International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Advancements in Audio Tampering and Forgery Detection Techniques

Harshith Chandrashekar, Aalisha Nishath Yousuf, Anjali V, Impana M

Department of Computer Science and Engineering in Artificial Intelligence and Machine Learning, Vidyavardhaka
College of Engineering Mysuru, India.

Abstract: Audio has become a crucial part of communication, legal proceedings, digital documentation etc. Maintaining authenticity of audio recordings is an acute concern in this digital age. Audio forensics deals with acquisition, analysis and authentication of audio recordings. This plays a critical part in evidence presented in legal or investigative contexts. Detecting tampered audio has become extremely difficult with the availability of advanced audio editing tools and deep generative models. This literature review scrutinizes the advancement and current status of the techniques developed for detecting and localizing tampered audio. It deals with an extensive analysis of techniques based on Electrical Network Frequency (ENF), environmental and device-specific signatures, and signal processing to determine various types of tampering. Various machine learning and deep learning techniques are also used to assess audio signals. They detect defects in time-domain, frequency-domain features and temporal patterns present. Metadata and compression artifact analysis examines anomalies in audio file structure, codec parameters and compression histories. The review lays emphasis on the strengths and weaknesses of existing frameworks in audio forensics. This work encourages development of effective and reliable techniques for audio manipulation detection.

Keywords: Audio forensics, Audio tampering detection, ENF analysis, Deepfake audio, Spectrogram-based features, Deep learning, Audio manipulation localization, Forensic signal analysis.

I. INTRODUCTION

Speech delivers semantic content with speaker identity and purpose. Audio recordings have become crucial in law enforcement, judicial proceedings, daily interaction and journalism. Enhancement in deepfake technologies and audio editing software has empowered manipulation of audio recordings. These manipulations change the original meaning, fabricate completely false narratives, or impersonate individuals posing serious threats to security, privacy and integrity of judicial investigations. Audio forensics is a subfield of digital forensics that assesses the origin and authenticity of audio recordings. Main focus of traditional audio forensic methods is on detecting noise patterns, inconsistencies in pitch or disparities in background environments. These techniques though useful, are increasingly inadequate against modern tampering strategies especially near- perfect forgeries that imitate speech patterns, tone, and emotion. This literature review focuses on methods for detecting digital audio tampering in the context of forensic investigations. In forensic investigations it is crucial to have advanced and sophisticated technology driven tools which are proficient in automatically identifying such manipulations in difficult atmospheres. There is a critical need of automated, machine learning–based approaches that examines temporal, spectral and contextual features of the audio signal. These techniques are based on electrical network frequency (ENF) signal tracking, spectrogram analysis, multimodal cross-verification, shallow and deep feature fusion etc. The purpose of this review is to evaluate existing methodologies and thereby recognizing gaps in the literature. It extends to comparative analysis, identifying strengths and limitations with respect to forensic applicability, computational efficiency, robustness and accuracy.

II. LITERATURE REVIEW

[1] The research paper titled “Audio Splicing Detection and Localization Using Environmental Signature”, authored by Hong Zhao, Yifan Chen, Rui Wang, and Hafiz Malik, was published on arXiv in November 2014. It focusses audio splicing which is a tampering technique where speech segments are transcribed from multiple devices and integrated into one resulting in false content. The proposed methodology estimates the environmental signature of the recording environment. The incompatible frames representing splicing are recognized by estimating the normalized cross-correlation (NCC). It is system-independent, provides high accuracy and remains effective even under MP3 compression. Its flaws include dependency on unique

environmental differences which may lead to detection failure. The method was evaluated on TIMIT speech corpus and another dataset transcribed in four real acoustic environments.

[2] The research paper "Audio Splicing Detection and Localization Based on Acquisition Device Traces" authored by Daniele Ugo Leonzio, Luca Cuccovillo, Paolo Bestagini, Marco Marcon, Patrick Aichroth, and Stefano Tubaro. It was published in IEEE Transactions on Information Forensics and Security, Volume 18 (2023). It focuses on audio splicing. The methodology extracts device-specific features using Convolutional Neural Network (CNN). The system enforces a clustering algorithm to combine frames from the same device and deciding boundaries between clusters. A distance- based measure was used to filter the temporal localization of the splicing events. It achieves high accuracy when detecting splices and localizing their positions. The outcomes confirmed its effectiveness in real world forensic scenarios. The flaws comprise of failure if spliced segments are from the same device model or pipeline and the necessity of using labelled data to train CNN.

