Kalpa Publications in Computing

Volume 19, 2024, Pages 266–276

Proceedings of 6th International Conference on Smart Systems and Inventive Technology

Leveraging Deep Learning Architectures for Deepfake Audio Analysis

C.Siva Kumar[1], G. Siva Nageswara Rao[2], P.Vivek[3], G.Sivaram Deepak[4], M.V.Sai Pranay[5],G.Tharun Raj[6]

[1]Associate Professor, Dept of CSSE Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati, India
[2]Professor, Dept of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur (Dt), AP, India.
[3]UG Scholar, Department of CSSE Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati,India
[4]UG Scholar, Department of CSSE Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati, India
[5]UG Scholar, Department of CSSE Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati,India
[6]UG Scholar, Department of CSSE Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati,India
[1]sivakumar.c@vidyanikethan.edu, [2]sivanags@kluniversity.in, [3]vivekpatnam08@gmail.com
[4]deepakgandikota12@gmail.com, [5]pranaymiriyala2002@gmail.com, [6]tharuntrendz3626@gmail.com

Abstract

Deepfake content is created or changed artificially utilizing AI strategies to make it genuine. This research addresses the evolving challenge of detecting deepfake audio content, as recent advancements in deepfake technology have rendered it increasingly challenging to distinguish fabricated content. Leveraging machine and deep learning methodologies, specifically employing Mel-frequency cepstral coefficients (MFCCs) for sound component extraction, we focus on the Genuine-or-Fake dataset — a cutting-edge benchmark dataset generated through a text-to-speech (TTS) model. This dataset is arranged into sub-datasets because of sound length and spot rate. This study reveals that the Convolutional Neural Network (CNN) models exhibit the highest accuracy in identifying deepfake audio within the for-rerec and for-2-sec datasets. Meanwhile, the gradient

boosting model performs well in the for-norm dataset. This study illustrates the CNN model's outstanding performance on the for-original dataset, outperforming other cutting-edge models. This study advances the field of deepfake recognition, especially in the areas of audio manipulation, demonstrating the efficacy of CNN models in detecting fake content.

Keywords: Audio manipulation, Mel-frequency cepstral coefficients (MFCCs), Text-to-speech model, Convolutional Neural Network (CNN), Gradient boosting, Benchmark dataset, Deepfake acknowledgement.

1 INTRODUCTION

Nowadays, the expansion of deepfake technology has presented a formidable challenge in the realm of digital content authenticity. Deepfakes, synthetic media created or altered using sophisticated artificial intelligence (AI) techniques, have become increasingly indistinguishable from genuine content, raising concerns about misinformation and deception. As the technology evolves, detecting deepfakes, particularly in the audio domain, has become a pressing concern. This research addresses this evolving challenge by leveraging machine what's more, deep learning procedures, with a particular spotlight on the identification of deepfake sound content. The study employs advanced techniques, particularly utilizing MFCCs for sound element extraction. The research centres around the Fake or Genuine dataset, a cutting-edge benchmark dataset generated through a read-aloud model. This dataset is fastidiously ordered into sub-datasets based on audio length and bit rate, providing a comprehensive framework for calculating the performance of various detection models.

The experimental findings reveal the potential of CNN models in detecting deepfake audio within specific datasets, such as for-rece and for-2-sec, leveraging higher accuracy in fake news identification.

