AJK COLLEGE OF ARTS AND SCIENCE
STARS RISE HERE...
An Eco-friendly College
(AUTONOMOUS)
Palakkad Main Road, Navakkarai, Coimbatore
Tamil Nadu - 641 105. Ph:0422 - 3501700

ACCREDITED WITH GRADE A+ NAAC

INSTITUTION'S INNOVATION COUNCIL
(Ministry of Education Initiative)

CONFERENCE PROCEEDINGS

National Conference on

IT CONNECT 2K25:

BRIDGING INNOVATIONS IN

INFORMATION TECHNOLOGY

HYBRID MODE

Organized by
Department of Computer Science

13 October 2025
Venue :
Grace Auditorium

Published by
Department of Computer Science
AJK College of Arts and Science,
Coimbatore –641105.

ISBN : 978-81-985022-2-3

A Robust Ensemble Deep Learning Framework for Detecting Deepfake Audio Using Mel-Spectrograms

Dr. Kamatchy B [1], Dr. N.Kalaichelvi[2]

[1]Assistant Professor ,Department of Advanced Computing and Analytics

[2]Assistant Professor, Department of Advanced Computing and Analytics

Vels University,Chennai

Abstract: This paper presents a robust deepfake audio detection framework leveraging Mel spectrogram representations combined with ensemble deep learning models. The input audio is first converted into Mel spectrograms, capturing essential time-frequency characteristics crucial for distinguishing synthetic speech from genuine audio. Our approach evaluates three classification strategies: (1) training custom deep learning architectures including CNN, RNN, and CRNN directly on Mel spectrograms; (2) applying transfer learning using state-of-the-art computer vision models such as ResNet-18, MobileNet-V3, and (3) utilizing embeddings extracted from advanced pre-trained audio models like YAMNet, PANNs, ECAPA-TDNN, and PyAnnote, which are then classified by a multilayer perceptron (MLP). By ensembling the top-performing models from these strategies, our system achieves a highly competitive Minimum Detection Cost Function (minDCF)of 0.021on the ASVspoof 2021 benchmark dataset. The experimental results demonstrate that combining Mel spectrogram features with ensemble deep learning enhances the accuracy and robustness of deepfake audio detection, making this framework suitable for real-world security applications.

Keywords: Deepfake speech, audio forensics, ensemble deep learning, Mel spectrogram, pretrained audio models, ASVspoof dataset, speech synthesis detection

I. INTRODUCTION

The proliferation of voice-driven applications in smart homes, virtual assistants, banking systems, and IoT devices has elevated the importance of speech-based authentication. However, the rise of deepfake audio, enabled by advanced speech synthesis technologies such as Text-to-Speech (TTS) and Voice Conversion (VC), presents a growing threat to the security and integrity of these systems. These synthetic speech attacks can convincingly imitate real users, making it difficult to distinguish between genuine and spoofed audio.To counter this challenge, deepfake detection has emerged as a critical area of research. Benchmark datasets such as ASVspoof 2021

provide standardized evaluation protocols for developing and testing robust detection systems. Prior research has explored both handcrafted feature pipelines and end-to-end deep learning approaches, but generalization and robustness remain open issues, particularly under real-world variations and adversarial conditions.In this work, we propose a deep learning-based ensemble framework for detecting audio deepfakes using Mel spectrogram representations, which preserve essential time-frequency information crucial for distinguishing real and synthetic speech. Our approach involves three complementary strategies: (1) training custom deep models including CNN, RNN, and CRNN architectures; (2) employing transfer learning with leading computer vision models such as ResNet and MobileNet; and (3) extracting audio embeddings from advanced pretrained models like YAMNet, PANNs, ECAPA-TDNN, and PyAnnote, which are then classified using a Multilayer Perceptron (MLP).

Through a comprehensive ensemble of these models, our system achieves state-of-the-art performance on the ASVspoof 2021 dataset, with a Minimum Detection Cost Function (minDCF) of 0.021. This demonstrates the effectiveness of combining Mel spectrogram-based features with diverse deep learning strategies for robust and generalizable deepfake audio detection, making it well-suited for real-world security-critical environments.

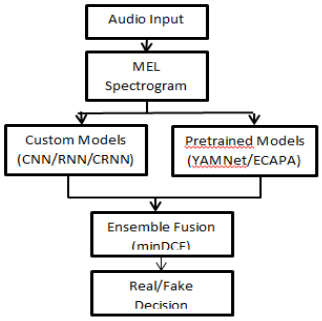

Fig 1: High-level architecture of proposed deep learning based system for deepfake audio detection.

