iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XI **Month of publication:** November 2024

DOI: https://doi.org/10.22214/ijraset.2024.65027

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XI Nov 2024- Available at www.ijraset.com

Advanced Sequential Modeling for DeepFake Audio Identification

Priyadarshan Dhabe[1], Nitin Choudhary[2], Ayush Vidhale[3], Yash Munde[4], Muaz Sayyed[5], Netal Zanwar[6]
Department of Information Technology Vishwakarma Institute of Technology Pune, India

Abstract: ***The research paper is about audio DeepFakes, which are fake audio clips that sound just like real people's voices. These can be used to spread false information or pretend to be someone else, which is a big problem.The paper looks at different ways to tell if an audio clip is a DeepFake or not. It uses advanced computer techniques, like generative adversarial networks (GANs), to do this. The paper tests different methods, like using raw sound waves, Mel-frequency cepstral coefficients (MFCCs), and linear frequency cepstral coefficients (LFCCs) as inputs. One of the methods tested is the Time-Domain Synthetic Speech Detection (TSSD) model, which takes raw audio waveforms as input. The tests are done on the WaveFake dataset, which has synthetic audio made by six different GAN architectures. The paper measures how well each method works using things like the equal error rate (EER), F1 score, and area under the receiver operating characteristic curve (ROC AUC). The results show that some methods, like the shallow CNN and TSSD architectures, are good at detecting audio DeepFakes. But, the paper also points out that there's still room for improvement, especially for cases where the fake audio is made to trick the detection methods.***
Keywords: Deepfake, GAN, MFCC, LFCC, CNN, TSSD

I. INTRODUCTION

Rapid progress in deep learning and generative models, has resulted in an increase in synthetic media, including audio deepfakes. Audio deepfakes are artificially created audio clips that mimic human speech so realistically that they can be mistaken for the real thing. These are usually created using deep neural networks that have been trained on large datasets of human speech recordings.While these audio deepfakes can be used for creative applications such as text-to-speech synthesis and voice conversion, they also pose significant risks. These risks include misuse for malicious purposes such as online harassment, identity theft, financial fraud, or spreading misinformation.The research paper highlights a particularly worrying use case where audio deepfakes could be used as a tool for information warfare. An example of this is an incident during the ongoing Russia-Ukraine conflict, where a deepfake video surfaced online, allegedly showing Ukrainian President Volodymyr Zelenskyy urging his forces to surrender. Although this was quickly debunked, it served to highlight the threat posed by such manipulated media. As technology improves, distinguishing between deepfake audio and real recordings is becoming an increasingly complex challenge. The research paper discusses the active area of research that is the detection of audio deepfakes. Various approaches have been proposed in recent years. Traditional methods often rely on hand-engineered audio features like mel-frequency cepstral coefficients (MFCCs) or linear frequency cepstral coefficients (LFCCs), which are then fed into statistical models or shallow machine learning classifiers. However, these techniques may struggle to keep up with the rapid evolution of deepfake generation methods, which are underpinned by sophisticated deep neural networks. In response to this, researchers have begun exploring deep learning-based detection frameworks that can automatically learn discriminative representations directly from audio data. Some approaches leverage sequential modeling architectures like recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, which are well-suited for processing the temporal nature of audio signals. Other methods employ convolutional neural networks (CNNs) to capture local patterns in spectrogram-like representations of the audio. End-to-end models that operate directly on raw waveform inputs have also been developed. Despite these efforts, reliably detecting audio deepfakes remains a formidable task, especially as generative models continue to improve. The research paper aims to conduct a comprehensive evaluation of various deep learning architectures for audio deepfake detection. It explores different model designs, input representations, and training strategies using a challenging benchmark dataset containing state-of-the-art deepfake samples. The goal is to gain insights into the strengths and limitations of current techniques, and to identify promising directions for future research in this critical domain. This expanded introduction provides a more detailed overview of the research paper's focus and objectives.

