Proceedings of the Nineth International Conference on Communication and Electronics Systems (ICCES-2024)
IEEE Xplore Part Number: CFP24AWO-ART; ISBN: 979-8-3503-7797-2

Detection of Deepfake Audio using Deep Learning

Dr.A. Jaya Lakshmi
Department of ECE,
Vardhaman College of Engineering
Shamshabad, Hyderabad.
Jayalakshmi417@gmail.com

V.Sindhuja
Department of ECE,
Vardhaman College of Engineering
Shamshabad, Hyderabad.
vallakatlasindhuja21ec@student.vardhaman.org

B. Meghana
Department of ECE,
Vardhaman College of Engineering
Shamshabad, Hyderabad.
meghanabolla21ec@student.vardhaman.org

G. Sweekar Gupta
Department of ECE,
Vardhaman College of Engineering
Shamshabad, Hyderabad.
guptaguggillasweekargupta21ec@student.vardhaman.org

2024 9th International Conference on Communication and Electronics Systems (ICCES) | 979-8-3503-7797-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCES63552.2024.10859752

***Abstract*—Deepfake technology employs AI algorithms to create misleading digital information, such as photos, movies, and audio recordings. Deepfakes can now create realistic-looking content, making detection more challenging. Most studies on detecting audio deepfakes use various machine learning and deep learning algorithms with the ASVSpoof dataset, despite recent advancements in video deepfake detection. MFCCs are used to extract useful acoustic information. The dataset was constructed using a text-to-speech model and then partitioned into a few datasets. Datasets are divided into subsets according to bit rate and audio length. The experimental results demonstrate that support vector machines (SVMs) outperformed other machine learning (ML) models in terms of accuracy. Our objective is to enhance the model's efficiency and detection accuracy through the use of deep learning techniques with CNN and LSTM neural networks. Using Generative Adversarial Networks (GAN) as a classifier, that can determine whether audio is real or fraudulent. And a Graphical User Interface (GUI) was further implemented for a better visual representation of the output,that is obtained**

***Index Terms*—ASVSpoof(Automatic speaker verification spoofing and countermeasures), MFCCs(Mel-frequency cepstral coefficients) ,LSTM(long-short term Memory)**

I. Introduction

Audio deepfake technology is not new in the public sector. It (i.e. audio spoofing) has been present since the 90s and reached some maturity with Google's WaveNet in 2016 whichgenerates very natural-sounding fake speeches. Researchers at UC Berkeley invented Voice Conversion in 2017, which changes a voice into the guise of someone else. This has eased the generation of deepfake audio.

There are audio deepfakes employed for fraud and disinformation. Voice Cloning — Scammers pretending to be someoneelse to extract money or personal information. They may evenget the technique of creating fake celeb footage that could mislead individuals.

Deepfake photos and videos have long threatened people's privacy, safety, and security. The deepfake technology wasn't restricted to photos and videos; it was further expanded to audio. Audio spoofing has become one of the major problems in our society. Audio spoofing refers to the process that involves a person trying to pretend to be a particular person or celebrity, which brings a lot of misinterpretation of the original information, further causing a threat to the safety and privacy of people in the public domain. Deepfake audio is now a greater issue for society. And this issue must be addressed to provide a safe and secure environment for people to work, and live in.

Audio deepfakes are used for several purposes by scammers, In most cases, audio deepfakes are used to mimic people and manipulate individuals into transferring money or revealing sensitive information. It did not stop with a threat to personal privacy. But further, it is impacting various sectors such as media, politics, and and cybersecurity.

Many studies have been performed on video, photo, and audio deepfakes. Most researchers utilized the ASVSpoof dataset, machine learning, and deep learning methods to identify fake audio. The detection of deepfake audio is done using sound patterns in datasets such as ASVSpoof. Historically, machine learning methods such as Support Vector Machines (SVM), random forest models, and decision tree models had been implemented for identifying fake sounds. Among these models, SVM outperformed the other machine learning models in terms of detection. However, these models generally fail to detect complex patterns in data.

Deepfake audio detection plays an important role in forensic investigations, cases involving criminal activity, and legal conflicts requiring audio evidence. Researchers aim to support law enforcement organizations forensic experts, and legal professionals in authenticating audio recordings by employing accurateand reliable detection procedures.

