Spectrogram-Based Analysis and Detection of Deepfake Audio Using Enhanced DCGANs for Secure Content Distribution

[1] Jovelin M. Lapates
Technological Institute of the Philippines
Aurora Blvd., Cubao, Quezon City, Philippines
qjmlapates@tip.edu.ph

[2] Bobby D. Gerardo, D.Eng
Northern Iloilo State University
Estancia, Iloilo, Philippines
bgerardo@nisu.edu.ph

[3] Ruji P. Medina, Ph.D
Technological Institute of the Philippines
Aurora Blvd., Cubao, Quezon City, Philippines
ruji.medina@tip.edu.ph

***Abstract*—While DCGAN as deep learning model utilizing spectrogram, allows for detection of deepfake audio, it is prone to overfitting which affects its ability to discriminate between real and fake audio. In this study, batch normalization is incorporated into both the generator and discriminator to address training instability. The datasets, consisting of real human speech and DeepFake renditions generated through Retrieval-based Voice Conversion (RVC), are categorized into 'REAL' and 'FAKE' (FoR) classes and preprocessed using Audacity and Sonic Visualizer. The paper introduces an enhanced DCGAN model for augmenting samples in voiceprint recognition and evaluates various spectrogram techniques—Mel-Spectrograms, GTCC, MFCC, and Chroma-CQT—to improve detection accuracy. The model achieved a training accuracy of 92.86% and a validation accuracy of 91.67%, underscoring the potential of advanced deep learning methods to ensure audio authenticity against deepfake threats.**

Keywords—waveforms, spectrogram, deepfake audio, CNN, DCGANs

I. INTRODUCTION

The proliferation of deepfake technology, particularly in audio, poses significant threats to the integrity and security of digital content distribution. Deepfake audio involves the synthesis of realistic-sounding speech that mimics a target speaker, and it can be used maliciously to disseminate false information, sway public opinions, or impersonate individuals for fraudulent activities. The increasing sophistication of these deepfakes necessitates the development of robust detection mechanisms to ensure the authenticity and reliability of audio content across various applications, including media, communications, and cybersecurity.

A spectrogram provides a comprehensive visualization of audio, depicting time, frequency, and amplitude within a single graph. A time-domain signal is split into equal-length segments by signal analysis software, which then applies the Fast Fourier Transform (FFT) [1] to each segment to transfer the data from the time domain to the frequency domain and create a spectrogram. Dynamic signals, such as output signals from communication devices, vary in amplitude over time, presenting two primary challenges: analyzing them, particularly through spectral analysis, and processing the signal, such as designing filters to achieve desired transformations. To analyze time-varying signals from devices, it is essential to establish a waveform that visualizes and converts the data into a useful format. Plotting the signal's amplitude on the vertical axis and time on the horizontal axis produces this waveform..

Converting an audio signal to a spectrogram is a crucial step in detecting deepfake audio. A neural network can automatically identify and extract significant voiceprint features associated with the study target for classification once the speech signal has been transformed into a spectrogram [2]. The visual representation of these spectrograms allows deep learning models, particularly convolutional neural networks (CNNs), to analyze and differentiate between authentic and synthesized audio. Using audio signal spectrograms has proven highly effective in detecting deepfake audio [1]. Artificial intelligence techniques are used to create realistic-looking but phony images, sounds, and videos, a phenomenon known as "deepfake technology.". Deepfake audio often exhibits inconsistencies in spectral patterns due to artificial synthesis methods, which can be detected by analyzing the spectrogram. Deepfakes are a serious threat to society, democracy, national security, and individual privacy, which emphasizes the necessity for efficient detection techniques to counter these possible dangers [3].

A family of machine learning models called Generative Adversarial Networks (GANs), first introduced by [4], are used to produce fake images. A GAN is consist of two networks: the discriminative network determines whether the generated images are legitimate, and the generative network uses an encoder and decoder to produce fake images. Several variations of GANs have been developed and used for supervised learning, semi-supervised learning, and picture synthesis. However, DCGANs, which are used to generate high-resolution images, are prone to model collapse and non-convergence, leading to poor generation quality. To ensure stable GAN training, the Wasserstein Distance in the discriminator is used by Wasserstein Generative Adversarial Networks (WGANs) to make sure that the parameter matrix meets Lipschitz requirements, which guarantees stable GAN training [5]. Excellent results were obtained using the Spectral Normalization (SN) approach, which was proposed in [6] for SNGAN, on datasets like LSUN, CIFAR-10, and ILSVRC2012 [2].