[3] The research paper titled "Exposing Speech Resampling Manipulation by Local Texture Analysis on Spectrogram Images" was authored by Yujin Zhang, Shuxian Dai, Wanqing Song, Lijun Zhang, and Dongmei Li. It was published in 2019 in the journal Electronics. It focuses on resampling in which the audio is altered by changing the sampling rate. This methodology incorporates spectrograms of speech signals using Short-Time Fourier Transform (STFT). The Local Binary Pattern (LBP) operator records local structural changes introduced by resampling. Support Vector Machine (SVM) classifier is used to transform the original speech into resampled speech. It records local texture patterns constant to global transformations. It is interpretable and easy to implement. The flaws include reliability on the transparency and resolution of the spectrogram. The classifier manifests high accuracy differentiating between resampled and non-resampled speech.

[4] The paper "Pyramid Feature Attention Network for Speech Resampling Detection" was authored by Xinyu Zhou, Yujin Zhang, Yongqi Wang, Jin Tian, and Shaolun Xu. It was published in June 2024 in Applied Sciences (Switzerland). This study focuses on issues of detecting speech resampling. The proposed methodology reckons the log-spectrogram which would be fed into a Feature Pyramid Network (FPN) to raise multi-scale feature maps. Finally, prepared feature maps are classified by a standard classifier. It is capable of capturing multi-scale spectro-temporal patterns. The deep learning approach is robust to MP3 compression. The limitations include computational complexity and spectrogram representations that degrade if the resampling artifacts are subtle or heavily. The model enhanced accuracy on a resampling corpus acquired from the TIMIT speech dataset.

[5] The research paper titled "Audio Tampering Forensics Based on Representation Learning of ENF Phase Sequence" which was authored by Chunyan Zeng, Yao Yang, Zhifeng Wang, Shuai Kong, and Shixiong Feng, and published in 2022 in the International Journal of Digital Crime and Forensics. It emphasizes on passive audio authentication using the electrical network frequency (ENF) phase sequence. Initially ENF component is removed from audio using digital signal processing and its phase sequence is calculated through discrete Fourier transform. A bidirectional Long Short-Term Memory (BiLSTM) architecture models consecutive patterns in ENF behaviour. The approach is device-independent and content-agnostic. The flaws include susceptibility to the strength and quality of ENF present. This method outperforms operating ENF-based detectors on benchmark datasets in terms of accuracy and robustness.

[6] The research paper titled "Shallow and Deep Feature Fusion for Digital Audio Tampering Detection", authored by Zhifeng Wang, Yao Yang, Chunyan Zeng, Shuai Kong, Shixiong Feng, and Nan Zhao appeared in the EURASIP Journal on Advances in Signal Processing on August 13, 2022. It focuses on the limitations of existing ENF-based and machine learning based audio tampering detection methods. The methodology consists of a band-pass filter around 50/60 Hz to isolate the ENF component. In parallel, these framed sequences are fed into a CNN to extract deep features and capture local ENF variations. The deep and shallow features are fused using a sigmoid-based attention mechanism. The fused representation is fed to a DNN classifier for binary classification to check tampering. The approach has increased architectural complexity. The method achieved 97.03% accuracy on the Classical set and 88.31% accuracy on GAUDI-DI in contrast to four baseline methods.

[7] The research paper titled "Digital Audio Tampering Detection Based on Deep Temporal- Spatial Features of Electrical Network Frequency" was authored by Chunyan Zeng, Shuai Kong, Zhifeng Wang, Xiangkui Wan, Yunfan Chen and published in April 2023 in the journal Information (MDPI), Volume 14, Issue 5. The authors focus on Electrical Network Frequency (ENF) to verify the authenticity of an audio file. Authors proposed a dual-path deep learning framework that learns temporal and spatial features from ENF phase sequences. It has two main components: a Residual Dense Temporal Convolutional Network (RDTCN) and a Convolutional Neural Network (CNN). The RDTCN is used for extracting temporal dependencies in ENF phase data. The CNN is used for capturing spatial features from structured ENF matrices. Later these two components are integrated using an attention-based feature fusion module. The

framework is entirely trained using labelled datasets. The drawbacks are dependent on quality and presence of ENF signals within the audio. The model achieved high accuracy and precision on three datasets - Carioca, New Spanish, and ENF_Audio.

