SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V12I3P104

Volume 12 Issue 3, 36-44, March 2025

Review article

Deep Learning-Based Techniques for Identification of Audio DeepFake with Open Issues: A Meta-Analysis

Krity Duhan[1], Abhishek Kajal[2]

[1,2]*Department of Computer Science and Engineering, GJUS&T, Hisar, Haryana, India.*

[1]*Corresponding Author : krityduhan7@gmail.com*

Received: 21 October 2024 Revised: 24 December 2024 Accepted: 25 February 2025 Published: 31 March 2025

Abstract *- Amidst the ongoing advancements in artificial intelligence, a particularly fascinating and alarming progress is the rise of deepfake speech. The emergence of deepfake technology poses a substantial risk to national security, democratic systems, society as a whole and individual privacy. Consequently, it is imperative to create better methods for identifying and mitigating possible deepfake threats. Audio counterfeiting detection is a rapidly developing and important subject. A growing amount of literature is focused on studying deepfake detection computations, which have demonstrated successful results. However, it is important to note that the issue is still far from being completely settled. As synthetic voice generation technology improves, audio deepfake is growing as perhaps the most widespread form of deception. Therefore, the task of differentiating between counterfeit and authentic audio recordings is growing increasingly difficult. Hence, the significance of a system capable of promptly identifying genuine or deceptive audio cannot be exaggerated. In this paper, the evaluation of audio-based deepfake identification methods has been surveyed, and their comparative analysis is being done based on the dataset usage, metrics for evaluation like AUC, EER, a language considered for the dataset taken and the factors such as MFCC, and CQCC. Moreover, the open challenges and future research directions have been highlighted.*

Keywords *- Audio, Deepfake, ASV spoof, GAN, Deep Learning, CNN, RNN, ResNet.*

1. Introduction

Due to the increasing accessibility of technology, a substantial number of deepfake videos are being disseminated via social media platforms. Deepfake is a term used to describe digitally altered media, like pictures or videos, in which a person's appearance has been substituted with the resemblance of someone else. Deepfake is an overly concerning problem growing in significance within contemporary society. The technique known as deepfake has been commonly employed to superimpose the faces of well-known stars from Hollywood onto explicit images and videos. Deepfake technology has been employed to generate deceptive information and spread rumours concerning politicians [1-4]. Furthermore, deepfakes were previously utilized in the context of the 2020 US campaign to alter videos of Joe Biden, specifically altering footage to depict him with his tongue protruding. The detrimental applications of deepfakes can significantly impact our society and contribute to disseminating misleading information, particularly on social media platforms. Generative Adversarial Networks (GANs) are advanced Deep Learning (DL) models that are enforced to produce counterfeit videos and pictures that are challenging for humans to differentiate from genuine ones. These models are practised for training on a dataset and subsequently generate counterfeit pictures and videos.

The larger the size of the data set, the model may generate the higher number of credible images and videos. Audio deepfakes, notably those incorporating human speech, pose a significant threat due to modern society's widespread use of speech as a biometric identifier [5]. The voice-based systems are vulnerable to audio spoofing attacks. The associated methodologies may encompass audio playback, artificial speech synthesis, voice transformation, and so forth. Disguise-based methods are frequently used in voice-based crimes, including vishing, which involves attempting to bypass voice authentication systems and making fraudulent calls. Using voice disguises a significant danger to autonomous voice biometric platforms [6, 7]. Advancements in voice conversion techniques and text-to-speech software have made it easier to synthesize human speech. This paves the way for a future where audio will be just as important as video in detecting deepfakes [8-11]. In voice forensics and artificial intelligence domains, multiple automated identification methods have been developed for identifying deepfakes [12, 13]. Nevertheless, there is a noticeable lack of experiments conducted on humans in comparison to machines' ability to understand altered audio, particularly controlled speech. The freshly matured voice biometrics technique has limitations, including the utilization of tools to imitate voices. Identifying these artificial voices could make the proof valid in a court of

legislation;. At the same time, the technically rigorous methods employ established procedures and demonstrate their ability to conduct research accurately and gain recognition from the academic community, the court will probably recognize them. There is a fundamental deficiency in the procedures and methods used to accurately detect voice deep fakes [14]. While deepfake videos have received significant attention, identifying audio deepfakes has gained comparatively little consideration. Recently, voice manipulation has advanced significantly. Synthetic voices pose a risk not only to robotic verification of the speaker's infrastructure but additionally to voice-assisted equipment used with the Internet of Things (IoT) [15].

Also, there was a reported case where bank robbers copied a speech made by a company executive to deceive other employees into sending large sums of money to a hidden account [16]. The coming ten years of deepfake detection are expected to bring a distinct challenge in the form of voice cloning. Hence, it is crucial to not only concentrate on identifying video signal tampering but also to scrutinize audio forgeries, which is lacking in current approaches. The discussion on deepfake identification approaches is constrained and does not include a discussion on audio deepfakes. The detection of audio deepfake has thus become a highly researched field, with the emergence of sophisticated techniques and DL methods. The paper categorization has been done accordingly. Section 2nd covers the audio deepfake identification process. Literature work about the DL-based methods in the realm of audio deepfakes has been showcased in the 3rd section. The comparative analysis of literature based on factors of importance has been executed under the 4th section. The open research issues have been highlighted under section 5th. At last, the conclusion is given in the section 6th.

1.1. Voice Deep Fake Identification Process

Despite a time when technology increasingly blurs the distinction between actuality and deception, deepfakes have emerged as a growing concern. The significance of detecting deepfakes has increased, particularly in sectors that require customer identity verification. The possible exploitation of deepfake audio highlights the crucial importance of identification in upholding security and trust. Audio-based deepfakes are additionally employed to produce authentic-sounding audio recordings of individuals speaking in various languages and fabricated recorded dialogues with the intention of manipulating public discussions. Since this technology could cause harm, we must recognize its risks and take precautions to prevent exploitation of society. An audio deepfake, sometimes also referred to as "synthetic voice", is a voice that is being generated, and it maps very accurately with someone's real voice. It copies the person's linguistic patterns, pronunciation, the flow of speaking, and other vocal characteristics to create a similar one to the normal voice, which is hard to discover. Audio or voice deepfakes are normally created with the help of ML algorithms like Generative Adversarial Networks (GANs). The usage cases of audio-based deepfake are quite enormous, and it covers numerous application areas like the media industry and social media. It is likewise used in various illegal activities such as impersonation, money deception, etc.

To ensure the identification or recognition of audio deepfakes, there is a need for methods that comprise ML models, forensic methods for audio, and further analysis of the behavioural aspects of individuals to separate the original and the fake audio recordings. Thus, selecting the best and the right ML method is particularly important for accurate audio deepfake identification. The main ML or DL models used in deepfake identification for the audio samples are Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid ones. The utilization of pre-trained models for audio classification is very fruitful in the initial stages. Also, incorporating techniques of signal processing and statistical analysis based to identify deviations in audio samples, like spectral characteristics, can add more advantages to the process of deepfake identification. Feature extraction is also essential for differentiating original audio against curated audio. The modelling process used for the identification can utilize either Mel-frequency Cepstral Coefficients (MFCCs), spectrogram images, or a hybrid of both as input features.

The original audio samples have different and unique background noise patterns, but the curated or deepfake patterns have few variations or problems, such as no background noise or any additional noise aspects combined. The software used for the audio deepfake identification: "The deepfake detection" software uses the spectral characteristics of an audio signal to identify any anomalies based on the mixing of the signals. Using the deepfake for audio generation poses a few issues pertaining to frequency ranges, pitch of person and fluency, which leads to the identification using advanced technologies by the experts.

2. Related Work

The term audio deepfakes is often used to refer to the act of artificially creating sounds using sophisticated ML and DL methods that mimic real audio. There is a need to have a comprehensive knowledge of how these are developed to identify an audio deepfake with more precision. The reason for this is that there has been an increased number of cases of fake audio being used more frequently in fraudulent activities. In this context, we highlight some key advancements made around detecting audio deepfakes. This work [17] presents a Quadratic Support Vector Machine (Q-SVM) algorithm for differentiating between unnatural and real audio. The authors used the binary categorization technique to group the sounds into two types: natural sounds generated by nature and artificial sounds created by humans. On Q-SVM, other techniques were surpassed in terms of an accuracy of 97.56% and a misclassification rate of 2.43%. In this project, authors

developed a Random Forest (RF) assisted Support Vector Machine (SVM) method that exploits distinctive characteristics of synthetic speech for predicting them in advance [18]. They trained their models using data from the 2019 ASV spoof challenge dataset [19].

The significance of removing the features in ML models manually and doing adequate preprocessing before training to achieve the best performance has been emphasized by some scholars. However, this method is cumbersome and can lead to inconsistencies that cause scientists to develop advanced DL methods. To tackle this issue, authors in [20] devised a pioneering method for identifying synthetic audio using two CNN models, namely Efficient-CNN and RES-Efficient-CNN. In [21], authors created a categorization model called Deep4SNet, which utilized a 2D CNN model to distinguish between real and synthetic audio in an audio dataset. The Deep4SNet model demonstrated a detection accuracy of 98.5% when identifying copy and artificial audio. Later [22], the authors conducted a comparison between CNN and the random method in terms of their performance when recognizing fake audio signals. In another study conducted [23], authors evaluated the effectiveness of CNN and Bidirectional LSTM models with ML models. This approach was focused on addressing the artificial nature of the dataset term AR-DAD [24] through imitation. The researchers evaluated the efficacy of CNN and LSTM models in discriminating genuine voices from imitators. While the CNN method had lower accuracy than the ML models, it improved at identifying false correlations.

In [25] authors provided a one-dimensional CNN and Siamese CNN for the need to identify counterfeit audio. For the 1-D CNN, the model's input consisted of speech log possibilities. The Siamese CNN consisted of two similar CNNs, like the 1-D CNN. However, they were combined using an entirely connected layer that included a SoftMax layer. Both models were evaluated on the ASV dataset, and it was found that the suggested Siamese CNN executed better than the GMM and 1-D CNN. Specifically, the Siamese CNN improved the Equal Error Rate (EER) by 55% when utilizing the LFCC features [26]. Nevertheless, using Constant Q Cepstral Coefficients (CQCC) features resulted in a slight decrease in performance. Additionally, it was discovered that the model lacks sufficient robustness and only functions effectively with a particular feature type.

In a separate study, the authors introduced a different CNN architecture in reference [27]. This model involved converting the audio data into scatter plots, which were then used as input for the CNN framework. The suggested model has been trained on the Fake or Real (FoR) dataset [28]. Although the suggested model was trained using data from different generation computations to address the generalization problem of DL-based techniques, it failed to meet the efficacy criteria for other models reported in the existing works. The accuracy, which was measured at 88%, and the EER, which was measured at 11%, were both inferior to the performance of the other DL models that were evaluated in the experiment. Therefore, it is necessary to enhance the model and incorporate additional data transformers.

In [29], authors created a Deep Neural Network (DNN) method called Deep-Sonar to analyze the neuron behaviours of Speaker Recognition (SR) systems when exposed to artificially generated fake audio produced by Artificial Intelligence (AI). Nevertheless, the performance of DeepSonar was significantly impacted by ambient noise in real-world conditions. Another work by authors [30] employed DNNs to compare between fake and truthful voices. Remarkably, this collectiveness leads to an impressive 94% preciseness in identifying audio produced by AI tools. However, the DNN fails to include a significant amount of artifact information when considering the feature representation.

In [31], the authors introduced a novel framework that combines Transfer Learning (TL) and the ResNet-34 technique to detect manipulated English-based voices. The TL model underwent pretraining on the CNN network. In an analogous way, the authors in [32] examined feature and image-based methods for categorizing artificially generated fake audio. This work utilized two novel DL models, namely the Temporal Convolutional Network (TCN) and the Spatial Transformer Network (STN). The TCN model shows an elevated level of success in accurately differentiating between counterfeit and authentic audio, achieving an impressive accuracy of 92%, whereas the STN model achieved a lower accuracy of 80%. Khalid and his colleagues introduced a novel dataset called FakeAVCeleb [33], specifically designed for Deepfake research [34]. The researchers examined unimodal techniques that incorporated five classifiers to assess their effectiveness in identification. The researchers determined that all the single-mode classifiers were ineffective in detecting counterfeit audio. In [35], the authors emphasize the necessity of creating a system for identification that relies on residual CNN.

Another method proposed by authors in [36] utilizes variations of the Squeeze-Excitation Network (SENet) and ResNet to combat spoofing. The ASV dataset is being used to assess the model, revealing that the method achieved a relative improvement of over 17% in fake audio. Nevertheless, the model demonstrated a t-DCF cost and EER of zero when evaluated under a logical access scenario, suggesting a significant issue of overfitting. In [37], the authors introduced an SSAD model inspired by the PASE+ method. The dataset was assessed and achieved an EER of 5.31%. Although the SSAD demonstrated satisfactory efficiency and scalability, its effectiveness was comparatively inferior to other DL methods. The primary cause for concern in deepfake is that it uses GAN to produce artificial audio that mimics real individuals. Due to

the multitude of methods for producing artificial speech, the task of identifying synthetic speech is exceedingly challenging. However, it is possible to generate synthetic speech by employing simple cut-and-paste techniques that achieve waveform concatenation, a process that is frequently available in the form of open-source toolkits. Additionally, it can be generated utilizing the source or bandpass filter model of the voice stream [38]. Recently, numerous presentations have been made on approaches based on CNNs for generating synthetic audio. These yield highly authentic outcomes that are challenging to differentiate from actual speech for human listeners and automated systems. The audio anti-spoofing investigation team has focused on the broader subject of detecting artificial language synthesis. In [39], the authors introduced a lip-oriented visual speaker verification method to safeguard against both human-oriented and deepfake risks. This method can effectively identify deepfake attacks, even if the attacker has a prior idea of the video creation mechanism. The reason for this is that the attacker has insufficient knowledge of the victim to mimic his/her speech patterns correctly in a spontaneous speech. It is also found that their method has the capability of detecting and rejecting various manipulation methods, which can produce the highest number of deepfake efforts. Moreover, they also present a learning-based approach for genuine and deceptive deepfake content detection [40]. Furthermore, they analyze both modalities in a video, relevant cues related to the emotion expressed, to decide whether it is fake or genuine. Lastly, researchers verify their model by measuring its performance with the Area Under the Curve (AUC) metric. This distinguishing method employs both audio and video techniques, along with extracted emotions from these modalities using them for identifying deepfakes. The training and validation process follows a procedure introduced by the authors [30], who developed models yielding high accuracy rates. The speech-denoising element utilizes Multilayer-Perceptron and CNN architectures to clean and preprocess audio effectively. These architectures achieve accuracy rates of 93% and 94%, respectively. The task of converting text using Natural Language Processing (NLP) achieved an accuracy of 93%. Speaker tagging was performed using the RNN model, which achieved an accuracy of 80%. The last component employs a CNN structure to accurately distinguish between genuine and counterfeit sounds. In [41], the authors introduce a revised version of the ResNet model called Res2Net. The model is assessed by employing various acoustic features, and the most optimal performance is achieved using CQT features. This model demonstrates superior performance in detecting audio manipulation, although its ability to generalize needs further enhancement.

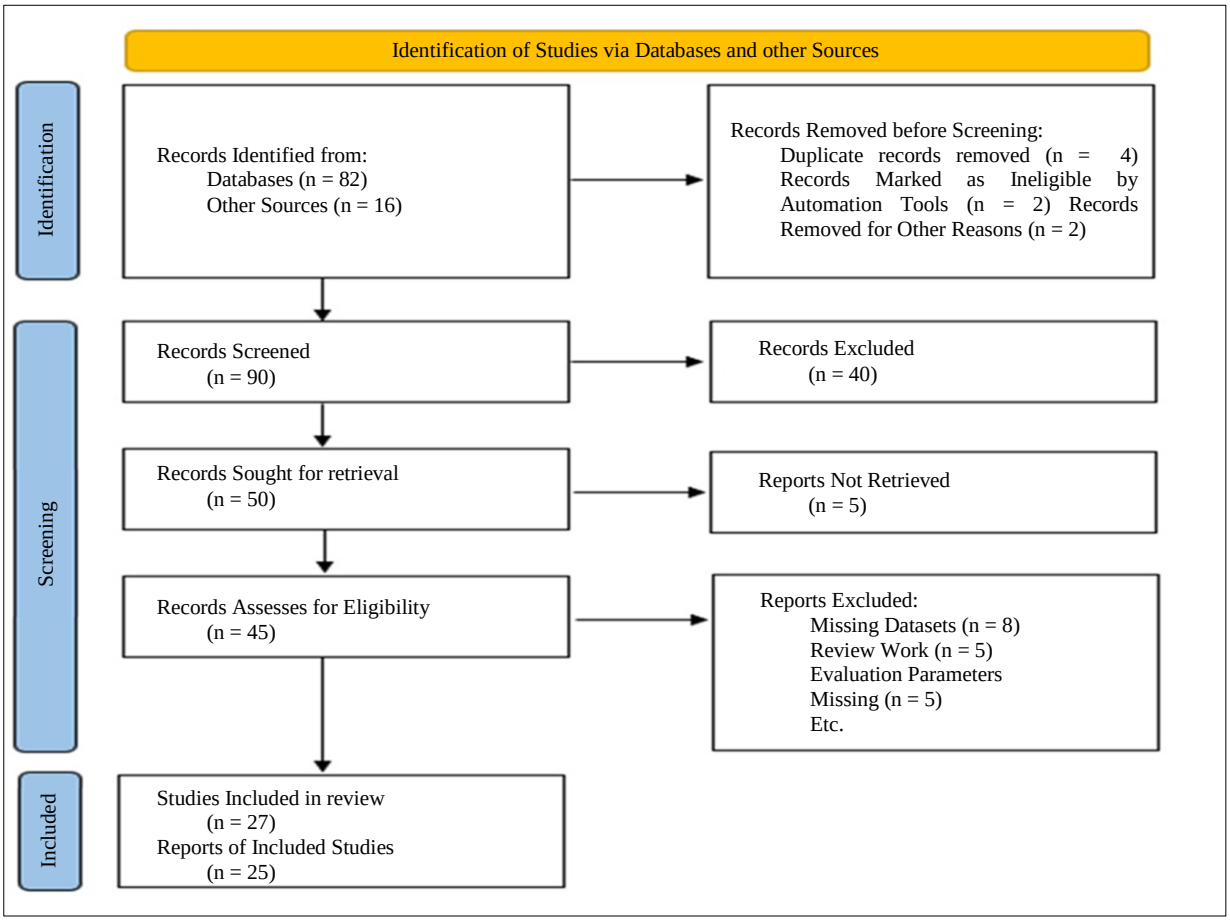

Fig. 1 Categorization of reviewed literature work for audio-based deepfake identification

In [42], the GMM and LCNN classifiers have been trained to identify forged speech using CQCC. In [43], the authors present a technique for identifying altered speech. Firstly, a signal compounding method is employed to enhance the variety of the training data through data augmentation. The technique enhances the precision of identifying counterfeit audio but necessitates a substantial amount of training data.

In their study [44], the authors proposed a framework that introduces a knowledge distillation loss function to improve the model's learning capability. This method is highly efficient in terms of computational resources and can identify previously unseen fraudulent alterations. However, its performance has not been assessed on samples with a prominent noise level. In [45], bi-spectral analysis is conducted to detect distinct and atypical spectral correlations found in speech samples created by GANs.

In [46], authors suggest an approach for detecting synthetic speech by utilizing inconsistencies. The researchers utilize a worldwide 2D-DCT characteristic to train a residual network for the purpose of identifying manipulated speech. This model exhibits superior generalization capability; however, its performance deteriorates when exposed to noisy samples. In [47], the authors present a model that uses an ensemble approach to detect synthetic speech.

DL models such as LCNNs and ResNets are employed to calculate deep features. These features are subsequently combined to distinguish between genuine and fake voices. The suggested model resists fake speech identification but must be evaluated using standard datasets. In [48], the authors introduce a DL-based framework designed to detect audio deepfakes. This work enhances the performance of detecting fake audio but is hindered by its significant computational burden.

In [49], the authors propose a method for detecting audio spoofing. They look at energy levels and short-term zero-crossing rates to find the silent spots in each audio signal. This approach [49] prevents over-fitting but requires significant computational resources. The aggregate grouping of the research work has been presented in Figure 1. This represents how many papers are initially identified and then the remaining ones that are decisively preferred to be included in this work.

2.1. Comparative Evaluation

This section covers the evaluation and analysis of literature work pertaining to the voice-oriented deepfake highlighted in the above section based on factors depicted in Tables 1 to 3 to reflect the main inferences drawn from the same. Table 1 demonstrates that effectiveness is more significantly influenced by the technique type instead of the feature utilized. Nevertheless, the ability to scale ML methods is not guaranteed, particularly when dealing with a substantial amount of audio files, because of the intensive training and manually extracted features involved.

Conversely, the utilization of DL algorithms necessitated specific modifications to the audio files to guarantee their compatibility with the algorithms. Several speech processing factors, like MFCCs, are being used extensively. The outcome of implementing the MFCC is a matrix that comprises feature vectors derived from each frame. Most of the research has focused on evaluating audio content in the English language. Similarly, Table 2 describes the datasets used primarily for the audio deepfake identification. The discussed techniques for identification employ models that require initial training on a data sample. Various datasets were documented in the literature, accompanied by the corresponding identification technique, whereas other studies prioritized clarifying the data they used and its inherent features.

Table 1. Comparative evaluation of audio deepfake identification methods based on techniques used and factors

Reference Work	Language Type English	CNN Based	ML Based	CQCC	MFCC
[17]	✓	×	✓	×	✓
[18]	✓	×	✓	×	×
[20]	✓	✓	×	×	×
[21]	✓	×	×	×	×
[22]	✓	✓	×	×	×
[23]	×	×	×	×	×
[25]	×	✓	✓	×	✓
[29]	✓	×	×	×	✓
[31]	✓	×	×	×	×
[33]	✓	×	×	×	✓
[36]	✓	×	×	✓	×
[37]	✓	×	×	✓	×

Table 2. Comparative evaluation of audio deepfake identification methods based on dataset usage

Reference Work	Language Type English	CNN Based	ASV Dataset	FoR Dataset	Other Dataset
[24]	×	×	×	×	✓
[28]	✓	×	×	✓	×
[34]	✓	✓	✓	×	✓
[18]	✓	×	✓	×	×
[39]	✓	✓	×	×	✓
[40]	✓	×	×	×	✓
[30]	✓	✓	×	✓	×

Table 3. Comparative evaluation of audio deepfake identification methods based on metrics

Reference Work	ResNet	CNN Based	CQCC	MFCC	EER	AUC	ASV Dataset
[15]	×	×	×	✓	✓	×	✓
[17]	×	×	×	×	×	×	×
[31]	✓	×	×	×	✓	×	✓
[41]	✓	×	×	×	✓	×	✓
[42]	×	×	✓	×	✓	×	✓
[43]	×	✓	×	×	✓	×	✓
[44]	×	✓	×	×	✓	×	✓
[45]	×	×	×	×	×	✓	×
[46]	✓	×	×	×	✓	×	✓
[47]	✓	✓	×	×	✓	×	✓
[48]	✓	×	×	×	✓	×	✓

From Table 2, it can be inferred that most datasets have been specifically created for the English language. The researchers primarily utilize the ASV and FoR datasets to identify audio deepfakes in most cases. Also, English language-based data is evaluated for deepfake identification in all cases. Table 3 highlights the evaluation of techniques based on metrics like AUC EER along with the other factors such as the dataset used, technique employed and the factors considered.

It is being inferred from Table 3 that whether the technique used is Resnet or CNN-based for audio deepfake identification, the metric that is being considered for the evaluation is EER. While AUC is being used in a very nominal manner also, most of the works have considered the ASV dataset for the evaluation. Moreover, less attention was paid to the CQCC and MFCC parameters during the analysis.

3. Open Challenges and Future Works

The emergence of deepfake technology has been recognized as a pressing threat that requires immediate action. In this scenario, the task of identifying deepfakes remains a significant obstacle. The current detection techniques primarily rely on ML and DL algorithms, utilizing features gathered from deepfake images and videos. Nevertheless, the precision and resilience of deepfake identification techniques remain inadequate. Here, we have defined the main challenges and futuristic aspects to work on in the field of audio deepfakes.

3.1. Advancing Technologies

Both the methods for creating deepfakes and the methods for detecting them are constantly evolving. Nevertheless, the existing detection methods still fall short in terms of accuracy. Additionally, the technology used to generate deepfakes is continuously advancing. Additionally, the existing methods for detecting deepfakes are not amazingly effective.

3.2. Datasets Quality

Most current methods for detecting deepfakes rely on DL algorithms, which necessitate large training datasets. Enhancing the quantity and quality of datasets can result in more accurate detection outcomes. A large amount of data is necessary to effectively detect deepfakes, both for training the detection models and evaluating their performance. Nevertheless, the accessibility of high-quality deepfake images and videos is currently restricted. These datasets typically lack diversity and an inadequate number of scenarios, rendering them unsuitable for detecting deepfakes in real-world scenarios. Additionally, the task of creating a comprehensive dataset that encompasses a wide range of instances for the explicit purpose of detecting deepfakes is demanding, primarily because of the potential privacy issues that may arise.

3.3. Lack of Standard for Evaluation of Models

Although evaluations of current deepfake identification methods have been conducted, there currently exists no widely recognized and consistent criterion for detecting deepfakes. Existing deepfake datasets exhibit variations in resolution, duration, and a dearth of diversity. Thus, it is crucial to possess standardized benchmark datasets to ensure precise identification of deep fake content. Furthermore, it is crucial to implement automated benchmark test methodologies to assess the efficacy of both deepfake generation and detection techniques.

3.4. Gathering Audio Datasets in Natural Environments

Most audio deepfake identification datasets are not obtained from real-life situations, which results in a lack of alignment with authentic utterances captured or generated under actual circumstances. The statements' actual circumstances may be unfavourable and exhibit a wider range of variation than the simulated circumstances.

3.5. Creating Extensive Multilingual Datasets

The prior datasets primarily consist of single-language data, with the majority being English deepfake audio datasets, along with a few others in Chinese or Japanese. The detection techniques may be influenced by the language used. However, it is imperative to develop detection systems that are not reliant on any specific language in practical applications. To enhance the reliability of counterfeit detection systems in various languages, it is necessary to assess the effectiveness of counterfeit detection models in situations involving multiple languages and a combination of languages within a single context.

3.6. Enhancing the Capacity for Generalization and Resilience of Detection Models

Despite previous research efforts in audio counterfeiting detection that have yielded promising results, the current detection models still exhibit inadequate generalization and robustness. However, their performance significantly deteriorates when evaluated on a dataset, including fake attacks, acoustic circumstances, or languages not encountered during training.

3.7. Enhancing the Comprehensibility of Detection Outcomes

Most current studies concentrate on differentiating counterfeit audio from genuine audio. Nevertheless, there is also a desire to exceed the limitations of binary classification that distinguishes between real and fake and instead accurately identify the specific sections of a partially manipulated speech that are fake and determine the origin of the fake audio. Furthermore, it is imperative to have the ability to comprehend the detection outcomes in real-world scenarios, such as audio forensics and attribution.

4. Future Works

Due to advancements in deepfake generation techniques, the presence of biometric features such as eye blinking may offer limited assistance in detecting deepfakes. Future advancements in deepfake detection will focus on creating systematic approaches that integrate multimodal signals from manipulated images and videos. These methods aim to achieve improved performance and durability.

The utilization of novel DL algorithms in the creation of deepfake content has significantly enhanced the quality of both images and videos. Conventional methods for detecting deepfakes are not adequately dependable or efficient, particularly when identifying deepfake videos. Developing strong deepfake detection techniques that can withstand adversarial attacks is imperative.

Another avenue to explore for deepfake detection is the utilization of datasets. Deepfake video recognition is more complex and requires more attention than deepfake image detection. DL models rely on large-scale datasets for effective detection methods. Significant improvements in the detection accuracy of deepfake media can only be achieved through the utilization of sufficient training data to extract significant characteristics.

5. Conclusion

The ease with which people can share pictures and videos on social networking sites has helped deepfake technology become more popular. This is particularly crucial in the present era as the accessibility of deepfake creation tools is increasing, and social media platforms readily facilitate the distribution and sharing of such fabricated content. However, when audio-based deepfake is considered, the work is not very elaborate. In this paper, a comparative evaluation of audio-based deepfake works by researchers has been analyzed based on the datasets used, metrics for the evaluation, the language used for the analysis and the techniques employed by them.

The researchers primarily utilize the ASV spoof and FoR datasets to identify audio deepfakes in most cases. Also, English language-based data is evaluated for deepfake identification in almost all cases. It can be concluded that EER is the measure used to judge whether the method used to find audio deepfakes is Resnet-based or CNN-based. At the same time, AUC is only used in a basic way. Most of the works have also used the ASV dataset for evaluation. On top of that, the CQCC and MFCC parameters got less attention during the analysis. Based on this investigation, it is evident that additional progress is required in the field of fake audio identification to devise a technique capable of identifying artificiality in various dialects or practical background noises in the future.

References

[1] Lakshmanan Nataraj et al., "Detecting GAN Generated Fake Images Using Co-Occurrence Matrices," *arXiv Preprint*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[2] Sheng-Yu Wang et al., "CNN-Generated Images are Surprisingly Easy to Spot... for Now," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 8695-8704, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] Chih-Chung Hsu, Chia-Yen Lee, and Yi-Xiu Zhuang, "Learning to Detect Fake Face Images in the Wild," *International Symposium on Computer, Consumer and Control*, Taichung, Taiwan, pp. 388-391, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[4] Mehdi Mirza, and Simon Osindero, "Conditional Generative Adversarial Nets," *arXiv Preprint*, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[5] Marwan Albahar, and Jameel Almalki, "Deepfakes: Threats and Countermeasures Systematic Review," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 22, pp. 3242-3250, 2019. [Google Scholar] [Publisher Link]

[6] Jonat John Mathew, "Towards the Development of a Real-Time Deepfake Audio Detection System in Communication Platforms," *arXiv Preprint*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Staffy Kingra, Naveen Aggarwal, and Nirmal Kaur, "Emergence of Deepfakes and Video Tampering Detection Approaches: A Survey," *Multimedia Tools and Applications*, vol. 82, no. 7, pp. 10165-10209, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Yuxuan Wang et al., "Tacotron: Towards End-To-End Speech Synthesis," *arXiv Preprint*, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[9] Sercan Arik et al., "Neural Voice Cloning with A Few Samples," *Advances in Neural Information Processing Systems*, vol. 31, 2018. [Google Scholar] [Publisher Link]

[10] Yipin Zhou, and Ser-Nam Lim, "Joint Audio-Visual Deepfake Detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, pp. 14800-14809, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Abu Qais et al., "Deepfake Audio Detection with Neural Networks Using Audio Features," *International Conference on Intelligent Controller and Computing for Smart Power*, Hyderabad, India, pp. 1-6, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Andreas Rössler et al., "Faceforensics++: Learning to Detect Manipulated Facial Images," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, pp. 1-11, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[13] A. Saleema, and Sabu M. Thampi, *Voice Biometrics: The Promising Future of Authentication in The Internet of Things*, Handbook of Research on Cloud and Fog Computing Infrastructures for Data Science, pp. 360-389, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[14] Ali Javed et al., "Towards Protecting Cyber-Physical and IOT Systems from Single-and Multi-Order Voice Spoofing Attacks," *Applied Acoustics*, vol. 183, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] Muteb Aljasem et al., "Secure Automatic Speaker Verification (SASV) System through sm-ALTP Features and Asymmetric Bagging," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3524-3537, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Mridul Sharma, and Mandeep Kaur, "A Review of Deepfake Technology: An Emerging AI Threat," *Soft Computing for Security Applications*, Singapore, pp. 605-619, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Yang Gao et al., "Generalized Spoofing Detection Inspired from Audio Generation Artifacts," *arXiv Preprint*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] Clara Borrelli et al., "Synthetic Speech Detection Through Short-Term and Long-Term Prediction Traces," *EURASIP Journal on Information Security*, vol. 2021, no. 1, pp. 1-14, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Massimiliano Todisco et al., "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," *arXiv Preprint*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[20] Nishant Subramani, and Delip Rao, "Learning Efficient Representations for Fake Speech Detection," *Proceedings 34th AAAI Conference on Artificial Intelligence*, vol. 34, no. 4, pp. 5859-5866, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[21] Dora M. Ballesteros et al., "Deep4SNet: Deep Learning for Fake Speech Classification," *Expert Systems with Applications*, vol. 184, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[22] Emily R. Bartusiak, and Edward J. Delp, "Frequency Domain-Based Detection of Generated Audio," *arXiv Preprint*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Mohammed Lataifeh et al., "Arabic Audio Clips: Identification and Discrimination of Authentic Cantillations from Imitations," *Neurocomputing*, vol. 418, pp. 162-177, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Mohammed Lataifeh, and Ashraf Elnagar, "Ar-DAD: Arabic Diversified Audio Dataset," *Data in Brief*, vol. 33, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[25] Zhenchun Lei et al., "Siamese Convolutional Neural Network Using Gaussian Probability Feature for Spoofing Speech Detection," *Interspeech*, pp. 1116-1120, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[26] Heinz Hofbauer, and Andreas Uhl, "Calculating A Boundary for The Significance from The Equal-Error Rate," *International Conference on Biometrics*, Halmstad, Sweden, pp. 1-4, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[27] Steven Camacho, Dora Maria Ballesteros, and Diego Renza, "Fake Speech Recognition Using Deep Learning," *Applied Computer Sciences in Engineering: 8th Workshop on Engineering Applications*, Medellín, Colombia, pp. 38-48, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[28] Ricardo Reimao, and Vassilios Tzerpos, "For: A Dataset for Synthetic Speech Detection," *International Conference on Speech Technology and Human-Computer Dialogue*, Timisoara, Romania, pp. 1-10, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[29] Run Wang et al., "Deepsonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices," *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle WA, USA, pp. 1207-1216, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[30] R.L.M.A.P.C. Wijethunga et al., "Deepfake Audio Detection: A Deep Learning Based Solution for Group Conversations," *2nd International Conference on Advancements in Computing*, Malabe, Sri Lanka, vol. 1, pp. 192-197, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[31] Joao Monteiro, Jahangir Alam, and Tiago H. Falk, "Generalized End-To-End Detection of Spoofing Attacks to Automatic Speaker Recognizers," *Computer Speech & Language*, vol. 63, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[32] Janavi Khochare et al., "A Deep Learning Framework for Audio Deepfake Detection," *Arabian Journal for Science and Engineering*, vol. 47, no. 3, pp. 3447-3458, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[33] Hasam Khalid et al., "Evaluation of an Audio-Video Multimodal Deepfake Dataset Using Unimodal and Multimodal Detectors," *Proceedings of the 1st Workshop on Synthetic Multimedia-Audiovisual Deepfake Generation and Detection*, Virtual Event China, pp. 7-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[34] Hasam Khalid et al., "FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset," *arXiv Preprint*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[35] Moustafa Alzantot, Ziqi Wang, and Mani B. Srivastava "Deep Residual Neural Networks for Audio Spoofing Detection," *arXiv Preprint*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[36] Cheng-I Lai et al., "ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual Networks," *arXiv Preprint*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[37] Ziyue Jiang et al., "Self-Supervised Spoofing Audio Detection Scheme," *Interspeech*, pp. 4223-4227, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[38] Kai-Da Xu et al., "60-GHz Third-Order On-Chip Bandpass Filter Using GaAs pHEMT Technology," *Semiconductor Science and Technology*, vol. 37, no. 5, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[39] In-Jae Yu et al., "Manipulation Classification for JPEG Images Using Multi-Domain Features," *IEEE Access*, vol. 8, pp. 210837-210854, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[40] Trisha Mittal, "Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues," *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle WA, USA, pp. 2823-2832, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[41] Xu Li et al., "Replay and Synthetic Speech Detection with Res2net Architecture," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, ON, Canada, pp. 6354-6358, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[42] Jiangyan Yi et al., "Half-Truth: A Partially Fake Audio Detection Dataset," *arXiv Preprint*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[43] Rohan Kumar Das, Jichen Yang, and Haizhou Li, "Data Augmentation with Signal Companding for Detection of Logical Access Attacks," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, ON, Canada, pp. 6349-6353, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[44] Haoxin Ma et al., "Continual Learning for Fake Audio Detection," *arXiv Preprint*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[45] Ehab A. AlBadawy, Siwei Lyu, and Hany Farid, "Detecting AI-Synthesized Speech Using Bispectral Analysis," *CVPR Workshops*, pp. 104-109, 2019. [Google Scholar] [Publisher Link]

[46] Yang Gao et al., "Generalized Spoofing Detection Inspired from Audio Generation Artifacts," *arXiv Preprint*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[47] Joao Monteiro, Jahangir Alam, and Tiago H. Falk, "Generalized End-To-End Detection of Spoofing Attacks to Automatic Speaker Recognizers," *Computer Speech & Language*, vol. 63, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[48] Tianxiang Chen et al., "Generalization of Audio Deepfake Detection," *Odyssey, The Speaker and Language Recognition Workshop*, pp. 132-137, 2020. [Google Scholar] [Publisher Link]

[49] Lian Huang, and Chi-Man Pun, "Audio Replay Spoof Attack Detection by Joint Segment-Based Linear Filter Bank Feature Extraction and Attention-Enhanced DenseNet-BiLSTM Network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1813-1825, 2020. [CrossRef] [Google Scholar] [Publisher Link]