Five GAN variants, DCGAN, Wasserstein GAN (WGAN) [5], WGAN with gradient penalty (WGAN-GP) [7], Least Squares GAN [8], and PGGAN [9]-have all been utilized to create fictitious 64x64 images. To validate the suggested technique, a total 10, 000 test photos with both fake and actual data and 385,198 training images were obtained [10]. The importance of spectrogram-based analysis lies in its ability to visually represent the frequency spectrum of audio signals over time, enabling detailed analysis of audio characteristics that are often imperceptible to the human ear. This method transforms audio signals into spectrograms, facilitating the identification of subtle anomalies and patterns that traditional audio analysis methods might miss.

II. RELATED WORKS

The increasing quality of deepfakes necessitates corresponding improvements in the performance of detection methods, as highlighted in [10]–[14] and summarized in Table 1.

TABLE 1. TECHNIQUES AND CLASSIFIERS FOR DETECTING DEEPFAKE

Methods	Classifiers/Techniques	Key Features	Dealing with (Kind of Data)	Datasets Used
Spectrogram Analysis	Convolutional Neural Networks (CNNs)	Time-frequency representation of audio signals	Audio signals (speech, music)	ASVspoof 2019, FakeAVCeleb
Deep Convolutional GANs (DCGANs)	Generative Adversarial Networks (GANs)	Generation and detection of synthetic audio	Synthetic and real audio data	Custom dataset for GAN training
Feature Extraction from Spectrograms	CNN-based Feature Extraction	Identification of spectral patterns and anomalies	Time-domain audio converted to spectrograms	LibriSpeech, VoxCeleb
Data Augmentation Techniques	Noise Addition, Time Stretching	Enhancing training data with varied samples	Augmented audio data	Augmented versions of existing datasets
Audio Classification	Transfer Learning, Fine-tuning	Pre-trained models adapted for specific tasks	Diverse audio datasets	Pre-trained models on large audio datasets (e.g., AudioSet)
Preprocessing combined with deep network	DCGAN, WGAN-GP and PGGAN.	Enhance the generalization ability of deep learning models	images	CelebA-HQ,
Pairwise learning	CNN catenated CFFN conto	Feature extraction using CFFN	images	Face images from CelebA,

Deep Convolutional Generative Adversarial Networks (DCGANs) offer a robust approach to detecting deepfake audio. By training on spectrogram representations, DCGANs learn to distinguish between real and synthetic audio samples by capturing unique features indicative of genuine audio and identifying discrepancies present in deepfakes. DCGANs can be integrated with other deep learning models, such Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), to further improve detection skills and allow for more thorough study of the temporal and spatial elements of audio data. This integrated process not only improves detection accuracy but also provides a scalable solution for analyzing large volumes of audio data, addressing the growing threat of deepfake audio in digital content distribution frameworks [15], [16].

To overcome limitations in data, [17] proposed a study utilizing DCGANs to enhance speech data using mel-spectrograms. This method enhances the robustness and accuracy of deepfake audio detection systems, providing a reliable way to counteract the increasing sophistication of audio deepfakes [15], [18]. By training DCGANs on spectrogram representations of audio, the model can learn to identify unique features of genuine audio and detect discrepancies in deepfake audio. This integrated process not only improves detection accuracy but moreover provides a scalable answer for analyzing large volumes of audio data.

III. METHODS

Datasets are sourced from Kaggle, consisting of two forms: real human speech and DeepFake renditions created using Retrieval-based Voice Conversion (RVC) [39]. These datasets are categorized into 'REAL' and 'FAKE' classes, with audio filenames indicating the original speaker and the converted voice. The tools required for converting audio to spectrograms include those for audio editing and pre-processing [20] and for converting audio files into spectrograms [21]. The original audio classification dataset [27] includes recordings of eight famous popular individuals, using AI to produce voice and genuine audio gathered from the internet and produced through RVC. The pre-processed audio files are saved in WAV format, and the corresponding spectrograms are saved as PNG images.