[8] The research paper titled "Digital Audio Tampering Detection Based on ENF Spatio-temporal Features Representation Learning" authored by Chunyan Zeng, Shuai Kong, Zhifeng Wang, Xiangkui Wan, and Yunfan Chen published in 2024 in Multimedia Tools and Applications. It highlights issues of ENF-based tampering detection. The proposed methodology includes removing a high-precision ENF phase sequence from audio by utilizing discrete Fourier analysis. A Convolutional Neural Network (CNN) extracts deep spatial features whereas bidirectional LSTM (BiLSTM) models extract temporal dependencies. Lastly, Multi-Layer Perceptron (MLP) classifier assesses if the audio has been altered. The detection accuracy is enhanced by 2.12%–7.12% over ENF-based methods on New Spanish, Carioca 1 and 2 databases. The setbacks include necessity of strong and clean ENF signal in the recording. It is an innovative deep learning framework that utilizes the temporal dynamics and spatial features of ENF phase sequences for robust digital audio tampering detection.

[9] Research paper titled "Audio Forgery Detection and Localization with Super-Resolution Spectrogram and Keypoint-Based Clustering Approach" authored by Beste Üstübioğlu, Gül Tahaoğlu, Güzin Ulutaş, Arda Üstübioğlu, and Muhammed Kılıç. It was published in the Journal of Supercomputing in June 2023. The work emphasized on audio copy-move forgery which is an audio manipulation technique. The audio file segment is cloned or removed to change the ambience of the file. To overcome the issue, a robust framework that uses Binary Robust Independent Elementary Features (BRIEF) was introduced. Next, the Ordering Points To Identify the Clustering Structure (OPTICS) clustering algorithm groups identical keypoint descriptors. Some flaws are that it uses high-resolution data and clustering algorithms resulting in high computational complexity. Detection efficiency is sensitive to parameter tuning. The method shows high performance across multiple datasets with precision, recall, accuracy and F1-score.

[10] The paper titled "Identification of Fake Stereo Audio", authored by Tianyun Liu and Diqun Yan, was first published as an arXiv preprint on April 20, 2021. It addresses the problem of differentiating between genuine and artificially generated stereo derived from mono sources. The methodology extracts 80-dimensional Mel Frequency Cepstral Coefficients (MFCC) features from stereo audio. These features are then used in two classification schemes: an SVM classifier with MFCC features and a lightweight CNN with three fully connected layers. The use of MFCCs allows for robust channel- specific spectral analysis. The SVM offers computational efficiency and CNN captures more complex patterns. However, the limitations include sensitivity to the feature extraction parameters. The classifiers achieved reliable detection accuracy and strong robustness under varying conditions on three independent datasets, each with five different cutoff frequencies. The paper provides a foundational forensic approach to detect fake stereo audio.

[11] The paper titled "Audio Deepfake Detection Based on a Combination of F_0 Information and Real Plus Imaginary Spectrogram Features" was authored by Jun Xue, Cunhang Fan, Zhao Lv, Jianhua Tao, Jiangyan Yi, Chengshi Zheng, Zhengqi Wen, Minmin Yuan, and Shegang Shao, presented as part of the 1st International Workshop on Deepfake Detection for Audio Multimedia held at ACM Multimedia in October 2022. It focuses on the issue of audio deepfake which utilizes artificial intelligence to imitate a person's voice. The methodology comprises of selective representations extracted from low-frequency F_0 band succeeded by higher-frequency real spectrogram and complementary imaginary spectrogram data. A two- stage fusion scheme unites intra-stream modelling and inter-stream fusion to form the final prediction. The system is evaluated on the ASVspoof 2019 Logical Access (LA) dataset with Equal Error Rate (EER) of 0.43%. Its flaws include increased complexity due to handling multiple subband streams and the need for adjusting fusion and modelling parameters.