II. LITERATURE SURVEY

[1] In their paper, Subramani and Rao propose efficient models for fake speech detection. They introduce four convolutional architectures and a novel multi-task problem, achieving high F1 scores. Using transfer learning, they tackle data sparsity in adversarial settings and demonstrate effective adaptation to new attack vectors. Their research offers promising strategies for detecting synthetic speech. [2] The paper "DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices" introduces DeepSonar, a method for detecting AI-synthesized fake voices. It leverages neuron behaviors of a deep neural network in a speaker recognition system. The approach shows high detection rates and robustness against manipulation attacks, providing new insights into multimedia fakes forensics. [3] This paper presents a novel approach for distinguishing between human and bot speakers. The authors propose a speech classification algorithm based on Convolutional Neural Networks (CNNs), which enables the automatic classification of human vs non-human speakers from the analysis of short audio excerpts. This is particularly relevant due to the increasing sophistication of automatic speech generation algorithms, which have led to more convincing machine-to-human interactions. [4] The paper introduces a method for detecting partially fake audio. The authors propose a fake span detection module that predicts the start and end positions of the fake clip within the audio. This approach enhances the model's discrimination capacity between real and partially fake audio, ranking second in the partially fake audio detection track of ADD 2022. [5] In the paper, authors Jee-weon Jung, Hye-jin Shim, Hee-Soo Heo, and Ha-Jin Yu propose an end-to-end deep neural network (DNN) that uses spectrograms as direct input for detecting replay attacks in speech. The approach explores complementary information and high-resolution details, assuming that the difference between bona-fide and replayed speech exists in these details. The authors also examine raw waveform and an i-vector-based system to verify whether other features are complementary to spectrograms. The method shows promising results in the ASVspoof 2019 physical access challenge, with t-DCF and equal error rates of 0.0570 and 2.45% for the evaluation set, respectively. [6] In the paper authors proposed a novel model structure, Res2Net, to improve the generalizability of anti-spoofing countermeasures. Res2Net modifies the ResNet block to enable multiple feature scales, which increases the possible receptive fields and results in multiple feature scales. This mechanism significantly improves the countermeasure's generalizability to unseen spoofing attacks and decreases the model size compared to ResNet-based models. The Res2Net model consistently outperforms ResNet34 and ResNet50 in both physical access (PA) and logical access (LA) of the ASVspoof 2019 corpus. [7] In the paper, authors proposed a classification method based on Hidden Markov Models (HMM) for audio keyword identification. This approach is an improvement over their previous work that used a hierarchical Support Vector Machine (SVM) classifier and segmented audio signals into 20 ms frames without any contextual information. The HMM-based classifiers treat specific sound as a continuous time series data and employ hidden states transition to capture context information. The authors also study how to find an effective HMM by determining its topology, observation vectors, and statistical parameters. Experimental results show that the proposed HMM-based method outperforms the previous hierarchical SVM. [8] The LJ Speech Dataset, created by Keith Ito and Linda Johnson, is a public-domain speech dataset. It comprises 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books. Each clip varies in length from 1 to 10 seconds, with a total length of approximately 24 hours. A transcription is provided for each clip, and the dataset also includes a normalized version of the text. The texts were published between 1884 and 1964 and are in the public domain. The audio was recorded in 2016-17 by the LibriVox project. This dataset can be used to train models for tasks such as Automatic Speech Recognition (ASR) or Text-to-Speech (TTS). [9] "WaveFake: A Data Set to Facilitate Audio Deepfake Detection" is a research paper by Joel Frank and Lea Schönherr presented at the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track in 2021. The paper addresses the potential harm caused by deep generative modeling, particularly in the audio domain, which has been relatively neglected compared to image-based deepfakes. The authors present a novel dataset, WaveFake, collected from six different network architectures across two languages. They also provide an introduction to common signal processing techniques used for analyzing audio signals. The paper includes an analysis of frequency statistics, revealing subtle differences between the architectures, especially among the higher frequencies. The authors also provide baseline models from the signal-processing community to facilitate further research. [10] In the paper, authors Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville propose a novel approach for generating high-quality coherent waveforms using Generative Adversarial Networks (GANs). They introduce architectural changes and simple training techniques that allow GANs to generate these waveforms reliably. The effectiveness of their approach is demonstrated through a subjective evaluation metric (Mean Opinion Score, or MOS) for high-quality mel-spectrogram inversion. The authors also show qualitative results of their model in speech synthesis, music domain translation, and unconditional music synthesis. Their model is non-autoregressive, fully convolutional, has significantly fewer parameters than competing models, and generalizes to unseen speakers for mel-spectrogram inversion.