TABLE 2. DATASETS FOR TRAINING, TESTING AND VALIDATION

Individual	Source	Length (MM:SS)
Joe Biden	Victory Speech2	10:00
Ryan Gosling	Golden Globes Speech3	1:33
Elon Musk	Commencement Speech4	10:00
Barack Obama	Victory Speech5	10:00
Margot Robbie	BAFTAs Speech6	1:19
Linus Sebastian	Down Monologue	9:30
Taylor Swift	Women in Music Speech	10:00
Donald Trump	Victory Speech 9	10:00

3.1 Audio to Spectrogram Conversion

The generation process of spectrogram [19] is shown in Fig. 1.

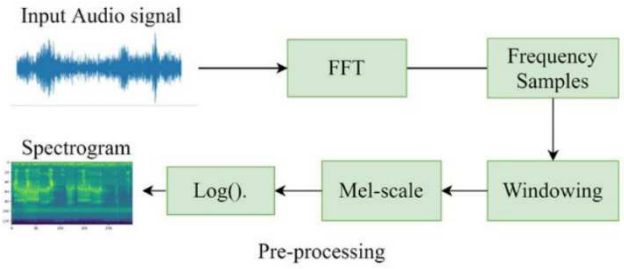

Fig 1. Conversion process of an input audio signal into spectrogram

The process begins with reading the audio file into a time-domain signal and applying the Fast Fourier Transform (FFT) to convert it into a frequency-domain representation, where the magnitudes of frequency samples represent the amplitude of the signal's frequency components. The signal is then divided into overlapping segments, each multiplied by a window function to reduce spectral leakage, and mapped to the Mel scale, spacing frequency bins more closely at lower frequencies and more widely at higher frequencies. Finally, a logarithmic transformation is applied to the Mel-frequency components to compress the dynamic range, resulting in a spectrogram that visually represents the time, frequency, and magnitude of the signal's components.

3.2 Audio Deepfake Detection Process

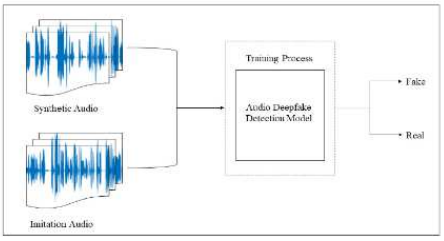

Fig. 2 Illustration of AD process [22]

The figure illustrates the general process for AD [30], [32], [33], [34] detection. Each audio clip is pre-processed and converted into Mel-spectrograms. After that, the detection model receives these features and uses them to do all necessary tasks, including training. In order to introduce nonlinearity, the output is sent through a fully connected layer with an activation function. This produces a prediction probability that classifies the audio as either real (class 1) or fake (class 0).

3.2.1 DCGANs Architecture

Following the successful applications of convolutional neural networks (CNNs) [23], [29], DCGANs were created by combining CNNs with GANs after researchers realized its potential [24] used a deconvolutional neural network architecture for the generator and fully convolutional networks in place of the multi-layer perceptron (MLP) structure that was originally used by [25]. Furthermore, batch normalization and ReLU activation were integrated into DCGAN. As a result, deconvolution was used more frequently in GAN generator architecture, and DCGAN became more well-liked [26].

One of the main architectural aspects of DCGAN is that strided convolutions in the discriminator and fractional-strided convolutions in the generator take the place of pooling layers. Both the discriminator and the generator use batch normalization. Deeper architectures are made possible by removing fully connected hidden layers. With the exception of the output layer, which use tanh to accommodate input images scaled to the range [-1, 1] from the original range [0, 255], the generator uses ReLU activation for all layers. All layers of the discriminator use LeakyReLU activation.