2 LITERATURE SURVEY

Researchers have recently explored diverse topics within digital forensics, audio forensics, and deepfake detection. This section provides an outline of recent evolutions and difficulties in these areas. The literature review begins with a focus on creating benchmark datasets for abnormality detection and uncommon event classification in sound forensics [1]. This aligns with the increasing significance of digital forensics, as highlighted in a survey covering best-in-class strategies, apparatuses, also, future directions in PC forensics [2]. A related survey explores the taxonomy, difficulties, and future directions in advanced video forensics [3], shedding light on the broader landscape of multimedia forensics. Privacyconcerns in web browsers, especially in the context of digital forensics, are discussed as a significant challenge [4]. Shiftinggears, the exploration of future smart cities and their technological requirements, applications, and challenges showcase the interdisciplinary nature of digital advancements [5]. Authorship identification using ensemble learning [6] and socialrelationship analysis through cutting-edge embeddings [7] demonstrate the evolving landscape of forensic techniques.Meanwhile, external threats such as AI-driven impersonation in cybercrime cases [8] and the creation and identification of deepfakes [9] highlight the urgency of developing robust countermeasures. The literature review delves into the specific area of automatic speaker verification spoofing and countermeasures, emphasizing the evolving challenges posed by adversaries. Additionally, it explores innovations in audio signal processing, such as fast

spectrogram inversion. The significance of appropriate sampling in audio signal processing is highlighted, drawing on historical perspectives. Moreover, the role of deep learning in ASV spoofing detection is discussed, integrating dynamic acoustic features and DNN classifiers. The challenge of replay attacks against voice assistants is addressed, showcasing the pervasivenature of security concerns in voice-based technologies. Start-to-end sound replay assault recognition utilizing deep convolutional networks with consideration is explored, presenting a specific solution to a prevalent issue. Utilizing pre-trained models on large audio datasets can improve the efficiency and effectiveness of deepfake detection models. Fine-tuning these pre-trained models on deepfake-specific datasets allows them to leverage existing knowledge while adapting to the specific task. Artificially manipulating training data (e.g., adding noise, and changing playback speed)can help the model generalize better and become more robust to real-world variations in audio recordings. The limitations and disadvantages of speech synthesis systems are discussed, with references to generative models like WaveNet and applications in creating adataset for audio deepfake detection.

3 PROPOSED WORK

A) System Architecture

The proposed system architecture is shown in Fig 1.

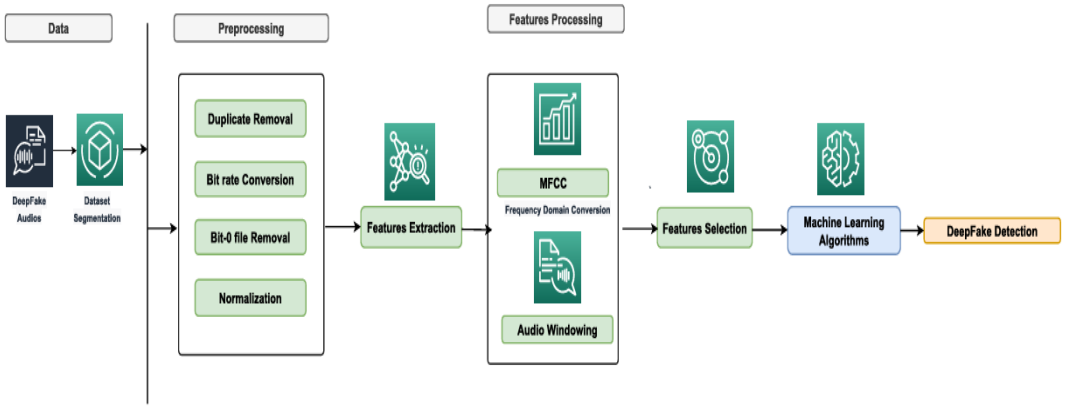

Figure 1: System Architecture

Proposed work

The proposed work is to develop an effective framework for the early identification of Detecting Deepfake Audio utilizing different AI methods. The paper plans to assess the exhibition of various ML algorithms like AdaBoost, Random Forest, Decision Tree Algorithm, K-Nearest Neighbors method, Gaussian Naïve Bayes, etc. Also, unique component scaling strategies on four standard ASD datasets (Babies, Young people, Kids, and Grown-ups). The paper additionally utilizes four different Elements of Choice Strategies/Attribute Evaluators: Info Gain, Gain Ratio, Relief F, Correlation Attribute Evaluator to rank the most important attributes.