[11] In the paper, authors introduced WaveNet, a deep neural network for generating raw audio waveforms. The model is fully probabilistic and autoregressive, with the predictive distribution for each audio sample conditioned on all previous ones. Despite the complexity, it can be efficiently trained on data with tens of thousands of samples per second of audio. When applied to text-to-speech, it yields state-of-the-art performance, with human listeners rating it as significantly more natural sounding than the best parametric and concatenative systems for both English and Mandarin. [12] In the paper, the authors proposed HiFi-GAN, a GAN-based model capable of generating high-fidelity speech efficiently. The model leverages sinusoidal signals with various periods, crucial for enhancing sample quality. A subjective human evaluation (mean opinion score, MOS) of a single speaker dataset indicates that HiFi-GAN demonstrates similarity to human quality while generating 22.05 kHz high-fidelity audio 167.9 times faster than real-time on a single V100 GPU. The generality of HiFi-GAN is shown in the mel-spectrogram inversion of unseen speakers and end-to-end speech synthesis. [13] In the paper, authors Durk P Kingma and Prafulla Dhariwal introduce Glow, a generative flow model that uses invertible 1x1 convolutions. This model extends previous work on reversible generative models and simplifies the architecture. Glow can generate realistic high-resolution images, supports efficient sampling, and discovers features that can be used to manipulate attributes of data. The authors demonstrate a significant improvement in log-likelihood on standard benchmarks, showing that a generative model optimized towards the plain log-likelihood objective is capable of efficient and realistic-looking synthesis. [14] In the paper authors, focus on enhancing the security of Automatic Speech Verification (ASV), a technology that verifies a person's identity based on their voice. Recognizing the need for robust and efficient countermeasures, they follow the setup in the ASVSpoof 2019 competition. They evaluate their system using two metrics, EER and t-DCF, aiming to achieve the highest system security standard as ASV safeguards valuable digital assets. [15] In the paper, authors K. N. R. K. Raju Alluri and Anil Kumar Vuppala present the IIIT-H spoofing countermeasures developed for the ASVspoof 2019 challenge. They use three instantaneous cepstral features, namely, single frequency cepstral coefficients, zero time windowing cepstral coefficients, and instantaneous frequency cepstral coefficients as front-end features. This approach aims to enhance the security of Automatic Speaker Verification (ASV) systems by providing robust and efficient countermeasures against spoofing attacks. [16] In the paper, authors present a robust approach to detecting spoofing attempts in Automatic Speaker Verification (ASV) systems. They use both deep neural networks and traditional machine learning models, combining them as ensemble models through logistic regression. The models are trained to detect logical access (LA) and physical access (PA) attacks on the dataset released as part of the ASV Spoofing and Countermeasures Challenge 2019. Their ensemble model outperforms all their single models and the baselines from the challenge for both attack types. [17] In the paper, authors propose a shift from traditional synthetic speech detection methods to end-to-end deep neural networks (DNNs). They introduce a lightweight neural network model, Time-domain Synthetic Speech Detection Net (TSSDNet), which outperforms state-of-the-art methods for the ASVspoof2019 challenge. The model, trained on ASVspoof2019, also demonstrates promising detection performance when tested on disjoint ASVspoof2015, significantly better than existing cross-dataset results. This work highlights the potential of end-to-end DNNs for synthetic speech detection, eliminating the need for hand-crafted features. [18] In the paper, authors introduce a residual learning framework to ease the training of substantially deeper networks. They reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. This approach allows them to train neural networks that are up to 152 layers deep—8x deeper than previous models, but with lower complexity. Their residual networks are easier to optimize and can gain accuracy from considerably increased depth. The effectiveness of this approach is demonstrated on the ImageNet dataset, where an ensemble of these residual nets achieves a 3.57% error rate. [19] In this paper, the authors present first application of RawNet2 to anti-spoofing. RawNet2 ingests raw audio and has the potential to learn cues that are not detectable using more traditional countermeasure solutions. They describe modifications made to the original RawNet2 architecture so that it can be applied to anti-spoofing. For A17 attacks, their RawNet2 systems results are the second-best reported, while the fusion of RawNet2 and baseline countermeasures gives the second-best results reported for the full ASVspoof 2019 logical access condition.