[12] Research paper titled "Deepfake Audio Detection by Speaker Verification" was authored by Alessandro Pianese, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva, and published in September 2022 as part of the IEEE International Workshop on Information Forensics and Security (WIFS). It focuses on evaluative constraints in audio deepfake detectors. This methodology uses speaker verification models to extract embeddings from audio samples. Two decision strategies Centroid-Based (CB) and Maximum-Similarity (MS) are used. The sample is categorized as hoax if the resemblance falls below a pre-defined point. The detector is universalized and robust to unseen deepfake generation methods. Its limitations include dependency on the availability of reference recordings of the target identity. It is flexible across formidable audio conditions. Results were consistently condescending under the generalization-focused evaluation. This proposal is more acceptable for situations where the target speaker is known and pre-enrolled.

[13] The research paper titled "Fake Speech Detection Using VGGish with Attention Block" authored by Tahira Kanwal, Rabbia Mahum, Abdul Malik AlSalman, Mohamed Sharaf, and Haseeb Hassan. It was published in June 2024 in the

EURASIP Journal on Audio, Speech and Music Processing. The existing fake speech detection models struggle to generalize across various types of fake speech. First, the audio samples are converted into Mel-spectrograms, which are then fed into the VGGish model. This model incorporates Convolutional Block Attention Module (CBAM) which is an attention mechanism. The resulting features train a classification model that is capable of identifying fake speech. The use of CBAM enhances the feature representation of the model. The limitations include the model's sensitivity to preprocessing steps and high computational costs. The model was evaluated on the ASV spoof 2019 dataset and achieved an Equal Error Rate (EER) of 0.07% for LA and 0.52% for PA, with a detection accuracy of 99.78%, precision of 98.86%, and recall of 99.94%. These results remarkably outperformed the baseline models on the dataset.

[14] Research paper titled "Deepfake Audio Detection Using Spectrogram based Feature and Ensemble of Deep Learning Models" by Lam Pham, Phat Lam, Truong Nguyen, Huyên Nguyễn, and Alexander Schindler was published in September 2024 at the 5th IEEE International Symposium on the Internet of Sounds. The methodology consists of three deep-learning approaches. First, the transfer learning is done from established computer-vision models and audio pre-trained models. The top- performing models are fused into an ensemble. The system encapsulates multi-perspective frequency– time representations. Multiple model types ensure complementarity and robustness. The system reduces of individual weaknesses while increasing their individual strengths. It introduces higher computational load due to multiple models and spectrogram variants. Experimental evaluation on the ASVspoof 2019 benchmark dataset shows that the model achieved an extremely low Equal Error Rate (EER) of 0.03. The paper introduces a highly effective, ensemble-based framework for audio deepfake detection. The figures Fig.1 and Fig.2 show relative time and accuracy of each technique respectively.

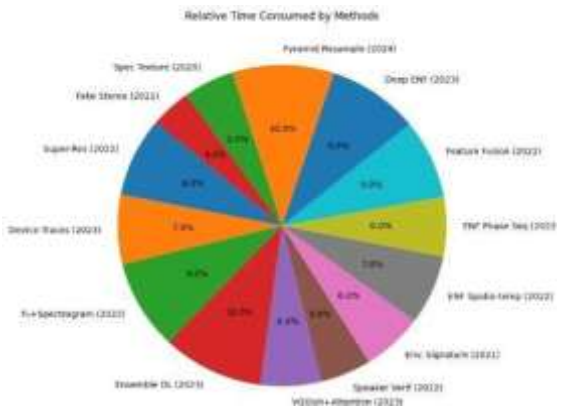

Fig.1: Relative time consumed of each technique

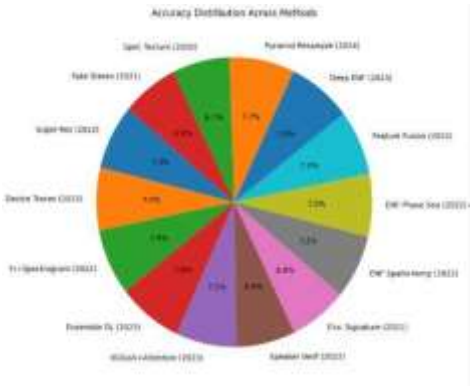

Fig.2: Relative accuracy of each technique

III. RESEARCH GAPS

After reviewing the listed papers, the following limitations and gaps are observed in current research:

