Grenze International Journal of Engineering and Technology, June Issue

Audio Forgery Alert: Uncovering Artificial Audio with Deepfake Detection

R. Bhavana[1], S. Kalaivani [2], B. Dharanidhar[3], K. Keerthi Reddy[4], K. Lokesh[5] and S.B. Keerthi Christopher[6]
[1-6]Madanapalle Institute of Technology & Science, Madanapalle, India
Email: bhavana.ramisetty2002@gmail.com, kalaivanis@mits.ac.in, dharanidhar224@gmail.com, kollikeerthireddy@gmail.com, lokesh95159@gmail.com, sbkeerthichristopher@gmail.com

***Abstract*—Deepfakes, the use of artificial intelligence to create fake media content, have become a big concern in recent days due to their potential misuse in various fields, like politics, entertainment, and finance. While deepfakes have primarily been associated with video manipulation, recent advances in audio manipulation techniques have made it possible to create highly realistic audio deepfake. This paper proposes a deep learning based approach for detecting audio deepfakes. Specifically, the paper introduce a novel architecture, called Audio Deepfake Detection Network (ADDN), which combines convolutional neural networks (CNNs) and attention mechanisms to detect deepfakes in audio recordings. Our approach is inspired by recent advances in image deepfake detection, which have shown that attention mechanisms can be effective in identifying subtle differences between real and fake images. This proposed approach has several implications for the field in deepfake detection. First, by combining CNNs and attention mechanisms, ADDN is able to learn both local and global features in audio recordings, making it more robust to various types of audio deepfakes. Second, ADDN is computationally efficient, making it suitable for real-time deepfake detection. The findings shows the promise of leveraging deep learning methodologies for detecting audio deepfakes, indicating significant areas for further exploration and advancement in this domain.**

***Index Terms*— Deepfakes, Audio Manipulation, Attention Mechanism, Features.**

I. INTRODUCTION

Deepfakes, a term coined by combining "deep learning" and "fake," have emerged as a significant phenomenon in digital media. They refer to artificially generated or morphed media created using techniques such as deep learning, aimed at producing content that appears realistic yet is manipulated, particularly in audio or video formats. In an era marked by rapid technological advancements, the advent of deepfakes poses a considerable threat to the authenticity of digital content. The potential misuse of deepfakes can have severe repercussions, including the spread of misinformation, fraudulent activities, and instances of blackmail. Hence, the development of robust and effective deepfake detection methods is crucial to uphold the integrity of digital content.

Audio deepfake detection involves the process of identifying and distinguishing synthetic or altered audio content from genuine recordings. This plays a vital role in preserving trust in audio-based media and mitigating the potential negative impacts associated with the misuse of deepfake technology.

Deepfake detection methods for audio can be broadly categorized into signal-based and machine learning-based

Grenze ID: 01.GIJET.10.2.596

approaches. Signal processing-based methods rely on analyzing spectral features of audio signals and detecting inconsistencies introduced during the deepfake generation process. On the other hand, machine learning-based models leverage various algorithms, including Support Vector Machines (SVM), Random Forests, or deep learning models, to learn patterns and characteristics of real and manipulated audio.

Our approach integrates cutting-edge techniques such as CNNs, attention mechanisms, and innovative normalization layers to enhance detection accuracy and system robustness. By utilizing ResBlock modules, our model can capture intricate patterns in audio signals, effectively distinguishing between authentic recordings and deepfake-generated content with high precision. Key components of our system include the AudioMiniEncoder module, which extracts essential features from spectrogram representations, and the AttentionBlock module, facilitating spatial positions to attend to each other, thereby improving the model's ability to detect anomalies indicative of deepfake manipulation. Additionally, the incorporation of QKVAttentionLegacy modules enhances the system's attention mechanism, allowing it to focus on relevant information while filtering out spurious signals.

II. Literature Survey

[1] Jiangyan Yi and Jianhua Tao conducted an extensive review on audio deepfake detection, emphasizing the distinctions between various types of audio deepfakes. Their survey covers competitions, datasets, features, classifications, and evaluation methods for cutting-edge detection techniques. They provide a detailed overview of existing datasets related to audio deepfake detection and conduct a comprehensive comparison of detection methods.