III. METHODOLOGY

The methodology of deepfake audio detection is as follows:

A. Dataset

For the experiments, WaveFake dataset introduced by Frank and Schönherr is used. This serves as a challenging benchmark for audio deepfake detection, containing both real human speech recordings as well as a diverse set of synthetic audio samples generated by state-of-the-art generative models.

The real audio data is sourced from the LJSpeech dataset, which comprises 13,100 short clips of a single female speaker reading passages from non-fiction books, amounting to 24 hours of recordings. The deepfake audio data consists of 117,985 synthetic clips spanning 196 hours, generated to directly correspond to the real LJSpeech samples.

To evaluate model performance under varying conditions, two different dataset splits proposed in the WaveFake paper was considered:

1) In-distribution Setup: 80% of the MelGAN samples and 80% of LJSpeech are used for training, with the remaining 20% held out for testing. This results in a balanced 1:1 ratio of real to fake samples during training.
2) Out-of-distribution Setup: All deepfake samples except those from MelGAN, along with 80% of LJSpeech, are used for training. The test set remains the same 20% MelGAN and LJSpeech split. This represents a more challenging scenario with a 7.4:1 imbalance of real to fake samples in the training data.

B. *Model Architectures*

The following deep learning architectures for audio deepfake detection:

1) Gaussian Mixture Model (GMM): As a baseline, we train separate GMMs on MFCC features extracted from real and fake audio, using 128 mixture components each. Classification is performed by thresholding the log-likelihood ratio between the two models.
2) Vanilla Recurrent Neural Network (RNN): This sequential model processes input audio frame-by-frame, updating a hidden state representation that captures temporal dependencies. The final hidden state is passed through feed-forward layers to produce a binary prediction.
3) Bidirectional LSTM: Extending the vanilla RNN, we explore bidirectional long short-term memory (BiLSTM) networks which can better model long-range dependencies in both forward and backward directions of the input sequence.
4) Shallow Convolutional Neural Network: Inspired by previous work, we implement a lightweight CNN architecture that operates on 2D representations of the audio signal, either MFCCs or LFCCs. This allows the model to learn local discriminative patterns while maintaining some shift-invariance.
5) Time-Domain Synthetic Speech Detection (TSSD): The TSSD model is an end-to-end framework that takes raw audio waveforms as input, passing them through a series of 1D convolutions and ResNet-style modules before a final classification output.

For the recurrent and convolutional models, extract relevant input features like waveforms, MFCCs, or LFCCs from the raw audio data. Double delta coefficients are computed to augment MFCC/LFCC features with first and second order temporal derivatives.

C. *Training Procedure*

All deep learning models are implemented in PyTorch using the Adam optimizer with an initial learning rate of 0.0005 and weight decay of 0.0001. We use mini-batch sizes of 256 audio samples. The loss function is the binary cross-entropy loss between the predicted label scores and the ground truth labels. For the out-of-distribution setup where the training data is heavily imbalanced, we calculate a positive weight factor based on the real-to-fake ratio. This positive weight is then incorporated into the loss to act as if there was an equal number of real and fake samples during training.

We train each model separately on the in-distribution and out-of-distribution dataset splits. Validation performance is monitored, and the model checkpoint with the lowest validation loss is used for evaluating the test set.