Here are the formulas for the evaluation of different spectrograms. Mel-Spectrograms, GTCCs, MFCCs, and Chroma-CQT are used to extract key spectral and temporal features from audio signals to detect deepfake audio. These features highlight different aspects of the audio's frequency content, aiding in distinguishing genuine audio from manipulated versions.

a. Mel-Spectrograms;

$$S(f,t) = \quad log(\sum_{k=1}^{K}|X_k(f,t)|^2) \qquad (1)$$

where: $Xk(f,t)$ represents the magnitude of the Fourier transform of the k-th Mel filter applied to the windowed audio signal.

b. Gammatone Frequency Cepstral Coefficients (GTCC)

$$GTCC_i = \sum_{n=0}^{N-1} log|X_k|(\sum_{k=1}^{K}|Y_{k(n)}^2|).cos(\frac{\pi i(n+0.5}{N}) \qquad (2)$$

where: $Y_k(n)$ represents the output of the *k-th* gammatone filter at time index n, and N is the number of DCT coefficients.

c. Mel-Frequency Cepstral Coefficients (MFCC);

$$MFCC_i = \sum_{j-1}^{J} log(\sum_{k=1}^{K}|X_k(f_j|^2)).cos(\frac{\pi i(j+0.5}{J}) \qquad (3)$$

where: $X_k(f_j)$ represents the magnitude of the Fourier transform of the k-th Mel filter applied to the audio signal at frequency bin f_j, and J is the number of DCT coefficients

d. Chroma Constant-Q Transform (Chroma-CQT)

$$Chroma(c,t) = \quad (\sum_{f\in ChromaBand(c)}|X(f,t)|^2) \qquad (4)$$

where: $X(f,t)$ represents the magnitude of the constant-Q transform of the audio signal at frequency f and time t, and ChromaBand(c) represents the frequency bins corresponding to the c-th chroma band.

The DCGANs model performance is evaluated based on the following evaluation metrics:

Accuracy to measure the percentage of sample properly identified relative to the total number of samples.

$$accuracy = \frac{Number\ of\ correctly\ classified\ samples}{Total\ number\ of\ samples} x\ 100\%\ (5)$$

Precision is a metric that expresses the percentage of actual positive predictions among all positive predictions the model generates.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (6)$$

Recall, which is often referred to as sensitivity or true positive rate, quantifies the percentage of actual positive data instances that are genuine positive forecasts.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (7)$$

F1 Score**,** is the precision and recall harmonic mean.

$$F1\ Score = 2x\frac{Precision * Recall}{Precision + Recall} \quad (8)$$

The detection framework's performance is validated using 10-fold cross-validation to ensure reliability and generalizability across different datasets.

IV. RESULTS AND DISCUSIION

For the experiments conducted in this paper, the authors used an Intel(R) Core (TM) i7-7800x CPU at 3.50GHz, with 16 GB of memory, and a T4 GPU with 11GB of memory. The software requirements included Python 3.6.6, TensorFlow 1.10.0, and Google Colab.

3.1 Dataset Analysis

Audio segments [27] are cropped, preprocessed, and converted from waveforms into spectrograms [28], [33]. The bit rates of the audio samples range from 1411 kbps to 1536 kbps. Furthermore, the sample rates span between 48000 Hz and 44100 Hz.

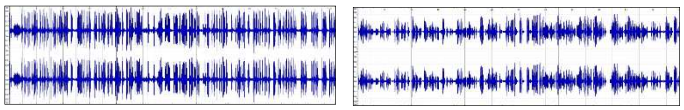

Fig 3. Sample Authentic vs Synthesized Audio Waveform

Fig. 3 represent the original audio waveforms of Linus's speech, while the Fig. 4 illustrate the transformed audio waveforms where Linus's speech is converted to sound like Biden's voice. After following the conversion steps from waveforms to spectrograms, the resulting spectrograms are presented below.

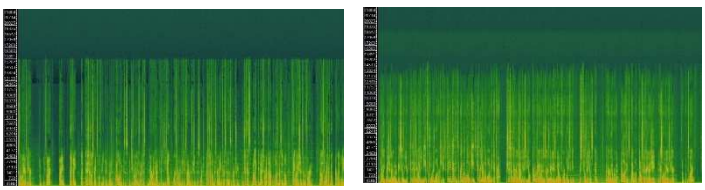

Fig. 4 Sample Authentic vs Synthesized Audio Spectrogram

TABLE 3. DCGAN PARAMETERS

Layer	Output Shape (N, H W)	Param #s
conv2d (Convo2D)	(None, 62, 62, 64)	1792
max_pooling2D	(None, 31, 31, 64)	0
conv2d_1 (Conv2D)	(None, 29, 29, 128)	73856
max_pooling2d_1	(None, 14, 14, 128)	0
flatten	(None, 25088)	0
dense	(None, 128)	3211392
dropout	(Non, 128)	0
dense_1(Dense)	(None, 2)	258

The table presents the parameters of the DCGAN. From left to right, each row describes the layer's characteristics, its output format, and how many parameters it has.. The output shape is represented as (N, H, W), where N is the number of feature maps produced, the width is W, while the height is H [31].