1. **Over-reliance on Specific Features:** Many methods focus only on one type of feature. This leads to limited generalization when the forgery does not involve those specific artifacts or the audio is captured in varying environmental or device conditions.
2. **Vulnerability to Adversarial Attacks:** Deep learning models such as those relying on spectrograms are susceptible to adversarial manipulations. In this, the attackers can intentionally smooth out traces or noise.
3. **Poor Performance in Noisy or Real- World Scenarios:** Most of the models work well in clean or controlled datasets. However, they fail in real forensic cases, wherein background noise, compression artifacts, and recording devices vary considerably.
4. **Insufficient Handling of Multi-source or Composite Forgeries:** Audio splicing, deepfake synthesis and copy-move techniques merge content from different sources. Current models are inadequate to accurately localize tampered segments or differentiate layered manipulations.
5. **Dependence on Supervised Learning with Labelled Data:** Numerous models need large labelled datasets for training. It makes them hard to adapt to unseen or rare manipulation scenarios.
6. **Limited Explainability:** Deep neural networks like attention blocks or ensemble methods often lack interpretability. Hence, forensic experts are hesitant to rely on them for legal cases.
7. **Device and Environment Dependency:** Approaches based on device traces or environmental cues are prone to fail when metadata is stripped or when devices are varied or unknown.

IV. PROPOSED METHODOLOGY

The proposed methodology suggests a comprehensive and adaptable hybrid framework for audio tampering detection. The system is designed as a modular architecture capable of detecting multiple manipulation types within a single unified model. It also provides fine-grained localization of the tampering. This approach eliminates the need for separate tools and improves detection efficiency. The model leverages feature-agnostic deep representations that are supported by handcrafted features. This combination allows the system to capture both subtle signal anomalies and domain-specific characteristics. This methodology embodies advanced preprocessing and data augmentation strategies to ensure robustness. These techniques prepare the model to perform reliably across a wide range of environments, devices, and recording conditions. A critical aspect of the framework is the development of large, diverse, and well-annotated datasets that include multiple manipulation types, tampering locations, and realistic environmental factors. These datasets will improve model training, generalization and serve as benchmarks for forensic research. The methodology also explores multimodal and contextual forensics by integrating audio transcripts, speaker verification, and metadata such as device information and timestamps. Cross-checking the sources helps in detecting semantic inconsistencies and enhances the accuracy of tampering detection. Explainability tools are integrated for transparency and trustworthiness. These tools allow forensic analysts and legal authorities to understand the motive behind the model's predictions. Hence, the findings more interpretable and admissible in court. The framework integrates adversarial training. This exposes the model to synthetic attacks and ensemble defence methods. This enhances its resilience to new and unanticipated tampering techniques. Overall, this methodology offers a scalable, interpretable, and resilient audio forensic solution that can detect and localize multiple types of tampering while withstanding real-world challenges. This makes it uniquely suited to support forensic investigations and ensure the credibility of audio evidence in legal contexts.

Uniqueness of the proposed methodology is as follows:

1. **Hybrid Framework:** It uses both handcrafted acoustic features and deep learning to improve detection accuracy. This combines traditional audio features with deep learning models, resulting in better detection performance than using either approach alone.
2. **Noise & Condition Robustness:** The model is designed such that it works efficiently even when the audio has compression artifacts, background noise, or comes from different devices which makes it dependable in real-world situations.
3. **Manipulation-Type Classification:** It not only detects tampering but also identifies the specific type of alteration done, providing more detailed information for investigations.
4. **Enhanced Analysis:** It leverages temporal, spectral and phase patterns to boost sensitivity. The system focusses at different aspects of the audio. This more accurately identifies anomalies.
5. **Dataset Contribution:** It shares a labelled tampered audio dataset which includes examples of manipulated audio to support future forensic research.
6. **Multi-level Feature Fusion:** It combines environmental signatures, shallow acoustic features and deep representations to get complete understanding of potential tampering.
7. **Self-supervised and Semi-supervised Learning:** The model trains itself using unlabeled audio recordings by learning what normal patterns sound like, allowing it to detect irregularities without needing a large set of labelled examples.

