Deepfake Audio Detection Using CNN-Transformer Hybrid Model with Data Augmentation

Archana Kadam
Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune

Anushka Yadav
Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune

Shraddha Zoman
Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune

Tanvi Unhale
Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune

Rutika Umale
Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune

Abstract:

The emergence of deepfake audio generated through advanced machine learning models such as GANs and speech synthesis networks presents serious threats to digital security and trust. In this paper, we propose a CNN-Transformer hybrid architecture for detecting deepfake audio signals. The CNN extracts local spectral features while the Transformer captures long-range temporal dependencies across audio sequences. Evaluated on the ASVspoof 2019 dataset, the model achieved a classification accuracy of 91.47%, outperforming conventional models including LSTM (90.00%), CNN-LSTM (91.39%), and TCN (86.96%). A detailed classification report and confusion matrix further demonstrate the robustness of the proposed approach. The approach builds upon trends observed in prior works using spectral learning, adversarial learning, and hybrid audio forensics architectures.

Keywords:

CNN-Transformer Hybrid, Data Augmentation, Deepfake Audio Detection, Spectrogram Analysis.