II. RELATED WORK

Advanced detection algorithms are beneficial in detecting deepfake content, which is difficult to identify because of

its a realistic replication of media. This study demonstrates that while Gradient Boosting works well with the for-normal dataset and its sub-datasets, Support Vector Machines perform best with the original dataset using VGG-16[1].[2]Advanced synthetic speech makes it harder and harder to detect audio deepfakes. Using six classifiers and three dataset subsets, this study achieves 26 percent greater accuracy than earliermethods by utilizing the Fake-or-Real dataset and a varietyof feature extraction and optimization strategies.[3] In social media, forensics, and cybersecurity, synthetic voices are essential for speech therapy and may even facilitate fraud and false information. This work effectively separates real from fake audio using deep learning and machine learning withMel-Frequency Cepstral Coefficients (MFCCs).While deep learning successfully solves challenging problems in computervision, human-level control, and data analytics, it also brings up privacy and security risks. Deepfake technology produces incredibly realistic media by utilizing CNNs and GANs. Using VGG-16 and RNN networks for image recognition, this study investigates deepfake production techniques and detection methodologies, emphasizing the necessity for cutting-edge technology to maintain media integrity [4].[5] Speech recognition and computer vision have benefited from recent deep-learning breakthroughs. This work uses text-to-speech (TTS) systems to study speaker diarization and synthetic speech. Our system successfully distinguishes between synthetic and real speech by utilizing Deep Neural Networks with datasets such as FakeOrReal and Urban-Sound8K.Speech denoising, speaker diarization with RNNs and NLP, and precise audio categorization with CNNs are important elements.[6]It has been suggested to use Convolutional Neural Networks (CNNs)to distinguish between artificial and human voices in order to identify phony talks. Both national security and personal safety may be threatened by deepfake audio. The effective detection of deepfake voices is made possible by the conversion ofaudio signals into 2D images (Spectrogram, MFCC, FFT, and STFT) to reduce computation. These images are thennumerically processed for CNN input.[7]Since deep learning makes it possible to create extremely realistic deepfakes, efficient detection is essential to preventing misuse in contextssuch as the manipulation of public opinion and legal evidence.Although CNNs are extensively utilized, this study emphasizes the need for better methodologies and more research. It analyzes machine learning-based deepfake detection techniques across audio, picture, video, and hybrid media.[8]The increasing realism of audiovisual content brought about by deep learning innovations, such as VAEs and GANs, has raisedquestions regarding the legitimacy of digital media. While most attention is now focused on photo and video deepfakes, this study intends to close this gap by offering a thorough review of approaches for both detecting and creating audio deepfakes, which will help with the development of detection models in the future.[9]The research suggests employing Mel spectrograms and Convolutional Neural Networks (CNNs) forartificial speech detection to tackle the ethical and social concerns associated with synthetic media. The best CNN architecture demonstrated high efficacy across multiple voice datasets, achieving up to 98% accuracy with the WaveFake dataset and 94% accuracy with FoR.Voicecloning systems are employed in gaming and marketing, and they can produce speech that is identical to a small sample size, which presents security problems. This paper examines techniques for forensically identifying deepfake audio by utilizing spectrogram, MFCC, Mel-spectrum, and Chromagram analysis of audio properties. Chromagram and Spectrogram were the best tools for custom designs, and MFCC features were where VGG-16 shines [10].[11]The deep learning method for deepfake audio detection presented in this work combines filters like Mel and DCT with a variety of spectrogram transformations, including STFT, CQT, and WT. It uses pre-trained audio models with MLP classification, leverages transfer learning from vision models and directly applies CNN, RNN, and C-RNN to spectrograms. On the ASVspoof 2019 dataset, the top ensemble model demonstrated enhanced detection accuracy with a competitive Equal Error Rate (EER) of 0.03.[12]Realistic deepfake material may now be produced and shared more quickly thanks to recent developments in machine learning. Conventional techniques frequently concentrate on either visual or aural detection. In response, we present AVFakeNet, a framework for integrated audio-visual deepfake detection that leverages Dense Swin TransformerNet (DST-Net). We show notable gains in accuracy over five datasets.[13][14]New developments in machine learning have produced increasingly believable deepfakes. This paper proposes two frameworks: AVFakeNet using Dense Swin Trans- former Net (DST-Net) and Multimodaltrace for cooperative audiovisual learning. Multimodaltrace produced noteworthy findings on other datasets and obtained 92.9 percentage accu-racy on Fake Celeb.[15]Custom Convolutional Neural Network(cCNN) and two pre-trained models(VGG16 and MobileNet) are among the network topologies that are suggested to address synthetic audio recognition. cCNN outperformed all baseline models and set new benchmarks with its highest accuracy of
97.23 percent on the deepfake audio dataset. Large data sets can be analyzed using machine learning, however manyof the methods have grown more complicated. This study used fundamental components like buttons and text boxes to create an intuitive machine-learning software with a straightforward graphical user interface. This system proved to be more user-friendly than previous open-source solutions [16-17].