TABLE 4. METRICS WHILE APPLYING A CLASSIFER BASED ON RULES

Class	Metric		
	Precision	***Recall***	***F1-Score***
Real	0.370	0.590	0.450
Fake	0.880	1.000	0.930
Weighted Average	0.770	0.880	0.820

Metrics were evaluated using a single rule-based classifier to split predictions based on the 2nd Mel Frequency Cepstral Coefficient. Using this feature, a mean accuracy of 88% was achieved over 10-fold cross-validation.

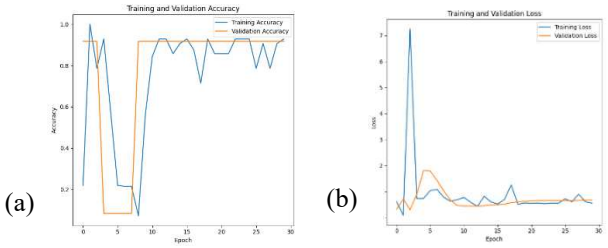

Fig. 5 Graph of a) Training/Validation Accuracy b) Training/Validation Loss

The training and validation data were processed over 30 epochs. Upon evaluating the model on the test data, the results are reflected in Fig. Over 30 epochs, the DCGAN model exhibits a noticeable upward trend in training and validation accuracy, as shown in Figure 8, indicating effective learning on the training data. Concurrently, the training and validation loss curves demonstrate a low and stable error rate, highlighting that the network is learning at a reasonable rate. By the end of the 30 iterations, the accuracy rate improves, and the loss value

decreases, confirming successful model generalization and convergence.

TABLE 5. OBSERVED STATISTICS IN THE DATASET ACROSS THE TWO DATA CLASSES

Attribute	**Real**			**Fake**		
(Overall Average)	***Mean***	***Med***	***Std.***	***Mean***	***Med***	***Std.***
Mel-Spectrograms	-52.50	-52.8	14.83	-56.7	-58.16	15.2
GTCC	-21.2	-5.66	94.3	-19.59	-3.98	91.56
MFCC	-21.29	-5.66	94.3	-19.5	-3.98	91.5
Chroma-CQT	0.491	0.45	0.28	0.458	0.41	0.29

The spectrogram evaluation metrics reveal distinct characteristics for each audio feature type. The Mel-Spectrogram values cluster around -54 dB with moderate variability, indicating consistent energy distribution across the frequency spectrum. The GTCC and MFCC metrics, sharing identical values, show significant variability with a mean around -20 and high standard deviation, reflecting diverse signal characteristics. In contrast, the Chroma-CQT metrics are more consistent, with values centered around 0.47, indicating stable harmonic content. These values are derived from the logarithmic scale (dB) used in audio signal processing, where negative values indicate lower energy or intensity levels.

The training results demonstrate a well-performing model with 92.86% for training and 91.67% for validation accuracy, indicating that the model has effectively learned the patterns in the data and generalizes well to unseen data. The test loss of 0.3137 and the lower validation loss further suggest that the model is not overfitting and maintains a balanced performance.

V. CONCLUSIONS

The study highlights the increasing threat posed by deepfake audio technology, which can convincingly mimic real speech, posing risks to misinformation, fraud, and cybersecurity. To combat these threats, the research explores the use of spectrograms—visual representations of audio frequencies over time—to detect deepfake audio. By converting audio signals into spectrograms, DCGANs can analyze and identify inconsistencies characteristic of synthetic speech. This approach forces the ability of deep learning models to process and classify complex patterns in the spectrograms, enhancing the accuracy of deepfake detection.

The study employs various techniques to improve detection, including data augmentation, preprocessing, and the integration of multiple spectrogram types such as Mel-Spectrograms, GTCC, MFCC, and Chroma-CQT. The study demonstrates that DCGANs, can effectively discriminate between real and fake audio, achieving an accuracy of 92.86%. The model's performance is validated through extensive testing and 10-fold cross-validation, showing promising results in distinguishing genuine audio from deepfake samples. Overall, the study underscores the potential of DCGAN methods to address the growing challenge of deepfake audio in digital content, ensuring the integrity and authenticity of audio communications.