B) Dataset Collection

Deepfake audio recognition utilizing MFCC elements and ML. Here's a data description for the specified dataset and processing steps:

FOR-REREC DATASET

This dataset is named FOR-REREC and is specifically collected for deepfake audio detection purposes. It likely contains recordings that have been re-recorded or manipulated to simulate different acoustic environments or conditions.

FOR-2SEC

All audio examples in the dataset are handled to have a standardized duration of 2 seconds. This uniform duration ensures consistency in feature extraction and model training.

FOR-NORM

The feature extraction process involves using MFCCs to address the sound signals. The extracted MFCC features are normalized (FOR-NORM) to bring them to a common scale, reducing the impact of variations in amplitude and intensity across different audio samples.

FOR-ORIGINAL

This category refers to the subset of the dataset that contains the original, unaltered audio recordings. These samples serve as the baseline or genuine instances against which manipulated or deepfake audio can be compared. The inclusion of original data is crucial for training the machine learning model to distinguish between genuine and manipulated audio.

C) Pre-processing

In the preprocessing phase (Fig 2)for deepfake audio detection, we employ a multi-faceted approach to enhance the robustness of our model. Initially, Exploratory Data Analysis /(EDA) is conducted to visualize the data distribution, aiding in understanding potential patterns and variations. To augment the dataset and improve model generalization, we introduce noise, stretch, shift, and pitch variations to the audio samples. These augmentations help the model better adapt to diverse scenarios and variations that may be encountered in real-world situations. To capture essential sound features, MFCCs are taken using Standard Scaler. MFCCs are particularly effective in representing the spectral characteristics of audio signals.

This comprehensive strategy sets the stage for our experimental investigation, where Convolutional Neural Network (CNN) models excel in discerning deepfake audio within specific sub-datasets, contributing to the advancement of deepfake detection in the realm of audio manipulation.

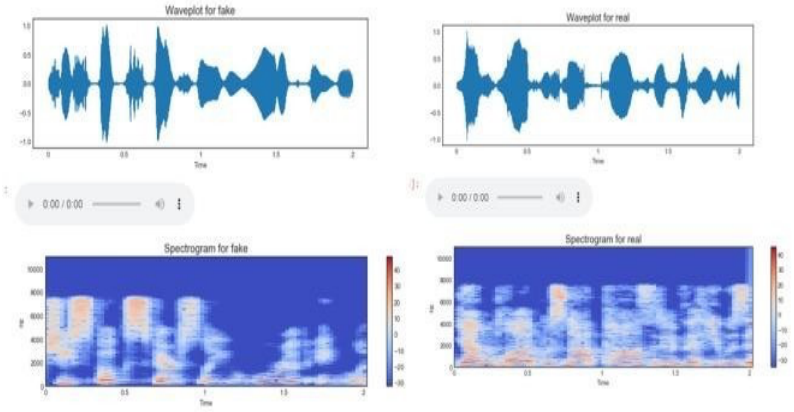

Figure 2: F1 Data Processing

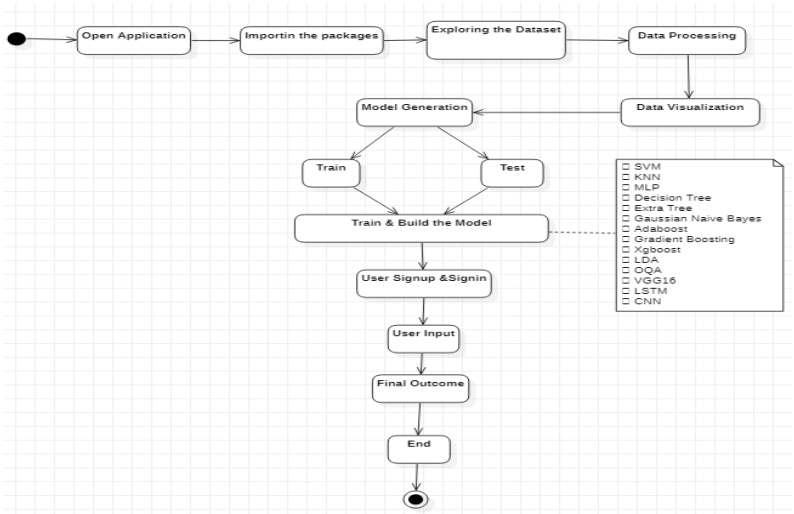

Figure 3: Flow Process of the proposed work

The simple procedure is to train datasets by using algorithms with different datasets and the outcome is displayed at the front end. Flow Process of the proposed work is shown in Figure 3.