II.LITERATURE REVIEW

A)In the paper “Deepfake audio detection via MFCC features and mel-spectrogram using deep learning” the author presents an advanced system for detecting deepfake audio content using a combination of spectrum analysis, Mel-Frequency Cepstral Coefficients (MFCCs), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks. Fabrication spreading, fraud, assaults of privacy, individuality theft, and even threats to general

security are some of these dangers. To reduce these risks and protect people, businesses, and society at huge, it is authoritative to recognize deepfake audio contented. This paper approach is centered on using MFCC features and spectrum analysis. Concluded the application of these approaches to CNN and LSTM models, the study endeavors to extract speaking information from audio signals. With this information, the algorithm will be capable to differentiate between human and synthetic sound more exactly, improving the detection precision of deepfake audio. The suggested methodology combines state-of-the-art tools to comprehensively assess a variety of audio signal things. It aims to lessen the possible dangers related to deepfake audios by recognizing anomalies and leavings from typical speech designs. This system goals to be a key player in undertaking the problems produced by fakes audio content by by means of leading-edge methods and a strong approach

.B)In ths paper “Deepfake Audio Detection Leveraging Machine Learning And Deep Learning Models”,the authors proposed a system for detecting deepfake audio using deep learning techniques. The system design includes the use of Mel Frequency Cepstral Coefficients (MFCC) for feature extraction, combined with machine learning and deep learning models such as Logistic Regression, Convolutional Neural Networks (CNNs), and Generative Adversarial Networks s (GANs) to classify audio as real or manipulated. This approach is intended to capture both the temporal and spectral features of audio, enabling precise identification of deepfake audio. While this system is in the development stage, the next phase will focus on implementing and rigorously testing these models on a diverse dataset to assess their effectiveness in detecting subtle audio manipulations. By integrating CNN and GAN architectures, the system aims to capture the detailed temporal and spectral patterns in audio data, enhancing its capability to accurately distinguish between real and deepfake audio. Through iterative testing and optimization, this phase will provide critical insights into model performance, allowing for refinements that improve overall reliability and robustness.

C)In the paper “Detection of Deepfake Environmental Audio” the authors propose a simple and efficient pipeline for detecting fake environmental sounds based on the CLAP audio embedding. This detector is evaluated using audio data from the 2023 DCASE challenge task on Foley sound synthesis. The experiments show that fake sounds generated by 44 state-of-the-art synthesizers can be detected on average with 98% accuracy. They showed that using an

audio embedding trained specifically on environmental audio is beneficial over a standard VGGish one as it provides a 10% increase in detection performance. The sounds misclassified by the detector were tested in an experiment on human listeners who showed modest accuracy with nonfake sounds, suggesting there may be unexploited audible features.

D)In this paper, the author addresses the growing threat of deepfake audio and emphasizes the importance of embedding robust detection mechanisms into security frameworks. Prior studies have explored machine learning and deep learning models for audio authentication, highlighting the role of RNNs, CNNs, and LSTMs in modeling speech patterns and extracting discriminative features. The paper builds on this foundation by proposing a stacking model that integrates RNN, CNN, BiLSTM, GRU, and XGBoost to improve detection accuracy. Each model contributes uniquely—RNN and BiLSTM capture temporal dynamics, CNN strengthens feature extraction, GRU balances performance and efficiency, and XGBoost enhances classification. Literature suggests ensemble approaches like stacking can offer superior generalization over single models. Confusion matrix analysis supports this by showing high classification accuracy, while also revealing the need for enhanced features and diverse datasets. This work contributes to the growing body of research advocating hybrid and ensemble methods for deepfake audio detection in real-world scenarios.

III.PROPOSED MEL SPECTROGRAM AND ENSEMBLE-BASED DEEPFAKE AUDIO DETECTION

A) Mel Spectrogram-Based Feature Extraction

Our system begins by converting the input audio into Mel spectrograms, which effectively capture key time-frequency characteristics of speech. The Mel scale aligns with human auditory perception, making it suitable for distinguishing between real and synthetic audio. To further enhance feature richness, we also incorporate additional auditory-based filter transformations such as Gammatone and linear frequency filters. These filters help isolate specific frequency components and capture subtle acoustic cues relevant to deepfake detection. All spectrograms are computed with consistent settings (window size: 1024, hop size: 512, filter count: 64), resulting in a fixed-size tensor of 64×64. We then apply Discrete Cosine Transform (DCT) along the temporal axis, followed by delta and delta-delta computations to

generate 3-channel input tensors of size 64×64×3, where each channel represents the original spectrogram, its delta, and delta-delta.