[2] Yinlin Guo proposed a novel audio deepfake detection method using the WavLM self-supervised model and Multi Fusion Attentive (MFA) Classifier. Their approach, trained on ASVspoof 2019 LA data and evaluated on ASVspoof 2019 LA, ASVspoof 2021 LA, and ASVspoof 2021 DF datasets, achieves state-of-the-art results on ASVspoof 2021 DF and competitive performance on ASVspoof 2019 and 2021 LA datasets. However, limitations include dataset diversity, generalization to various attacks, narrow evaluation metrics, insufficient analysis of failures, undiscussed preprocessing rationale, high computational resource dependency, and external dependency risks.

[3] Yuankun Xie and Haonan Cheng introduced an ASDG approach for audio deepfake detection to improve the ability to detect unfamiliar instances of manipulated speech. Their method incorporates an LCNN based feature generator for feature classification, single side domain learning to blur the line between authentic and altered speech across various domains, and the utilization of triplet loss to segregate fake speech distributions while amalgamating real speech distributions. The outcomes demonstrate that ASDG surpasses standard models in tasks spanning different domains, reducing the Equal Error Rate (EER) by up to 39.24% compared to RawNet2.

[4] Yuankun Xie presents W2V-ASDG, a robust Audio Deepfake Detection (ADD) model addressing cross domain limitations. Combining a self-supervised frontend (W2V2-XLS-R) and a domain generalization backbone (ASDG), the system learns a domain-invariant feature representation for real and fake speech. Utilizing diverse datasets for training and evaluation, including ASVspoof2019LA, WaveFake, FakeAVCeleb, IWA, ASVspoof2021DF, and FAD, the model achieves an impressive 4.60% Equal Error Rate (EER). However, limitations arise from a primarily English focused training dataset, potentially impacting generalization to other languages and emerging attack types.

[5] Zaynab M. Almutairi underscores the importance of audio deepfake detection within the field of forensics and introduces a novel approach named Arabic AD. This technique employs self-supervised learning methodologies to identify synthetic and mimicked voices. It involves the development of a synthetic dataset featuring a single speaker delivering Modern Standard Arabic with precision. Three experiments are conducted to compare the proposed method with well-known benchmarks, showing that Arabic AD performs better with low EER rates and high detection accuracy.

[6] Davide Salvi addresses deficiencies in current forensic methods for examining multimodal forgeries in videos and introduces a novel audio-visual deepfake dataset named TIMIT-TTS. This dataset comprises synthetic speech produced through Text to Speech (TTS) and Dynamic Time Warping (DTW) methods, serving as a resource for creating and evaluating forensic algorithms designed to identify multimodal deepfakes. The study emphasizes the necessity for multimodal forensic detectors and increased availability of multimodal deepfake data.

[7] Soniya Salman discusses challenges and solutions related to deepfake generation and detection, emphasizing the importance of detecting fake audio and video content. The paper provides an overview of deepfake detection

models and datasets, proposing a detailed classification of different types of deepfakes and presenting an extensive framework for deepfake generation and detection.
[8] Ousama A. Shaaban explores challenges posed by audio deepfakes and detection techniques, utilizing diverse datasets including Baidu, Mozilla TTS, and ASVspoof 2019. The study employs methodologies like speaker adaptation, multi-speaker sequence-to-sequence models, continuous learning, and ML models (SVM, RBF, XGBoost) for detecting imitation-based fakeness. Results show a PyTorch based system achieving 97% precision in detecting audio deepfakes.
[9] Ramubara conducts a comprehensive literature survey on deepfakes, focusing on their generation, detection, and impacts across visual, audio, and textual formats. The survey outlines the systematic approach used to collect and analyze relevant publications, emphasizing the need for a unified detection framework integrating visual, audio, and text detection methodologies, along with addressing the potential impacts of deepfakes in various domains.
[10] Lin Zhang and Xin Wang examine vulnerabilities in automatic speaker verification, including voice conversion, replay, tampering, adversarial assaults, and text-to-speech synthesis. They propose modifications to improve detection accuracy in a spoofing situation known as "Partial Spoof" (PS) and suggest improvements in feature extractors and segment labeling to detect and pinpoint short, generated spoof speech segments.

III. DATASET

The model has been trained on real audios and fake audios generated by the tortoise model, which is used for deepfake audio generation. When the model is performing at its best, the weights are saved and stored into a .pth file. These saved weights are loaded into the model for training instead of training it every time.

IV. PROPOSED WORK

The functioning of the Audio Deepfake Detection Network (ADDN) entails a sequence of steps and elements that collectively contribute to identifying deepfake audio.