D. *Evaluation Metrics*

To assess the performance of the models, several standard metrics were reported for audio deepfake detection benchmarks:

1) Equal Error Rate (EER): This measures the rate at which the false positive rate (falsely classifying real audio as fake) is equal to the false negative rate (falsely classifying fake audio as real). Lower EER values indicate better overall classification accuracy.
2) F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance across both classes. Higher F1 scores, ranging from 0 to 1, are preferable.
3) Area Under the Receiver Operating Characteristic Curve (ROC AUC): The ROC AUC quantifies the model's ability to distinguish between the real and fake classes across all possible decision thresholds. An AUC of 1 represents perfect separability, while 0.5 is equivalent to random chance.

By evaluating the models using these complementary metrics on both the in-distribution and out-of-distribution test sets, we can gain a comprehensive understanding of their detection capabilities under varying real-world conditions. The proposed model is shown in Fig 1. This systematic analysis allows us to identify strengths, limitations, and potential avenues for improvement in future audio deepfake detection research.

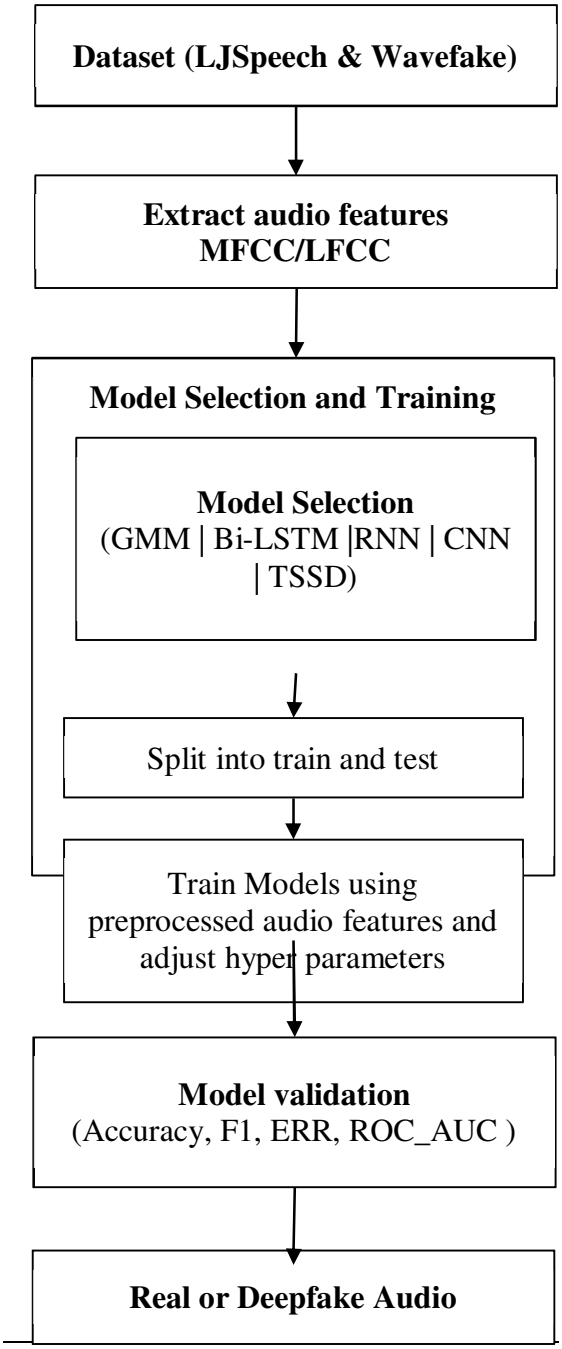


Fig1. Proposed Workflow of the model

IV. RESULTS

The performance of the various models on the audio deepfake detection task is summarized in Table 1 for the equal error rate (EER) metric and Table 2 for the F1 score and ROC AUC metrics.

Table 1: Equal Error Rates (EER) on the WaveFake dataset. Lower is better.

