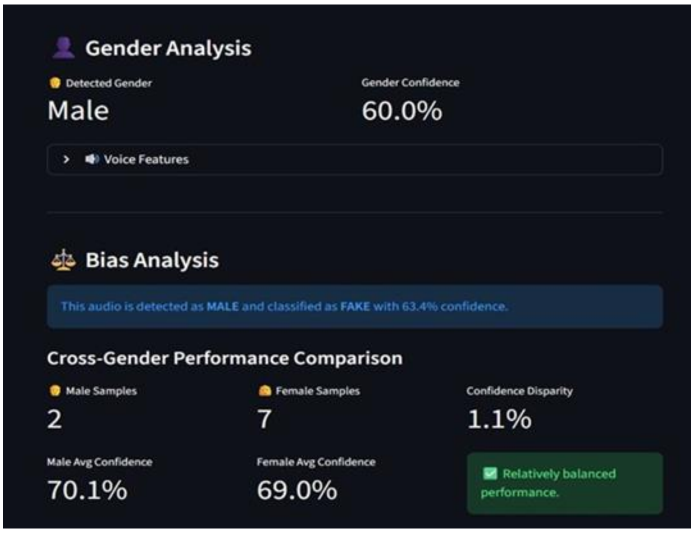


Figure2 :- Gender Analysis and Bias Analysis

Each model is trained separately on male-only, female-only, and mixed-gender datasets to understand performance variations between genders. Feature engineering and normalization techniques are applied to improve learning efficiency and reduce noise impact. Hyperparameter tuning and cross-validation are performed to achieve optimal detection accuracy and robustness.



Gender Analysis

Detected Gender: Male

Gender Confidence: 60.0%

Bias Analysis

This audio is detected as **MALE** and classified as **FAKE** with 63.4% confidence.

Cross-Gender Performance Comparison

Male Samples: 2

Female Samples: 7

Confidence Disparity: 1.1%

Male Avg Confidence: 70.1%

Female Avg Confidence: 69.0%

✓ Relatively balanced performance.

Figure2:- Mel-Spectrogram

The models effectively detected audio deepfakes, though performance was slightly higher for male voices, indicating gender bias. Using balanced datasets reduced this disparity, improving fairness across genders. Deep learning models achieved higher overall accuracy, while traditional models handled minority gender samples more consistently. These findings highlight the importance of bias mitigation for reliable and ethical deepfake detection.

5. CONCLUSION

This project highlights the critical challenge of gender bias in audio deepfake detection systems. By analyzing real and synthetic speech data, we demonstrated that existing detection models can exhibit unequal performance across genders, potentially leading to unfair or unreliable outcomes. Implementing gender-aware training strategies and balanced datasets helps reduce this disparity, improving both fairness and accuracy. The study emphasizes the importance of ethical considerations in AI, advocating for models that are robust, unbiased, and socially responsible. Ultimately, this work contributes to safer and more trustworthy deepfake detection systems while laying the groundwork for future research in bias mitigation.

ACKNOWLEDGEMENTS

We express our sincere gratitude to our Principal, Dr. Anand Deshpande, and HOD, Dr. Dhanashree Kulkarni, for their encouragement and support throughout this research. We also thank our project guide, Prof. Priyanka Sheelavantar, for valuable guidance, motivation, and constant supervision. Lastly, we extend our appreciation to our parents and friends for their continuous encouragement.

6. REFERENCES

[1]Bird, J. J., & Lotfi, A. (2023). Real-time Detection of AI-Generated Speech for Deep Fake Voice Conversion. arXiv preprint arXiv:2308.12734. unpublished.

[2]Chadha, A., Kumar, V., Kashyap, S., & Gupta, M. (2021). Deepfake: an overview. In Proceedings of Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020 (pp. 557-566). Springer Singapore.

[3]Pu, M., Kuan, M. Y., Lim, N. T., Chong, C. Y., & Lim, M. K. (2022, May). Fairness evaluation in deepfake detection models using metamorphic testing. In Proceedings of the 7th International Workshop on Metamorphic Testing (pp. 7-14).

[4] Xu, Y., Terhörst, P., Raja, K., & Pedersen, M. (2022). A comprehensive analysis of ai biases in deepfake detection with massively annotated databases. arXiv preprint arXiv:2208.05845. unpublished.

[5]Trinh, L., & Liu, Y. (2021). An examination of fairness of ai models for deepfake detection. arXiv preprint arXiv:2105.00558. unpublished.