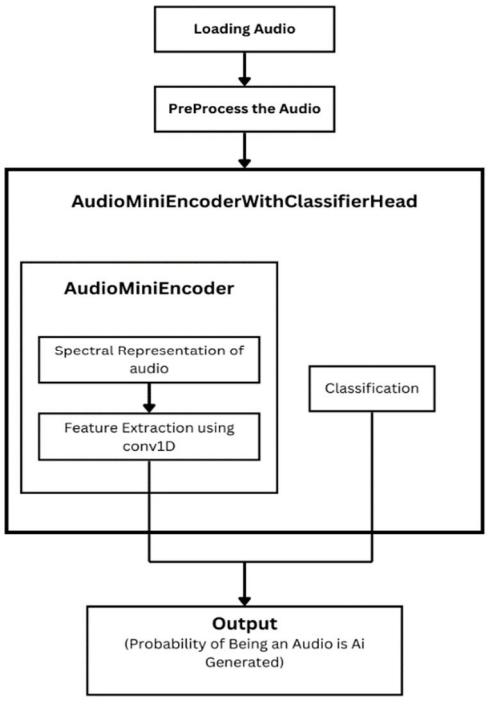

Fig 1 Architecture of the ADDN model

The model, depicted in Fig 1, initiates with data preparation, where audio spectrogram data undergoes processing and transformation utilizing modules like TorchMelSpectrogram to extract important features for the deep learning model. Through a sophisticated architecture combining convolutional neural networks (CNNs) and attention mechanisms, the model effectively discriminates between real and deepfake audio recordings with high degree of accuracy and precision.

This model consists of several modules and functions that play crucial roles in the detection process.

- GroupNorm32 Module: This module extends nn.GroupNorm to handle normalization in the model, ensuring proper group normalization for input channels.
- QKVAttentionLegacy Module: Responsible for performing QKV attention, this module applies attention mechanisms to input Qs, Ks, and Vs tensors to calculate attention weighted outputs.
- AttentionBlock Module: The AttentionBlock Module functions as a mechanism enabling spatial positions to interact with each other, thereby improving the model's capability to identify intricate patterns within the audio data. Within the AttentionBlock, the Softmax activation function is employed to calculate attention weights, enabling the model to prioritize relevant segments of the input data during processing. Through the utilization of SoftMax, the model effectively assigns significance to various elements within the input data, facilitating efficient feature extraction and pattern recognition.
- Upsample and Downsample Modules: These modules facilitate upsampling and downsampling operations within the network, allowing for spatial adjustments in the audio data representation.
- ResBlock Module: The ResBlock module is designed to handle residual connections within the model, allowing for the propagation of gradients and aiding in training deep networks effectively. It consists of convolutional layers, normalization, activation functions, and dropout to enhance feature extraction and learning capabilities. The activation functions used ResBlock is SiLU(Sigmoid Linear Unit)

  $$SiLU: f(x)=x.\sigma(x)$$

  where $\sigma(x)$ is the sigmoid function. It introduces a non-linearity to the model by applying a smooth curve to the input data, enhancing the model's ability to capture intricate patterns and correlations present within the audio spectrogram data.The SiLU function is preferred for its smooth gradient behavior, which aids in stable training and convergence of deep learning models.
- CheckpointedLayer Module: CheckpointedLayer module facilitates efficient memory usage during training by checkpointing certain computations to reduce memory consumption. It wraps a given module and selectively applies checkpointing based on specified arguments, optimizing memory usage while maintaining computational accuracy.
- AudioMiniEncoderModule:The AudioMiniEncoder module serves as the core component responsible for encoding audio spectrogram data into a lower dimensional embedding space .It utilizes convolutional layers, residual blocks, downsampling, and attention mechanisms to extract essential features from audio signals while reducing dimensionality.The module incorporates multiple layers of ResBlocks and Downsample layers to capture intricate patterns in the audio data and enhance representation learning.
- AudioMiniEncoderWithClassifierHead:The AudioMiniEncoderWithClassifierHead module extends the functionality of the AudioMiniEncoder by adding a linear classifier head for multi class classification tasks. It incorporates the encoder's output into a linear layer to predict class labels based on the extracted audio embeddings. Again softmax activation function is used inside this module. The softmax function, typically employed in neural networks for classification tasks especially multi-class classification, serves to normalize the network's output into a probability distribution across various classes. However, in this instance, it is applied to perform binary classification.