III. PROPOSED METHODOLOGY

Deepfake audio detection is a difficult problem that involves analyzing audio signals to detect altered or manipulated audio. Deepfake audio can be identified using a hybrid architecture that combines Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs).The audio spectrograms' characteristics are extracted using the CNN component. To detect deepfake audio, local patternsin the spectrograms must be captured by the CNN architectureand the temporal patterns in the audio streams are analyzed by the LSTM component. To detect deepfake audio, the

LSTM architecture is made to detect cumulative dependencies in the audio signals [4][5].

We implemented CNN and LSTM in our research rather than with specific characteristics. Convolutional Neural Networks (CNNs) are used to apply filters to audio spectrograms for analysis and to classify audio according to Mel- Frequency Cepstral Coefficients (MFCCs). Long-term sequences of audio can have temporal dependencies captured by LSTMs, which are specialized for sequence data.

• Length of the Audio = 10 .

1. Data collection: The initial input, a 10-second genuine audio recording, is located in the data section. Resulting from the completion of the preprocessing and input data modeltraining. After that, the other input data being provided is a false audio file.
2. Audio Preprocessing & Audio Feature Extraction: For the model to learn the patterns that differentiate the two, genuine and fake audio data must be fed into the system throughout the training process. Training a model to precisely identify audio deepfakes in practical applications is the aim.We fed the resulting spectral waveforms into a hybrid architecture model, which uses particular algorithms like CNN and LSTM, to identify patterns.
3. classification of fake or real audio: Using Generative Adversarial Networks (GAN) as a classifier, one can determine whether audio is real or fraudulent. Make useof this provided input data, which includes both false and genuine audio signals, to train a GAN.While the discriminatornetwork learns to differentiate between actual and fake signals, the generator network learns to create artificially created audio signals that imitate real data. Then, identify incoming audio signals as true or fraudulent by employing the discriminator network as a binary classifier.

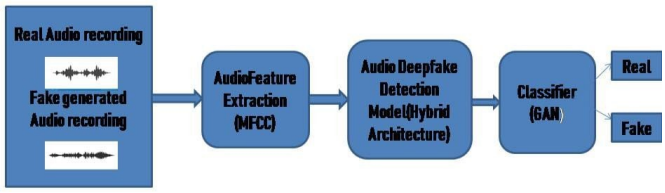

Fig. 1. Block Diagram of Detection of Deepfake Audio

The above fig.1 represents the input data is represented by the first block, which consists of both artificially created and actual human recordings.The audio feature extractor,located in the second block uses MFCC to extract the features of Audio through spectrograms.The Audio Deepfake Detection Model, located in the third block, uses the hydrid architecture includes CNN and LSTM analyzes the where CNN are particularly suitable for audio processing because they aremade to handle data having a grid-like architecture, like time-series data or pictures. CNNs can be used to evaluatethe spectrogram or Mel-spectrogram of an audio segment,which shows the audio data in a format closer to a 2D image and LSTMs can be applied to the analysis of the temporal relationships between audio samples in the contextof deepfake audio. The classifier, located in the fourth block, uses the Audio Deepfake Detection Model's output tocategorize audio as real or fake.

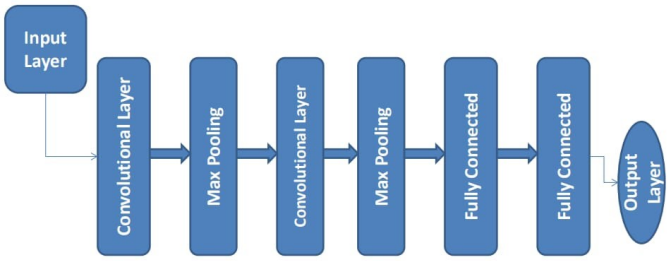

Fig. 2. CNN Architecture

The above Fig. 2 represents Preprocessed audio data, usually in the form of a spectrogram or Mel-spectrogram, is sent to the input layer. Convolution layers extract features such as edges and patterns by applying filters. Max pooling layers lower dimensionality and guard against overfitting, whereas activation layers (ReLU) induce non-linearity. Real or fake audio is classified binaryally by the output layer, whereasfully connected layers examine complicated connections.