Model (input feature type)	In-dist Setup	Out-of-dist Setup
GMM (w/ LFCC) [10]	0.148	0.220
RawNet2 (w/ wave) [22]	0.001	**0.008**
VanillaRNN (w/ wave)	0.350	--
Bi-LSTM (w/ wave)	0.264	--
Bi-LSTM (w/ MFCC)	0.040	--
Bi-LSTM (w/ LFCC)	0.004	**0.044**
ShallowCNN (w/ MFCC)	0.004	--
ShallowCNN (w/ LFCC)	**0.000**	0.093
TSSD (w/ wave)	0.001	0.056

Table 2: F1 Scores and ROC AUC on the WaveFake dataset. Higher is better.

	In-dist Setup	
Model (input feature type)	F1-score	ROC AUC
VanillaRNN (w/ wave)	0.649	0.653
Bi-LSTM (w/ wave)	0.742	0.750
Bi-LSTM (w/ MFCC)	0.960	0.960
Bi-LSTM (w/ LFCC)	0.996	0.996
ShallowCNN (w/ MFCC)	0.997	0.997
ShallowCNN (w/ LFCC)	**1.000**	**1.000**
TSSD (w/ wave)	0.999	0.999

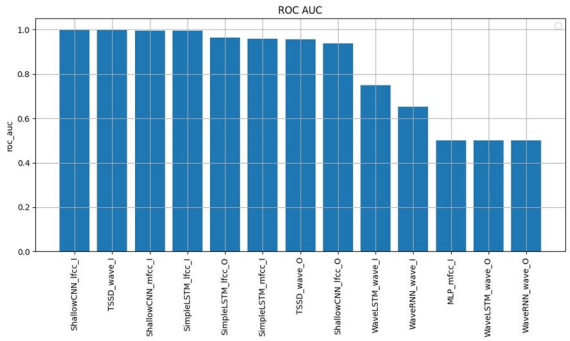


Fig 2. ROC Curve for different models

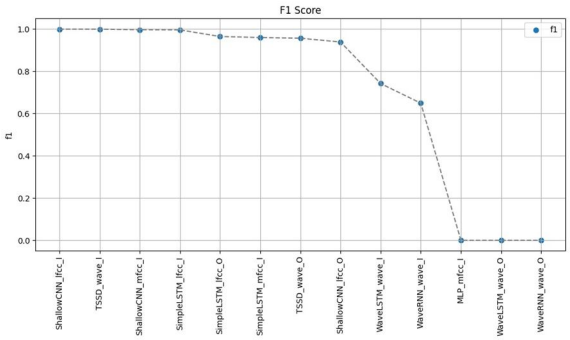


Fig 3. F1 SCore for different models

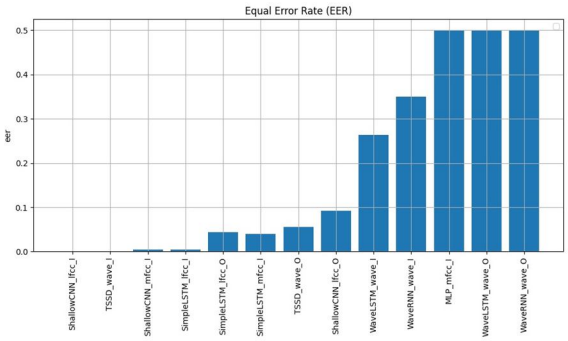


Fig 4. EER for different models

On the in-distribution test set, the shallow CNN model using LFCC features achieves the lowest EER of 0.000, along with a perfect F1 of 1.000 and AUC of 1.000, indicating excellent detection performance when the deepfake samples come from the same distribution as the training data. The TSSD model operating directly on raw waveforms also exhibits very strong in-distribution results, with an EER of 0.001, F1 of 0.999, and AUC of 0.999.