ACKNOWLEDGMENT

The researchers would like to acknowledge Bukidnon State University for the scholarship grant and the faculty and staff of the Technological Institute of the Philippines-Quezon City for their invaluable support and guidance throughout the duration of this study.

REFERENCES

[1] Khodzhaev, Z. (2024). A Practical Guide to Spectrogram Analysis for Audio Signal Processing. arXiv, arXiv:2403.09321. https://doi.org/10.48550/arXiv.2403.09321

[2] Xu, Z. J., Wang, R. F., Wang, J., & Yu, D. H. (2020). Parkinson's disease detection based on spectrogram-deep convolutional generative adversarial network sample augmentation. IEEE Access, 8, 206888-206900.

[3] Zhang, T. Deepfake generation and detection, a survey. Multimed Tools Appl 81, 6259–6276 (2022). https://doi.org/10.1007/s11042-021-11733-y

[4] Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, A. Courville, and Y. Bengio, ''Generative adversarial nets,'' in Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS), vol. 2. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.

[5] M. Arjovsky, S. Chintala, and L. Bottou, ''Wasserstein generative adversarial networks,'' in Proc. Int. Conf. Mach. Learn., vol. 70, Aug. 2017, pp. 214–223.

[6] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, ''Spectral normalization for generative adversarial networks,'' 2018, arXiv:1802.05957. [Online]. Available: http://arxiv.org/abs/1802.05957

[7] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. arXiv preprint arXiv:1704.00028, 2017

[8] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, pages 2794–2802, 2017.

[9] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.

[10] Nguyen TT, Nguyen CM, Nguyen DT, Nguyen DT, Nahavandi S (2019) Deep learning for deepfakes creation and detection. arXiv preprint arXiv:190911573

[11] Masood, M., Nawaz, M., Malik, K.M. et al. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Appl Intell 53, 3974–4026 (2023). https://doi.org/10.1007/s10489-022-03766-z

[12] Pham, L., Lam, P., Nguyen, T., Nguyen, H., & Schindler, A. (2024). Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models. arXiv:2407.01777 [cs.SD]. Retrieved from https://doi.org/10.48550/arXiv.2407.01777

[13] Khochare, J., Joshi, C., Yenarkar, B. et al. A Deep Learning Framework for Audio Deepfake Detection. Arab J Sci Eng 47, 3447–3458 (2022). https://doi.org/10.1007/s13369-021-06297-w

[14] Paul, D., Pal, M., & Saha, G. (2017). Spectral Features for Synthetic Speech Detection. IEEE Journal of Selected Topics in Signal Processing, 11, 605-617.

[15] Wani, T.M., Amerini, I. (2023). Deepfakes Audio Detection Leveraging Audio Spectrogram and Convolutional Neural Networks. In: Foresti, G.L., Fusiello, A., Hancock, E. (eds) Image Analysis and Processing – ICIAP 2023. ICIAP 2023. Lecture Notes in

Computer Science, vol 14234. Springer, Cham. https://doi.org/10.1007/978-3-031-43153-1_14

[16] Jayan Shah, Pratham Shah, Mustansir Godhrawala, S. B. N. J. B. P. V. K. Y. N. S. K. . (2024). Harmonizing Algorithms: An Approach to Enhancing Audio Deepfake Detection. International Journal of Intelligent Systems and Applications in Engineering, 12(3), 1297–1304. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/5520

[17] Baek, J. Y., & Lee, S. P. (2023). Enhanced speech emotion recognition using dcgan-based data augmentation. Electronics, 12(18), 3966.

[18] Arniriparian, S., Freitag, M., Cummins, N., Gerczuk, M., Pugachevskiy, S., & Schuller, B. (2018, September). A fusion of deep convolutional generative adversarial networks and sequence to sequence autoencoders for acoustic scene classification. In 2018 26th European signal processing conference (EUSIPCO) (pp. 977-981). IEEE.