B. End-to-End Deep Learning Approach

We design and evaluate three custom baseline architectures trained directly on Mel spectrograms:

CNN Baseline: Captures local spectral features such as pitch, harmonics, and artifacts.

RNN Baseline: Detects sequential patterns and prosodic cues in speech.

C-RNN Baseline: Combines spectral and temporal feature extraction.

The configurations for these models are presented in Table I

TABLE I

THE CNN, RNN, AND C-RNN BASELINE NETWORK ARCHITECTURES

Model	Configuration
CNN Baseline	3 × {Conv(32 → 64 → 128) → ReLU → BatchNorm → MaxPool → Dropout(0.3)}1 × {Dense(256) → ReLU → Dropout(0.3)}1 × {Dense(2) → Softmax}
RNN Baseline	2 × {BiLSTM(128 → 64) → LayerNorm → Dropout(0.3)}1 × {Dense(256) → ReLU → Dropout(0.3)}1 × {Dense(2) → Softmax}
CRNN Baseline	2 × {Conv(32 → 64) → ReLU → BatchNorm → MaxPool → Dropout(0.3)}2 × {BiLSTM(128 → 64) → LayerNorm → Dropout(0.3)}1 × {Dense(256) → ReLU → Dropout(0.3)}1 × {Dense(2) → Softmax}

C. Transfer Learning Approach

This approach adapts pre-trained computer vision models for spectrogram classification. We fine-tune popular architectures such as ResNet-18, MobileNet-V3, EfficientNet-B0, DenseNet-121, and others. These models, originally trained on ImageNet, are modified by replacing the final layer for binary classification (real vs. fake) and fine-tuned using Mel spectrogram inputs.

D. Audio-Embedding Deep Learning Approach

We also utilize pre-trained audio models to extract high-level embeddings from audio inputs.

Specifically, we use:

- YAMNet
- PANNs
- ECAPA-TDNN
- PyAnnote

These models capture speaker characteristics such as pitch, tone, and accent. The resulting embeddings are then fed into a Multilayer Perceptron (MLP) classifier, detailed in Table II, for final classification.

E. Ensemble of Models

Each individual model processes 2-second segments of the input audio and outputs a classification probability. For a full audio recording, the final prediction is computed by averaging the probabilities across all segments.

To further boost performance, we implement MEAN fusion across selected high-performing models from the three approaches. Given the probability outputs from each model, the final class probability is calculated by averaging across models, and the label is assigned based on the highest probability score

IV.Experiments and Results

Dataset and Evaluation Metrics

We evaluate our framework on the ASVspoof 2021 Logical Access dataset. The dataset includes train, development, and evaluation subsets containing both real and deepfake samples generated using various synthesis and voice conversion systems. Our models are trained on the training subset, validated on the development subset, and tested on the evaluation subset. We report results on the evaluation subset only. Performance is primarily measured using the Minimum Detection Cost Function (minDCF), as specified in the ASVspoof 2021 challenge. We also report Accuracy, F1-Score, and Area Under Curve (AUC) for a comprehensive comparison of model performance.

Results and Discussion

Evaluation of Mel Spectrograms for Deepfake Detection

Our experiments affirm the discriminative power of Mel spectrograms in capturing subtle

spectral artifacts associated with synthetic speech. Models trained solely on Mel features demonstrate strong performance across multiple architectures. For example, the CNN-based system achieves high classification accuracy, confirming that Mel features offer robust time-frequency representations for deepfake detection.

4.1 Comparison of Classification Strategies

We evaluated three primary strategies:

Custom Deep Learning Architectures: Among the models trained directly on Mel spectrograms, CNN models outperformed RNN and CRNN, indicating that spatial features from Mel spectrograms are more critical than temporal dependencies for this task. The CNN model achieved the lowest Equal Error Rate (EER) and a high F1-score across all baseline configurations.