Performance degrades for all models on the more challenging out-of-distribution test set containing novel deepfake samples not seen during training. However, the BiLSTM with LFCC features, shallow CNN with LFCC, and TSSD maintain reasonable detection capabilities with EERs below 0.1. Across both setups, models using handcrafted LFCC features tend to outperform those using MFCCs or operating solely on raw waveforms, suggesting the importance of this input representation. The baseline GMM system struggles compared to the deep learning approaches, indicating the need for more powerful models to capture the complex patterns differentiating real and synthetic audio. Overall, these results demonstrate the effectiveness of deep learning techniques, particularly the shallow CNN and TSSD architectures, for detecting state-of-the-art audio deepfakes when there is a match between training and test data distributions. However, the out-of-distribution performance highlights remaining challenges in generalizing to unseen deepfake generation methods. Further analysis is required to enhance model robustness in real-world deployment scenarios.

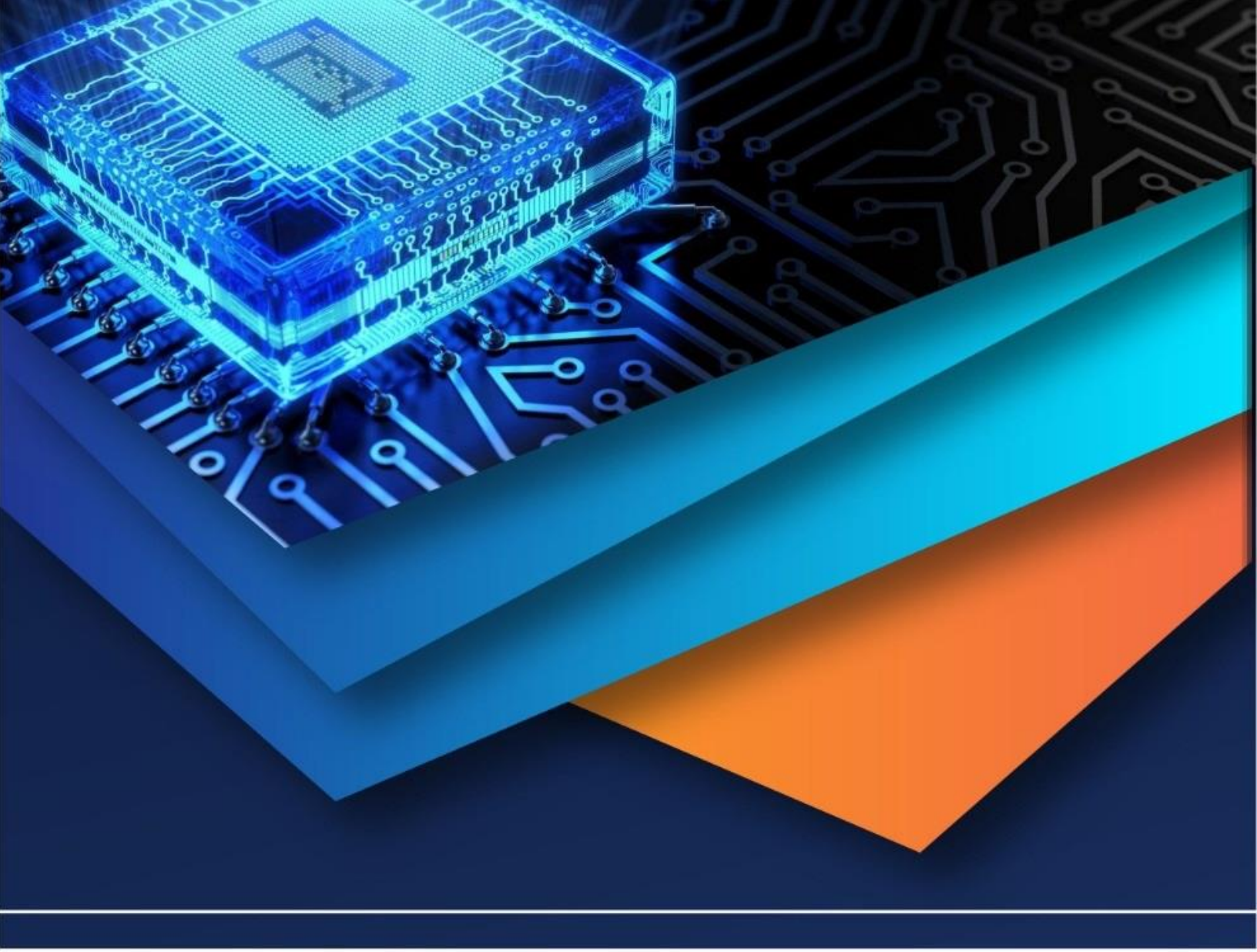
V. CONCLUSION

The research evaluated several deep learning models for detecting audio deepfakes on the challenging WaveFake dataset. While architectures like the shallow CNN and TSSD achieved strong in-distribution performance, all models struggled with out-of-distribution generalization to unseen deepfake generation methods. Key findings highlight the need for future work on transfer learning, ensemble methods, model interpretability, and more diverse benchmarks to enhance robustness against evolving deepfake threats. Reliable audio deepfake detection remains an open challenge requiring continued research efforts across scientific and societal domains to mitigate the risks posed by this technology.

REFERENCES

[1] Nishant Subramani and Delip Rao. Learning Efficient Representations for Fake Speech Detection. Proceedings of the AAAI Conference on Artificial Intelligence, 34:5859–5866, 04 2020.
[2] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, and Yang Liu. DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices, 2020.
[3] A. Lieto, D. Moro, F. Devoti, C. Parera, V. Lipari, P. Bestagini, and S. Tubaro. "Hello? Who Am I Talking to?" A Shallow CNN Approach for Human vs. Bot Speech Classification. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2577–2581, 2019.