Preprocessed audio data is fed into the input layer; this is usually in the form of spectrogram features or audio frames. In order to find patterns or abnormalities in the audio stream, the LSTM layer examines the sequence utilizing memory cell information. The output layer outputs the final binary result after the Dense layer has classified the audio as real or fake.

IV. Results and Disscusions

Through careful experimentation with 10 seconds of input data, the performance of our suggested hybrid architecture model was carefully assessed.To ensure an objective evaluation of valid model training, the input data was carefully processed for the detection analysis and worked ona graphical user interface (GUI) for integrated model selectionand visualization of detection results. These illustrations, which can help users in better understanding the system'soutput, include spectrograms, waveforms, and other graphical representations of audio data. The results we obtained show significant improvements over standard and advanced models in terms of efficiency and the ability to distinguish between genuine and fraudulent.

The output of the hybrid model combines CNN and LSTM networks, and a classifier employing a generative adversarial network processes it to separate artificial and real sound. Whereas LSTMs record temporal connections and patterns, CNNs extract spectral characteristics such as spectrograms. Combining CNN's local pattern identification with LSTM's long-term understanding of context enables the model to identify minute irregularities and inconsistencies in the audio.

As the graph shown above fig.3, the provided audio file converts speech into a graphical format, with the text reading ”Feel the fear anyway”and traines the model.

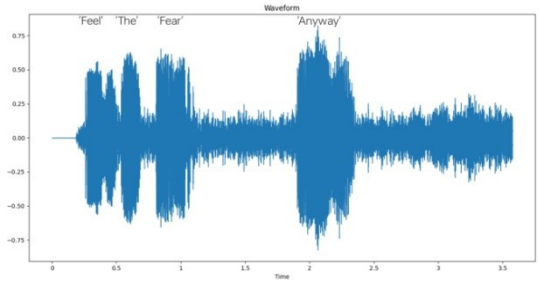

Fig. 3. Graphical Representation of real Audio

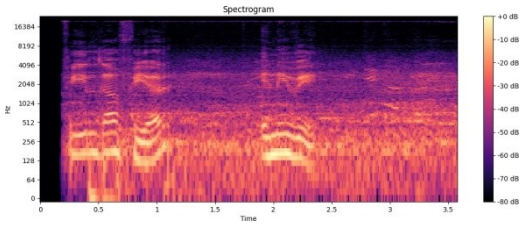

Fig. 4. spectrogram of real audio

The above figure.4 shows result is a real audio signal’s spectrogram.Time is on the x-axis,andfrequency is on the y-axis.The colour indicates the sound’s amplitude at that frequency and moment in time.

As the graph shown above,fig.5 the provided audio file converts speech into a graphical format, with the text reading ”Feel the fear and do it anyway”and traines the model and fig.6 represents the output is a spectrogram of a fake audio signal.One axis represents time, while the other represents frequency.The amplitude of the sound at that frequency and instant is indicated by the colour.To further corrupt the original audio, it appended a few words.

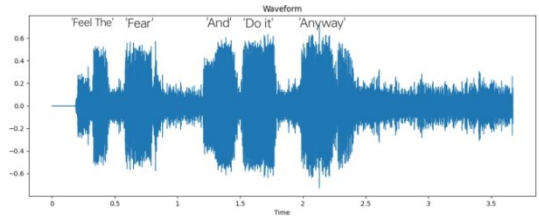

Fig. 5. Graphical representation of fake audio

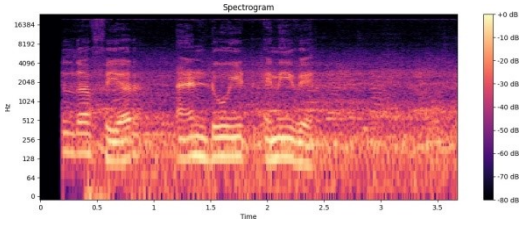

Fig. 6. spectrogram of fake audio

A. A graphical user interface

Through the use of spectrograms and the ability toinspect frequency components for anomalies, a graphical user interface (GUI) offers an interactive platform for audio research and deepfake identification. It provides CNN-LSTM model results in an intuitive manner and offers features like zooming and scanning. The GUI also makes it easier to select algorithms, change parameters, and upload audio files, which improves user interaction and increases the precision of deepfake audio detection[16].

