ABSTRACT

The proliferation of audio deepfakes, generated using advanced technologies such as WaveNet and Generative Adversarial Networks (GANs), poses significant threats to digital security, including identity theft, misinformation, and fraud. To address these challenges, this study proposes an end-to end framework for audio deepfake detection that leverages Mel Spectrograms as input features and the Xception model as the backbone architecture. The methodology includes optimized preprocessing techniques, such as normalization and resizing, and robust data augmentation strategies to enhance feature quality and model generalization. The framework was evaluated using the Automatic Speaker Verification (ASV) spoof 2021 dataset, achieving a high test accuracy of 95.86% with balanced precision, recall, and F1-scores for ‘real‘ and ‘fake‘ classifications. Comparative analysis demonstrated that the Xception model outperformed ResNet50 and MobileNetV2 in both accuracy and generalization. While the results highlight the robustness and efficiency of the proposed framework, future research could ex plore advanced preprocessing pipelines, hybrid architectures, and diverse datasets to further improve detection performance. This study provides a reliable and efficient solution for safeguarding against the growing threats posed by audio deepfakes.

Keywords: Audio Deepfakes, Mel Spectrograms, Xception Model, Deep Learning, Digital Security