D) Training & Testing

This phase uses different AI and ML Algorithms to complete the task:

Traditional Machine Learning Models such as SVM, KNN, MLP, Decision Tree, Extra Tree, Gaussian NaiveBayes, Adaboost, Gradient Boosting, XGBoost, LDA and OQA are employed here. Also, deep learning models such as VGG16, LSTM, and CNN are used.

4 Experimental Results

The Results show how important it is to use the Convolutional neural network algorithm of deep learning which is helpful to find the accuracy of 100% for datasets: for-rerec, for-2sec, for-norm, for-original. As we want to get full accuracy the other algorithms haven't shown as much impact as CNN and it delivered as we needed.

Overall, we can say that CNN shows the full results for datasets: for-rerec, for-2sec, for-norm, for-original. Below are the tables of performance evaluation. It consists of metrics like accuracy, precision, recall, and F1-score.

A) Performance Evaluation table

Accuracy: It is defined as its ability to recognize debilitated and solid examples precisely.

$$Accuracy = TP+TN/(TP+TN+FP+FN)$$

Precision: It is one mark of an ML model's performance- the nature of a positive forecast made by the model.

$$Precision = True\ positives/\ (True\ positives\ +False\ positives) = TP/(TP + FP)$$

Recall: Recall is a machine-learning metric that surveys a model's capacity to recognize all pertinent examples of aparticular class.

$$Recall = TP/(TP+FN)$$

F1-Score: It is an assessment estimation that assesses the accuracy of a model. It merges a model's accuracy and survey scores.

$$F1\ Score = (2 * Precision * Recall)/(Precision + Recall)$$

Performance metrics achieved using different datasets are shown in Table 1 – Table 4.

Algorithm	Accuracy	Recall	Precision	F1
SVM	0.821	0.821	0.833	0.820
MLP	0.967	0.967	0.968	NaN
Decision Tree	0.983	0.983	0.983	NaN
Extra Tree	0.895	0.895	0.896	1.000
Logistic Regression	0.979	0.979	0.979	0.979
Gaussian Naïve Bayes	0.749	0.749	0.756	0.749
Adaboost	0.820	0.820	0.824	0.820
Gradient Boosting	0.861	0.861	0.862	0.861
Xgboost	0.979	0.979	0.979	0.979
LDA	0.816	0.816	0.822	0.816
QDA	0.882	0.882	0.885	0.882
VGG16	0.504	NaN	NaN	NaN
LSTM	0.504	NaN	NaN	NaN
CNN	1.000	1.000	1.000	1.000

Table 1: Metrics for FOR-REREC dataset

Algorithm	Accuracy	Recall	Precision	F1
SVM	0.866	0.866	0.871	0.866
MLP	0.976	0.976	0.977	NaN
Decision Tree	0.987	0.987	0.987	NaN
Extra Tree	0.922	0.922	0.922	1.000
Logistic Regression	0.988	0.988	0.988	0.988
Gaussian Naïve Bayes	0.754	0.754	0.754	0.753
Adaboost	0.859	0.859	0.861	0.859
Gradient Boosting	0.885	0.885	0.885	0.885
Xgboost	0.988	0.988	0.988	0.998
LDA	0.836	0.836	0.839	0.836
QDA	0.936	0.936	0.936	0.936
VGG16	0.504	NaN	NaN	NaN
LSTM	0.504	NaN	NaN	NaN
CNN	1.000	1.000	1.000	1.000