Three buttons are displayed on the graphical user interface: the first button allows users to upload audio files as input data; the second allows users to listen to both real and fake audio; and the third button allows users to run authentication. By pressing the run authentication button triggers the modelto begin processing, providing the appropriate waveform and spectrogram as well as a message that opens up saying if the audio is real or fake.

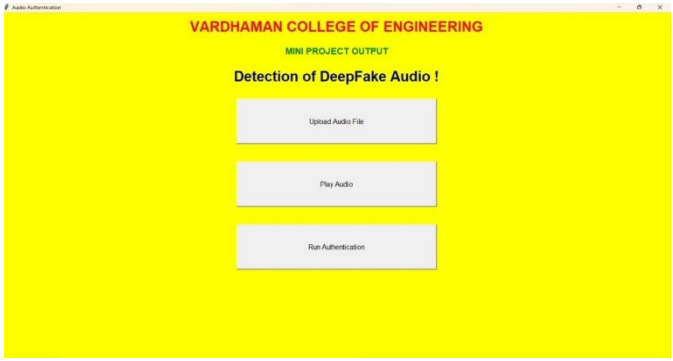

Fig. 7. Home page of GUI

The above figure.7 represents the Home page of Graphical User Interface consists of 3 buttons,the first one is Upload Audio File where the audio is given as input,the second one is play Audio where the audio is palyed for identification,the third one is Run Auttentication performs the operation and visulaizes the spectrograms,waveforms and gives text message of real or fake Audio.

The below figure.8 represents the detection of deepfake audio on graphical user interface,by pressing on run auttentication button it produces the graphical and spectrogram representaation of the real audio and a text message.

The below figure.9 represents the detection of deepfake

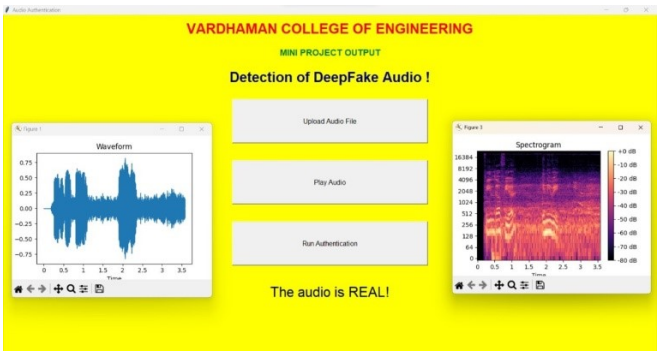

Fig. 8. Real Audio Detection

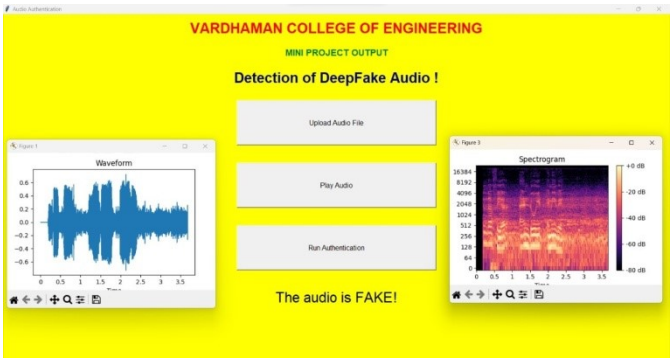

Fig. 9. Fake Audio Detection

audio on graphical user interface, by pressing on run authentication button it produces the graphical and spectrogram representation of the fake audio and a text message.

V. Conclusion

Prior to the final analysis, it was shown that the deep learning methods of identifying altered audio signals employing Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks were quite effective in identifying audio deepfakes. With CNNs and LSTMs working together, the model is able to capture temporal correlationsand properly express spectral information, which makes deepfake identification possible. In the audio spectrogram,the CNN component worked more effective in capturing local patterns and elements, while the LSTM component performed effectively in modeling long-term dependencies and temporal interactions within this signal. The method converts audio to a frequency-based format utilizing Mel-Frequency Cepstral Coefficients (MFCCs) to improve pitch and amplitude analysis for better deepfake identification. Generative Adversarial Networks (GANs) contribute by producing diverse training data, making the model more adaptive and robust to changes in audio quality and format. Combining CNNs, LSTMs, and GANs in a simple interface known as a graphical user interface can successfully verify audio authenticity. preventing the emergence of audio deepfakes will require the continued development of powerful but user-friendly detection algorithms.