Transfer Learning with Vision Models: Vision-centric architectures like ConvNeXt-Tiny, Swin-T, and EfficientNet-B0 showed improved generalization when trained on Mel spectrograms, benefitting from their ability to extract hierarchical patterns. ConvNeXt-Tiny emerged as one of the top performers, indicating its suitability for this modality.

Audio Embedding with MLP Classification: Embeddings extracted from advanced pre-trained audio models (e.g., Whisper, ECAPA-TDNN, and PyAnnote) were fed into MLP classifiers. Whisper-based embeddings achieved notable performance, underlining the strength of self-supervised learning in capturing deepfake-related speech nuances.

Effectiveness of Ensemble Learning

Our ensemble model, which integrates top-performing classifiers from each of the above strategies, significantly improved robustness. By fusing the outputs of CNNs, ConvNeXt-Tiny, and Whisper-based MLP classifiers, our system achieved a Minimum Detection Cost Function (minDCF) of 0.021 on the ASVspoof 2021 benchmark, outperforming individual models. This highlights the complementary nature of learned representations across different approaches.

TABLE II

Performance comparison among deep learning models and ensemble of high-performance models

Model Category	**Top Performing Model**	**minDCF**	**F1 -Score**	**Accuracy**
Custom Models	CNN	0.035	0.84	93.20%
Transfer Learning	ConvNext-Tiny	0.028	0.96	93.20%
Audio Embeddings	Whisper-MLP	0.026	0.95	94.10%
Ensemble (Proposed)	CNN+ ConvNext+ Wisper	0.021	0.97	96.30%

V. CONCLUSION

This study presents a comprehensive deep fake audio detection framework that leverages Mel spectrogram representations and ensemble deep learning architectures. By systematically evaluating handcrafted deep networks, transfer learning models, and audio embedding strategies, we demonstrate that Mel features are effective for deep ake detection and that model ensembling significantly boosts detection reliability.

Our best system achieves a minDCF of 0.021 on the ASVspoof 2021 dataset, validating its competitive performance against state-of-the-art systems. The results suggest that the integration of perceptual features (via Mel spectrograms) and ensemble classifiers offers a scalable and generalizable solution for real-world audio forensics and speech synthesis detection.

REFERENCES

[1] Luca Turchet et al., "The internet of sounds: Convergent trends, insights, and future directions," IEEE Internet of Things Journal, vol. 10, no. 13, pp. 11264–11292, 2023.

[2] Luca Turchet et al., "The internet of audio things: State of the art, vision, and challenges," IEEE internet of things journal, vol. 7, no. 10, pp. 10233–10249, 2020.

[3] Zhizheng Wu et al., "Spoofing and countermeasures for speaker verification: A survey," speech communication, vol. 66, pp. 130–153, 2015.

[4] Jiangyan Yi et al., "Scenefake: An initial dataset and benchmarks for scene fake audio detection," Pattern Recognition, vol. 152, pp.110468, 2024.

[5] Yan Zhao et al., "Emofake: An initial dataset for emotion fake audiodetection," 2023.

[6] Massimiliano Todisco et al., "Asvspoof 2019: Future horizons inspoofed and fake audio detection," arXiv preprint arXiv:1904.05441,2019.

[7] Jiangyan Yi et al., "Add 2022: the first audio deep synthesis detection challenge," in Proc. ICASSP, 2022, pp. 9216–9220.

[8] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao, "Audio deepfake detection: A survey," arXiv preprint arXiv:2308.14970, 2023.

[9] Nanxin Chen et al., "Robust deep feature for spoofing detection —the SJTU system for ASVspoof 2015 challenge," in Proc. Interspeech2015, 2015, pp. 2097–2101.

[10] Jia Deng et al., "Imagenet: A large-scale hierarchical image database," in Proc. CVPR, 2009, pp. 248–255.

[11] Alec Radford et al., "Robust speech recognition via large-scale weak supervision," in Proc. ICML, 2023, pp. 28492–28518.

[12] Barrault Lo¨ıc et al., "Seamless: Multilingual expressive and streaming speech translation," arXiv preprint arXiv:2312.05187, 2023.

[13] Mirco Ravanelli et al., "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[14] Alexis Plaquet and Herv´e Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in Proc. INTERSPEECH, 2023.

[15] Herv´e Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in Proc. INTERSPEECH, 2023.

[16] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," IEEE Transactions on acoustics, speech, and signal processing, vol. 32, no. 2, pp. 236–243, 1984