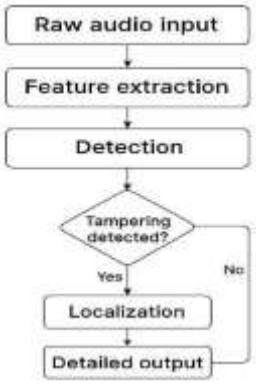

Fig 3: A general flowchart of the proposed methodology

V. CONCLUSION

This literature review explores various types of existing audio tampering detection techniques. Thus, tracing the budge from traditional signal processing methods to modern deep learning and multimodal approaches. Early methods relied on handcrafted features such as noise inconsistencies, pitch anomalies and Electrical Network Frequency (ENF) traces. With the advancement of machine learning hybrid models to combining shallow features with learned representations improves accuracy in complex manipulations like splicing and cloning. It took a major lunge in adopting deep learning where in the end-to-end architectures learned biased features directly from spectrograms or raw waveforms. These techniques refined performance in detecting minute manipulations. Recent work has relocated towards multimodal forensics. Cross-modal cues are leveraged with lip- sync inconsistencies or stereo spatial artifacts for improved robustness. Representation learning has

also gained importance. There is a gradual shift from handcrafted audio features to deep learned and multimodal representations. Each stage has bestowed to more accurate, interpretable and robust audio forensic tools for real-world investigations.

VI. REFRENCES

1. Zhao, H., Chen, Y., Wang, R., & Malik, H. (2016). Audio Splicing Detection and Localization Using Environmental Signature. Multimedia Tools and Applications.

2. (Assumed) Leonzio, D. U., Cuccovillo, L., Bestagini, P., Marcon, M., Aichroth, P., & Tubaro, S. (2023). Audio Splicing Detection and Localization Based on Acquisition Device Traces. IEEE Transactions on Information Forensics and Security, 18, 4157–4172.

3. Zhang, Y., Dai, S., Song, W., Zhang, L., & Li, D. (2019). Exposing speech resampling manipulation by local texture analysis on spectrogram images. Electronics, 9(1), 23.

4. Zhou, X., Zhang, Y., Wang, Y., Tian, J., & Xu, S. (2024). Pyramid Feature Attention Network for Speech Resampling Detection. Applied Sciences, 14(11), 4803.

5. Zeng, C., Yang, Y., Wang, Z., Kong, S., Feng, S., & Zhao, N. (2022). Audio tampering forensics based on representation learning of ENF phase sequence. International Journal of Digital Crime and Forensics, 14(1), 1–19.

6. Wang, Z., Yang, Y., Zeng, C., Kong, S., & Feng, S. (2022). Shallow and deep feature fusion for digital audio tampering detection. EURASIP Journal on Advances in Signal Processing, 2022(1), 69.

7. Zeng, C., Kong, S., Wang, Z., Li, K., & Zhao, Y. (2023). Digital Audio Tampering Detection Based on Deep Temporal-Spatial Features of Electrical Network Frequency. Information, 14(5), 253.

8. Zeng, C., Kong, S., Wang, Z., Wan, X., & Chen, Y. (2022). Digital Audio Tampering Detection Based on ENF Spatio-temporal Features Representation Learning. CoRR, abs/2208.11920

9. Ustübioğlu, B., Tahaoğlu, G., Ulutaş, G., Üstübioğlu, A., & Kılıç, M. (2024). Audio forgery detection and localization with super-resolution spectrogram and keypoint-based clustering approach. Journal of Supercomputing, 80(1), 486–518.

10. Liu, T., & Yan, D. (2021). Identification of fake stereo audio. Information, 12(7), 263.

11. Xue, J., Fan, C., Lv, Z., Tao, J., Yi, J., Zheng, C., Wen, Z., Yuan, M., & Shao, S. (2022, October). Audio Deepfake Detection Based on a Combination of F_0 Information and Real Plus Imaginary Spectrogram Features. In Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Lisbon, Portugal.

12. Pianese, A., Cozzolino, D., Poggi, G., & Verdoliva, L. (2022, December). Deepfake Audio Detection by Speaker Verification. In IEEE International Workshop on Information Forensics and Security (WIFS).

13. Kanwal, T., Mahum, R., AlSalman, A. M., Sharaf, M., & Hassan, H. (2024). Fake Speech Detection Using VGGish with Attention Block. EURASIP Journal on Audio, Speech, and Music Processing, 2024(1), 35.

14. Pham, L., Lam, P., Nguyen, T., Nguyễn, H., & Schindler, A. (2024). Deepfake Audio Detection Using Spectrogram-based Feature and Ensemble of Deep Learning Models. In Proceedings of the IEEE 5th International Symposium on the Internet of Sounds (IS2).