[4] Haibin Wu, Heng-Cheng Kuo, Naijun Zheng, Kuo-Hsuan Hung, Hung-Yi Lee, Yu Tsao, Hsin-Min Wang, and Helen Meng. Partially Fake Audio Detection by Self-attention-based Fake Span Discovery, 2022.
[5] Jee-weon Jung, Hye-jin Shim, Hee-Soo Heo, and Ha-Jin Yu. Replay attack detection with complementary high-resolution information using end-to-end DNN for the ASVspoof 2019 Challenge, 2019.
[6] Xu Li, Na Li, Chao Weng, Xunying Liu, Dan Su, Dong Yu, and Helen Meng. Replay and synthetic speech detection with res2net architecture, 2020.
[7] Min Xu, Ling-Yu Duan, Jianfei Cai, Liang-Tien Chia, Changsheng Xu, and Qi Tian. HMM-based audio keyword generation. In Pacific-Rim Conference on Multimedia, pages 566–574. Springer, 2004.
[8] Keith Ito and Linda Johnson. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.
[9] Joel Frank and Lea Schönherr. WaveFake: A Data Set to Facilitate Audio Deepfake Detection. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021.
[10] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. Advances in neural information processing systems, 32, 2019.

[11] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbren-ner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio, 2016.
[12] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems, 33:17022–17033, 2020.
[13] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. Advances in neural information processing systems, 31, 2018.
[14] Shentong Mo, Haofan Wang, Pinxu Ren, and Ta-Chung Chi. Automatic Speech Verification Spoofing Detection. 12 2020.
[15] K. N. R. K. Raju Alluri and Anil Kumar Vuppala. IIIT-H Spoofing Countermeasures for Automatic Speaker Verification Spoofing and Countermeasures Challenge 2019. In Gernot Kubin and Zdravko Kacic, editors, Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019, pages 1043–1047. ISCA, 2019.
[16] Bhusan Chettri, Daniel Stoller, Veronica Morfi, Marco A. Martínez Ramírez, Emmanouil Benetos, and Bob L. Sturm. Ensemble Models for Spoofing Detection in Automatic Speaker Verification, 2019.
[17] Guang Hua, Andrew Beng Jin Teoh, and Haijian Zhang. Towards End-to-End Synthetic Speech Detection. IEEE Signal Processing Letters, 28:1265–1269, 2021.
[18] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.
[19] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-End anti-spoofing with RawNet2. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6369–6373, 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY