Volume 9, Issue 3, March – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR766

Detection of Synthetically Generated Speech

Dr. Kavitha C[1]
[1]Head of Department and Professor, Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India

Pavan G[2]
[2]UG Scholar, Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India

Josh Kayyaniyil Joby[3]
[3]UG Scholar, Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India

R Vipul Nayak[4]
[4]UG Scholar, Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India

Rakesh Rathod[5]
[5]UG Scholar, Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India

Abstract:- Deepfake technology has made it increasingly difficult to discern real from fabricated audio, posing a significant challenge in the digital age. By employing sophisticated algorithms and voice recognition techniques, the system proposed in this article can analyse voice patterns and nuances to spot inconsistencies and anomalies, which are common indicators of deepfake voices and prevent scams and other types of cyber security issues.

Keywords:- *Deepfakes, Deepfake Audio, Synthetic Audio, Machine Learning, Acoustic Data.*

I. INTRODUCTION

In today's world, audio clips can be manipulated to impersonate individuals, leading to identity theft, reputational damage, and the spread of false information. By identifying and flagging such deceptive voice recordings, this project offers a vital layer of defence against malicious actors seeking to exploit audio for fraudulent purposes.

In an age where trust is paramount, the project places a strong emphasis on promoting authenticity in digital communication. Ensuring the honesty of voice-based interactions is critical to maintaining trust and credibility in personal and professional relationships.

By effectively addressing the challenges posed by deepfake technology, this initiative contributes to a more secure and reliable digital environment. People can communicate with confidence, knowing that their voices are not being manipulated or misused, and this trust fosters better relationships and information sharing in the digital realm.

To solve above mentioned problems, we are proposing a system with practical implementation of ML model into a application and demonstrate how we can secure the users from being spoofed.

II. LITERATURE REVIEW

Audio deepfakes are generally generated by using deep neural networks such as Generative Adversaria Network and other complex models.

- *Deep Fake Audio Detection via Mfcc*

This paper explores using Mel-frequency cepstral coefficients (MFCCs) and machine/deep learning models to detect deepfake audio. The models are evaluated on the Fake-or-Real dataset and achieve good accuracy, with SVM performing best on some subsets and VGG-16 on others. The approach shows promise for deepfake audio detection.

- *Contributions of Jitter and Shimmer in the Voice for Fake Audio Detection*

This paper investigates using jitter and shimmer voice features to detect fake audio, as these relate to prosody and can indicate unnaturalness. Various algorithms are explored for fundamental frequency estimation. Features are combined with Mel-spectrograms and fed to a neural network classifier. Results on the ADD 2022 and 2023 datasets show incorporating shimmer features can improve performance, indicating they provide complementary information.

- *Audio Splicing Detection and Localization Based on Acquisition Device Traces*

This paper tackles detecting and localizing audio splicing based on inconsistencies in traces left by the recording device model. A CNN extracts features, then clustering and distance measures localize splicing points. Enhancements handle multiple splices and refine localization.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR766

Experiments on a dataset built from MOBIPHONE show accurate detection and localization. Preliminary experiments also show potential for detecting splicing of real and synthetic speech.

III. METHODOLOGIES

A study explored using Mel-frequency cepstral coefficients (MFCCs) as audio features and machine learning models like SVMs, random forests, and XGBoost for detecting deepfake audio. Evaluated on the Fake-or-Real dataset with real and synthetic speech, SVMs performed best on some subsets, while transfer learning with VGG-16 showed top results on others. This indicates deep learning's potential for deepfake audio detection with sufficient training data.

The objective of audio device attribution is to relate a certain audio track to the device used for its acquisition. Certain studies refer to this challenge as the microphone classification task, although the classification encompasses the entire acquisition pipeline, not solely the microphone. The early works on audio source attribution focused on identifying microphones as the source devices. As a growing number of recordings originate from smartphones in recent times, the research in audio source attribution has shifted its focus to associating an audio recording with the mobile phone model used for its capture [4].

The second paper investigates using jitter and shimmer voice features that relate to prosody to detect fake audio. The motivation is that limitations in capturing rich prosody information can cause unnaturalness in synthesized speech. The authors test different fundamental frequency estimation algorithms and find that combining shimmer features with Mel-spectrograms improves performance when fed to a neural network classifier architecture. Testing on the challenging ADD 2022 and ADD 2023 datasets shows that the features provide the information to spectral features for fake audio detection.

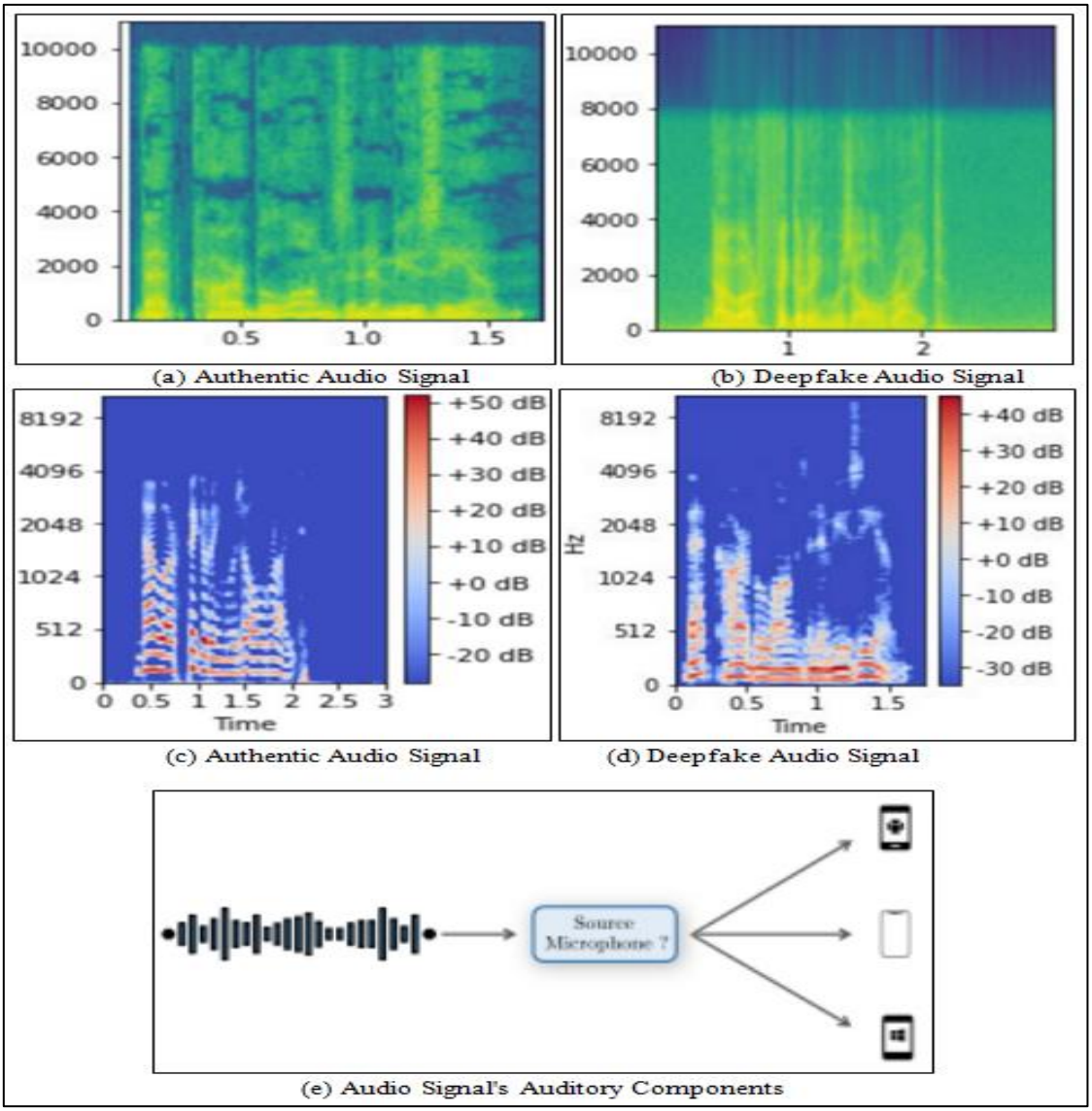


Fig 1 The Spectrograms in (a) and (b) Compare Deepfake and Authentic Audio Signals, Showing Noticeable Amplitude Differences. (c) and (d) Present the Amplitude in Decibels (e) for Better Understanding of the Audio Signal's Auditory Components

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR766

The third study concentrates on detecting and pinpointing audio splicing by employing a convolutional neural network to extract device-specific features across varying time windows. Through clustering, potential inconsistencies indicative of track splicing from multiple devices is identified. The localization of splicing points is achieved by detecting peaks in the distances between feature vectors. The evaluation, conducted on a dataset derived from the MOBIPHONE database, demonstrates the method's proficiency in detecting and localizing splicing points.

IV. PROPOSED METHODOLOGY

A key challenge with long audio is that fake segments may represent only a small portion of the entire file. Therefore, the first step is to employ a sliding window approach to divide the long recording into smaller segments for analysis. The segments should have 50% overlap to ensure no fake portion is missed at the boundaries. Each segment can then be analysed using a multi-faceted fake detection system. The first module will be based on the MFCC audio features and SVM classifier found effective in Paper 1. To complement this spectral approach and improve generalizability, a second module based on jitter, shimmer and other prosody features from Paper 2 will be included. The neural network classifier architecture from Paper 2 can be retrained on the prosody features.

Existing methods for detecting fake speech often do not directly analyse the differences between real and synthesized speech. The distinction typically arises from the fundamental challenge of making synthesized speech sound natural. Moreover, the unnatural qualities frequently stem from limitations in capturing the rich and diverse prosodic information. Analysing prosodic differences between real and fake speech thus shows promise for providing useful cues for fake audio detection.

Jitter and shimmer reflect inconsistencies in the pitch and amplitude between consecutive vocal cycles. They characterize audio frequency and amplitude perturbations (AFP). To illustrate AFP differences between real and synthesized speech, especially under degraded conditions, we compared segments of genuine and fake speech with identical linguistic content (/i/). After amplitude normalization to [0,1], Figure 1 shows the fake speech has much less stable amplitudes and periods over time.

Let F(i) be the frequency of the ith pitch period in an utterance. Parameters Lp with p=2,3,4 denote the number of consecutive periods used to compute AJp/CJp - specifically L_2=3, L_3=5, L_4=55 where N is the total number of periods. The jitter and shimmer metrics AJ_1/CJ_1, AJ_2/CJ_2, AJ_3/CJ_3, and AJ_4/CJ_4 are defined as:

$$AJ_1 = \frac{1}{N-1}\sum_{i=2}^{N}|F(i)-F(i-1)| \times \frac{1}{N}\sum_{i=1}^{N}F(i)\times 100 \quad (1)$$

$$CJ_1 = \frac{|F(i)-F(i-1)|}{N}\times\sum_{i=1}^{N}F(i)\times 100 \quad (2)$$

$$AJp = \frac{1}{N-Lp+1}\sum_{i=1+Lp-\frac{1}{2}}^{N-Lp-\frac{1}{2}}|F(i)-Fg(i)| \times \frac{1}{N}\sum_{i=1}^{N}F(i)\times 100 \quad (3)$$

$$CJp = \frac{|F(i)-Fg(i)|}{N}\times\sum_{i=1}^{N}F(i)\times 100 \quad (4)$$

$$Fg(i) = \frac{1}{Lp}\sum_{k=i-Lp-1/2}^{i+Lp-1/2}F(k) \quad (5)$$

Table 1 Architecture of the Deep Classifier for FAD, based on LCNN-BLSTM

Type	Kernel Shape	Output Shape	Param
Feature_CS3	-	[64, 404]	-
Feature_Mel-spectrogram	-	[64, 80, 404]	-
Conv2d_0	[1, 64, 5, 5]	[64, 64, 404, 80]	1.66k
MaxFeatureMap2D_1	-	[64, 32, 404, 80]	-
MaxPool2d_2	-	[64, 32, 202, 40]	-
Conv2d_3	[32, 64, 1, 1]	[64, 64, 202, 40]	2.11k
MaxFeatureMap2D_4	-	[64, 32, 202, 40]	-
BatchNorm2d_5	-	[64, 32, 202, 40]	-
Conv2d_6	[32, 96, 3, 3]	[64, 96, 202, 40]	27.74k
MaxFeatureMap2D_7	-	[64, 48, 202, 40]	-
MaxPool2d_8	-	[64, 48, 101, 20]	-
BatchNorm2d_9	-	[64, 48, 101, 20]	-
Conv2d_10	[48, 96, 1, 1]	[64, 96, 101, 20]	4.704k
MaxFeatureMap2D_11	-	[64, 48, 101, 20]	-
BatchNorm2d_12	-	[64, 48, 101, 20]	-
Conv2d_13	[48, 128, 3, 3]	[64, 128, 101, 20]	55.42k
MaxFeatureMap2D_14	-	[64, 64, 101, 20]	-
MaxPool2d_15	-	[64, 64, 50, 10]	-
Conv2d_16	[64, 128, 1, 1]	[64, 128, 50, 10]	8.32k
MaxFeatureMap2D_17	-	[64, 64, 50, 10]	-
BatchNorm2d_18	-	[64, 64, 50, 10]	-
Conv2d_19	[64, 64, 3, 3]	[64, 64, 50, 10]	36.93k
MaxFeatureMap2D_20	-	[64, 32, 50, 10]	-
BatchNorm2d_21	-	[64, 32, 50, 10]	-
Conv2d_22	[32, 64, 1, 1]	[64, 64, 50, 10]	2.11k
MaxFeatureMap2D_23	-	[64, 32, 50, 10]	-
BatchNorm2d_24	-	[64, 32, 50, 10]	-
Conv2d_25	[32, 64, 3, 3]	[64, 64, 50, 10]	18.50k
MaxFeatureMap2D_26	-	[64, 32, 50, 10]	-
MaxPool2d_27	-	[64, 32, 25, 5]	-
Dropout_28	-	[64, 32, 25, 5]	-
LSTM_1_blstm	-	[25, 64, 160]	154.88k
LSTM_1_blstm	-	[25, 64, 160]	154.88k
GAP	-	[64,160]	-
FC_1	-	[64,256]	103.68k
FC_2	-	[64,160]	41.12k
Feature_Concat	-	[64, 564]	-
FC_3	-	[64, 128]	41.09k
FC_4	-	[64, 2]	258
Total	-	-	653.41k

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR766

➢ *Architecture of the Deep Classifier for FAD, based on LCNN-BLSTM*

Approach proposed in Paper 3. The embeddings will be obtained from a CNN trained to classify real vs. fake samples as well as different fake generation methods. Significant peaks in the distance timeseries can indicate model changes. Periodic retraining of the CNN with updated datasets will address challenges such as adversarial attacks and evolving manipulation techniques, maintaining the system's efficacy over time.

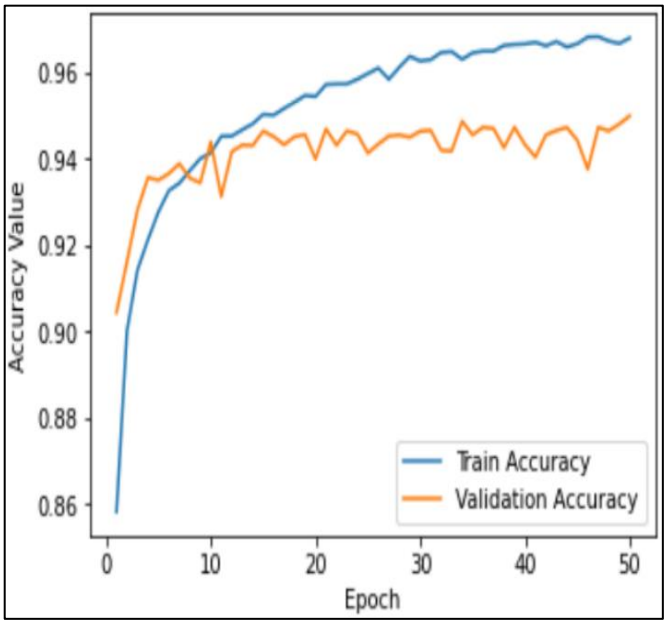
Finally, the outputs from the three modules will be fused using an ensemble classifier. Segments detected as fake by the majority of modules will indicate manipulated audio. The localization will be refined based on peaks found across all distance time series. After one pass of analysis, fake portions can be removed and the process repeated to find multiple edited sections.