References

[1] Hamza, A., Javed, A.R.R., Iqbal, F., Kryvinska, N., Almadhor, A.S., Jalil, Z. and Borghol, R., 2022. Deepfake audio detection via MFCC features using machine learning. IEEE Access, 10, pp.134018-134028

[2] Pham, L., Lam, P., Nguyen, T., Nguyen, H. and Schindler, A., 2024. Deepfake Audio Detection Using Spectrogram-based Feature and Ensemble of Deep Learning Models. arXiv preprint arXiv:2407.01777

[3] Kilinc, H.H. and Kaledibi, F., 2023, October. Audio Deepfake Detection by using Machine and Deep Learning. In 2023 10th International Conference on Wireless

[4] Nguyen, T.T., Nguyen, Q.V.H., Nguyen, D.T., Nguyen, D.T., Huynh-The, T., Nahavandi, S., Nguyen, T.T., Pham, Q.V. and Nguyen, C.M., 2022. Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223, p.103525.

[5] Wijethunga, R.L.M.A.P.C., Matheesha, D.M.K., Al Noman, A., De Silva, K.H.V.T.A., Tissera, M. and Rupasinghe, L., 2020, December. Deepfake audio detection: a deep learning-based solution for group conversations. In 2020 2nd International conference on advancementsin computing (ICAC) (Vol. 1, pp. 192-197). IEEE.

[6] Qais, A., Rastogi, A., Saxena, A., Rana, A. and Sinha, D., 2022, July. Deepfake audio detection with neural networks using audio features. In 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP) (pp. 1-6). IEEE.

[7] Heidari, A., Jafari Navimipour, N., Dag, H. and Unal, M., 2024. Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(2), p.e1520.

[8] Dixit, A., Kaur, N. and Kingra, S., 2023. Review of audio deepfake de- tection techniques: Issues and prospects. Expert Systems, 40(8),p.e13322.

[9] Valente, L.P., de Souza, M.M. and Da Rocha, A.M., 2024, May. Speech Audio Deepfake Detection via Convolutional Neural Networks. In 2024 IEEE International Conference on Evolving and Adaptive Intelligent Sys- tems (EAIS) (pp. 1-6). IEEE.

[10] Wani, T.M. and Amerini, I., 2023, September. Deepfakes audio detection leveraging audio spectrogram and convolutional neural networks. In In- ternational Conference on Image Analysis and Processing (pp. 156-167). Cham: Springer Nature Switzerland

[11] Iqbal, F., Abbasi, A., Javed, A.R., Jalil, Z. and Al-Karaki, J.N., 2022. Deepfake Audio Detection Via Feature Engineering and Machine Learn- ing. In CIKM Workshops.

[12] Ilyas, H., Javed, A. and 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, p.110124

[13] Raza, M.A. and Malik, K.M., 2023. Multimodaltrace: Deepfake detec- tion using audiovisual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 993-1000

[14] Anagha, R., Arya, A., Narayan, V.H., Abhishek, S. and Anjali, T., 2023, December. Audio Deepfake Detection Using Deep Learning. In 2023 12th International Conference on System Modeling and Advancement in Research Trends (SMART) (pp. 176-181). IEEE

[15] Mcuba, M., Singh, A., Ikuesan, R.A. and Venter, H., 2023. The effect of deep learning methods on deepfake audio detection for digital investigation. Procedia Computer Science, 219, pp.211-219

[16] Rosasn-Arias, L., Sanchezn-Perez, G., Toscano-Medina, L.K., Perez-Meana, H.M. and Portillo-Portillo, J., 2019, May. A graphical user interface for fast evaluation and testing of machine learning models performance. In 2019 7th International Workshop on Biometrics and Forensics (IWBF) (pp. 1-5). IEEE.

[17] Lakshmi AJ, Kishore G, Kumar T, Koushik V. Drone detection and classification using YOLOv8 and deep CNN. In: Ragavendiran SDP, Pavaloaia VD, Mekala MS, Cabezuelo AS, editors. Innovations and advances in Cognitive systems. ICIACS 2024. Information Systems Engineering and Management. Volume 15. Cham: Springer; 2024. https://doi. org/10.1007/978-3-031-69197-3_2.