Crucial elements of the model includes an Audio MiniEncoder module, responsible for extracting vital features from spectrogram representations of audio data, along with an AttentionBlock module that allows the model to concentrate on pertinent information while disregarding irrelevant signals. Moreover, the inclusion of ResBlock modules amplifies the network's capacity to capture complex patterns and variations within audio content, thereby enhancing its detection capabilities. Additionally, the model incorporates innovative techniques like relative position embeddings and group normalization, aimed at boosting performance and resilience against adversarial attacks. These advancements significantly contribute to the overall efficacy of the detection system, enabling it to adapt to various methods of deepfake generation and uphold accuracy across diverse scenarios.

V. RESULT

This ADDN is able to distinguish between the real and deepfake audios correctly by collecting the spectrograms and other features using attention mechanisms and ResBlocks and give the probability of the audio being AI

generated accurately,using the saved and loaded weights for training. This ADDN is compared with several baseline models, along with the traditional algorithms from machine learning and deep learning models.

VI. CONCLUSION

A deep learning based approach for detecting audio deepfakes is presented. Our proposed architecture, ADDN, combines CNNs and attention mechanisms to identify deepfakes in audio recordings with high accuracy. The proposed model can be used to detect audio deepfakes in various applications, such as voice recognition and security systems. Future work includes extending ADDN to other domains, such as video deepfake detection, and exploring other attention mechanisms, integrating multi modal information like audio and visual data to create a more comprehensive deepfake detection system that can handle various types of deepfake content, using transformer based architectures to further improve the performance.

REFERENCES

[1] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Zhang, Yan Zhao, "Audio Deepfake Detection: A Comprehensive Review", Journal of Audiovisual Forensics, 2023.

[2] Yinlin Guo, Haofan Huang, Xi Chen, He Zhao, Yuehai Wang, "Enhancing Audio Deepfake Detection using Self-Supervised WavLM and MultiFusion Attentive Classifier", Journal of Signal Processing and Machine Learning, vol. 1, 2024.

[3] Y. Xie, H. Cheng, Y. Wang, L. Ye, "Domain Generalization via Aggregation and Separation for Audio Deepfake Detection," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 344-358, 2024. DOI: 10.1109/TIFS.2023.3324724.

[4] Xie, Yuankun, Haonan Cheng, Yutian Wang, Long Ye, "Learning a Self-Supervised Domain-Invariant Feature Representation for Generalized Audio Deepfake Detection", Proceedings of the International Speech Communication Association, 2023, pp. 2808-2812. DOI: 10.21437/Interspeech.2023-1383.

[5] Z. M. Almutairi and H. Elgibreen, "Detecting Synthetic Audio of Arabic Speakers Using Self-Supervised Deep Learning," in IEEE Access, vol.11, pp.72134-72147, 2023.

[6] DOI: 10.1109/ACCESS.2023.3286864.

[7] D. Salvi, B. Hosler, P. Bestagini, M. C. Stamm, and S. Tubaro, "TIMIT TTS: A Text-to-Speech Dataset for Multimodal Synthetic Media Detection," in IEEE Access, vol. 11, pp. 50851-50866, 2023. DOI: 10.1109/ACCESS.2023.3276480.

[8] S. Salman, J. A. Shamsi, and R. Qureshi, "Deep Fake Generation and Detection: Challenges, Solutions, and Implications," in IT Professional, vol. 25, no. 1, pp. 52-59, Jan.-Feb. 2023. DOI: 10.1109/MITP.2022.3230353.

[9] O. A. Shaaban, R. Yildirim, and A. A. Alguttar, "Approaches for Audio Deepfake Generation and Detection," IEEE Access, vol. 11, pp. 132652-132682, 2023. DOI: 10.1109/ACCESS.2023.3333866.

[10] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa Dutse, S. Khan, and S. Parkinson, "Detection and Impact Assessment of Deepfakes in Visual, Audio, and Text Formats: A Comprehensive Survey," in IEEE Access, vol. 11, pp. 144497-144529, 2023. DOI: 10.1109/ACCESS.2023.3344653.

[11] Xie, Yuankun, Haonan Cheng, Yutian Wang, Long Ye, "Learning a Self-Supervised Domain-Invariant Feature Representation for Generalized Audio Deepfake Detection," Proceedings of the International Speech Communication Association, 2023, pp. 2808-2812. DOI: 10.21437/Interspeech.2023-1383.

[12] D. U. Leonzio, L. Cuccovillo, P. Bestagini, M. Marcon, P. Aichroth, S. Tubaro, "Detection and Localization of Audio Splicing Based on Device Traces," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 4157-4172, 2023. DOI: 10.1109/TIFS.2023.3293415.