Table 2 Accuracy Comparison for Machine Learning Models

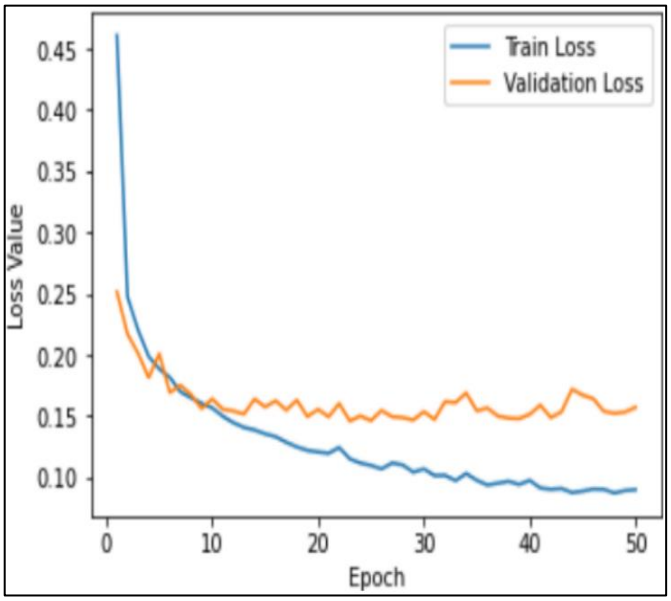
Models	for-2sec	for-norm	for-rerec
SVM	99.59	75.21	98.43
MLP Classifier	99.49	89.22	98.66
Decision Tree	87.52	65.10	82.12
Extra Tree Classifier	97.91	90.19	96.25
Logistic Regression	87.53	86.28	88.00
Gaussian Naive Bayes	79.77	80.16	82.14
Ada Boost	85.28	91.35	83.89
Gradient Boosting	92.29	93.50	88.86
XGBoost	92.22	94.25	88.92
Linear Discriminant Analysis	87.05	90.52	85.88
Quadratic Discriminant Analysis	96.22	65.15	95.59

Table 3 Comparison of the Accuracy of Machine Learning Models for Noisy Audio Signals

Models	for-2sec	for-norm	for-rerec
SVM	97.57	71.54	98.83
MLP Classifier	94.69	86.82	98.79
Decision Tree	87.13	62.16	88.28
Extra Tree Classifier	94.61	91.46	96.87
Gaussian Naive Bayes	88.20	81.81	81.91
Ada Boost	90.23	88.40	87.67
Gradient Boosting	94.30	92.63	93.51
XGBoost	94.52	92.60	93.40
Linear Discriminant Analysis	89.50	91.35	87.56
Quadratic Discriminant Analysis	96.13	61.36	96.91



(a) Model Accuracy



(b) Model Loss

Fig 2 The Figures (a) and (b) above depict the contrast between validation and training accuracy as well as loss.

This combined approach builds upon the complementary strengths demonstrated across the three papers to improve generalized fake audio detection in long files. The methodology can be extended by integrating additional forensic trace analyses from areas such as compression and environmental noise.

V. CONCLUSION

This paper has explored methods for detecting manipulated and fake audio, an increasingly critical challenge with recent advances in AI-based audio synthesis. We summarized key related works focused on deepfake audio detection using machine learning approaches. The methodologies were categorized based on usage of audio

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR766

features like cepstral coefficients versus voice prosody characteristics, as well as techniques leveraging artifacts from acquisition devices.

Our analysis showed that while spectral features remain an approach, there is significant complementary information in prosody and device traces. Furthermore, fusion of multiple detection modules can improve generalizability. The reviewed methods displayed promising results on datasets like Fake-or-Real and the ADD 2022/2023 challenges. However, real-world robustness remains an open research question.