[19] Karam, S., Ruan, S.-J., Ul Haq, Q. M., & Li, L. P.-H. (2023). Episodic memory based continual learning without catastrophic forgetting for environmental sound classification. Journal of Ambient Intelligence and Humanized Computing, 14(4), 1-11. https://doi.org/10.1007/s12652-023-04561-5

[20] Audacity. https://www.audacityteam.org/. Accessed June 09, 2024

[21] Sonic Visualizer. http://www.sonicvisualiser.org/ Accessed June 7, 2024

[22] Almutairi Z, Elgibreen H. A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. Algorithms. 2022; 15(5):155. https://doi.org/10.3390/a15050155

[23] Jovelin M. Lapates , "Corn Crop Disease Detection Using Convolutional Neural Network (CNN) to Support Smart Agricultural Farming," International Journal of Engineering Trends and Technology, vol. 72, no. 6, pp. 195-203, 2024. Crossref, https://doi.org/10.14445/22315381/IJETT-V72I6P120

[24] A. Radford, L. Metz, and S. Chintala (2016), “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv:1511.06434.

[25] Dash, A., Ye, J., & Wang, G. (2023). A review of generative adversarial networks (GANs) and its applications in a wide variety of disciplines: from medical to remote sensing. IEEE Access.

[26] He, Yibo, Kah Phooi Seng, and Li Minn Ang. "Generative Adversarial Networks (GANs) for Audio-Visual Speech Recognition in Artificial Intelligence IoT." Information 14, no. 10 (2023): 575. https://doi.org/10.3390/info14100575

[27] Bird, J. J., & Lotfi, A. (2023). Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion. arXiv:2308.12734 [cs.SD]. https://doi.org/10.48550/arXiv.2308.12734

[28] Levy et al. (2022). Classification of audio signals using spectrogram surfaces and extrinsic distortion measures. EURASIP Journal on Advances in Signal Processing, 2022:100. https://doi.org/10.1186/s13634-022-00933-9

[29] Karandikar, A., Deshpande, V., Singh, S., Nagbhidkar, S., & Agrawal, S. (2020). Deepfake Video Detection Using Convolutional Neural Network. International Journal of Advanced Trends in Computer Science and Engineering, 9(2), March-April. Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse629220 2020.pdf https://doi.org/10.30534/ijatcse/2020/62922020

[30] Khanjani Z, Watson G and Janeja VP (2023) Audio deepfakes: A survey. Front. Big Data 5:1001063. doi: 10.3389/fdata.2022.1001063

[31] Bartusiak, E. R., & Delp, E. J. (2022). Frequency Domain-Based Detection of Generated Audio. ArXiv. /abs/2205.01806

[32] Nasar, B. F., Sajini, T., & Lalson, E. R. (2020). A survey on deepfake detection techniques. International Journal of Computer Engineering in Research Trends, 7(8), August. E-ISSN: 2349-7084.

[33] Bayat, N., Khazaie, V. R., Keyes, A., & Mohsenzadeh, Y. (2021). Latent vector recovery of audio GANs with application in deepfake audio detection. DOI: 10.21428/594757db.1ee3922d.

[34] Ilyas, H., Javed, A., & Malik, K. M. (2023). AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection. Applied Soft Computing, 136, 110124. https://doi.org/10.1016/j.asoc.2023.110124

[35] Hassan, F. and Javed, A., 2021, April. Voice spoofing countermeasure for synthetic speech detection. In 2021 International Conference on Artificial Intelligence (ICAI) (pp. 209-212). IEEE.

[36] Javed, A., Malik, K.M., Malik, H. and Irtaza, A., 2022. Voice spoofing detector: A unified anti-spoofing framework. Expert Systems with Applications, 198, p.116770.

[37] Arif, T., Javed, A., Alhameed, M., Jeribi, F. and Tahir, A., 2021. Voice spoofing countermeasure for logical access attacks detection. IEEE Access, 9, pp.162857-162868.

[38] Javed, A., Malik, K.M., Irtaza, A. and Malik, H., 2021. Towards protecting cyber-physical and IoT systems from single-and multi-order voice spoofing attacks. Applied Acoustics, 183, p.108283

[39] Deep Voice - Deepfake Voice Recognition [Dataset] [Online]. Available: Kaggle. https://www.kaggle.com/datasets/birdy654/deep-voice-deepfake-voice-recognition