Table 4 Comparison between Proposed Approach and Existing Approach

Approaches	Features	Models	Accuracy (%)
Existing Approach [36]	MFCC-20	RF	62
		KNN	62
		XGB	59
		NB	67.27
Existing Approach [28]	Timbre Model Analysis (Brightness, Hardness, Depth, Roughness)	SVM	73.46
		DT (J48)	70.26
	STFT, Mel-Spectrograms, MFCC and CQT	RF	71.47
		VGG19	89.79
Proposed Approach	MFCC-40, Roll-off point, centroid, contrast, bandwidth	LSTM	91
		VGG16	93

- *Comparison between Proposed Approach and Existing Approach*

In conclusion, deepfake audio represents an emerging and multi-faceted problem at the intersection of machine learning, signal processing and multimedia forensics. Impactful future work includes aggregating diverse forensic finger printers, testing on wild fake examples, and pushing robustness in unconstrained scenarios. There remains great potential for AI techniques to not just synthesize realistic audio, but also detect manipulations.

REFERENCES

[1]. A. Hamza, A. R. Javed, F. Iqbal, et al., "Deepfake audio detection via MFCC features using machine learning," Access, IEEE, 2022.

[2]. K. Li, X. Lu, M. Akagi, et al., "Contributions of jitter and shimmer in the voice for fake audio detection," Access, IEEE, 2023.

[3]. D. U. Leonzio, L. Cuccovillo, P. Bestagini, et al., "Audio splicing detection and localization based on acquisition device traces," J. Sel. Topics Signal Process., vol. 13, no. 5, pp. 1084–1097, 2019.

[4]. A. Abbasi, A. R. R. Javed, A. Yasin, Z. Jalil, N. Kryvinska, and U. Tariq, ''A large-scale benchmark dataset for anomaly detection and rare event classification for audio forensics,'' IEEE Access, vol. 10, pp. 38885–38894, 2022.

[5]. A. R. Javed, W. Ahmed, M. Alazab, Z. Jalil, K. Kifayat, and T. R. Gadekallu, ''A comprehensive survey on computer forensics: Stateof-the-art, tools, techniques, challenges, and future directions,'' IEEE Access, vol. 10, pp. 11065–11089, 2022.

[6]. A. R. Javed, Z. Jalil, W. Zehra, T. R. Gadekallu, D. Y. Suh, and M. J. Piran, ''A comprehensive survey on digital video forensics: Taxonomy, challenges, and future directions,'' Eng. Appl. Artif. Intell., vol. 106, Nov. 2021, Art. no. 104456.

[7]. A. Ahmed, A. R. Javed, Z. Jalil, G. Srivastava, and T. R. Gadekallu, ''Privacy of web browsers: A challenge in digital forensics,'' in Proc. Int. Conf. Genetic Evol. Comput. Springer, 2021, pp. 493–504.

[8]. A. R. Javed, F. Shahzad, S. U. Rehman, Y. B. Zikria, I. Razzak, Z. Jalil, and G. Xu, ''Future smart cities: Requirements, emerging technologies, applications, challenges, and future aspects,'' Cities, vol. 129, Oct. 2022, Art. no. 103794.

[9]. A. Abbasi, A. R. Javed, F. Iqbal, Z. Jalil, T. R. Gadekallu, and N. Kryvinska, ''Authorship identification using ensemble learning,'' Sci. Rep., vol. 12, no. 1, pp. 1–16, Jun. 2022.

[10]. S. Anwar, M. O. Beg, K. Saleem, Z. Ahmed, A. R. Javed, and U. Tariq, ''Social relationship analysis using state-of-the-art embeddings,'' ACM Trans. Asian Low-Resource Lang. Inf. Process., Jun. 2022.

[11]. C. Stupp, ''Fraudsters used Ai to mimic CEO's voice in unusual cybercrime case,'' Wall Street J., vol. 30, no. 8, pp. 1–2, 2019.

[12]. T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. HuynhThe, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, ''Deep learning for deepfakes creation and detection: A survey,'' 2019, arXiv:1909.11573.

[13]. Z. Khanjani, G. Watson, and V. P. Janeja, ''How deep are the fakes? Focusing on audio deepfake: A survey,'' 2021, arXiv:2111.14203.

[14]. Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, ''ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge,'' in Proc. Interspeech, Sep. 2015.

[15]. T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, ''The ASVSPOOF 2017 challenge: Assessing the limits of replay spoofing attack detection,'' in Proc. 18th Annu. Conf. Int. Speech Commun. Assoc., 2017, pp. 2–6.