Table 2: Metrics for FOR-2SEC dataset

Algorithm	Accuracy	Recall	Precision	F1
SVM	0.869	0.869	0.872	0.869
MLP	0.976	0.976	0.976	NaN
Decision Tree	0.985	0.985	0.985	NaN
Extra Tree	0.930	0.930	0.930	1.000
Logistic Regression	0.983	0.983	0.983	0.983
Gaussian Naïve Bayes	0.804	0.804	0.804	0.804
Adaboost	0.868	0.868	0.868	0.868
Gradient Boosting	0.893	0.893	0.893	0.893
Xgboost	0.983	0.983	0.983	0.983
LDA	0.856	0.856	0.859	0.856
QDA	0.956	0.956	0.956	0.956
VGG16	0.504	NaN	NaN	NaN
LSTM	0.504	NaN	NaN	NaN
CNN	1.000	1.000	1.000	1.000

Table 3: Metrics for FOR-NORM dataset

Algorithm	Accuracy	Recall	Precision	F1
SVM	0.866	0.866	0.871	0.866
MLP	0.976	0.976	0.977	NaN
Decision Tree	0.987	0.987	0.987	NaN
Extra Tree	0.922	0.922	0.922	1.000
Logistic Regression	0.988	0.988	0.988	0.988
Gaussian Naïve Bayes	0.754	0.754	0.754	0.753
Adaboost	0.859	0.859	0.861	0.589
Gradient Boosting	0.885	0.885	0.885	0.885
Xgboost	0.988	0.988	0.988	0.988
LDA	0.836	0.836	0.839	0.836
QDA	0.936	0.936	0.936	0.936
VGG16	0.504	NaN	NaN	NaN
LSTM	0.504	NaN	NaN	NaN
CNN	1.000	1.000	1.000	1.000

Table 4: Metrics for FOR-ORIGINAL dataset

B) Accuracy comparison for all datasets

We choose accuracy as our main metric to compare the algorithm's performance (Fig 4 – Fig 7).

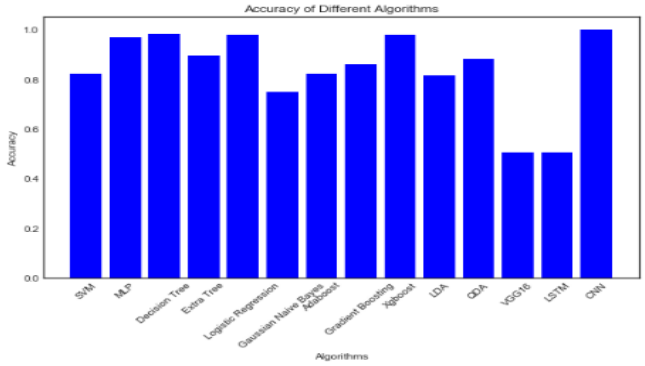

Figure 4: Accuracy graph for FOR-REREC dataset

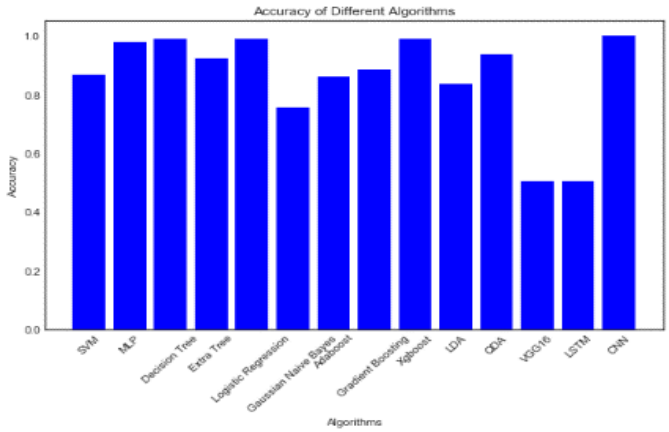

Figure 5: Accuracy graph for FOR-2SEC dataset

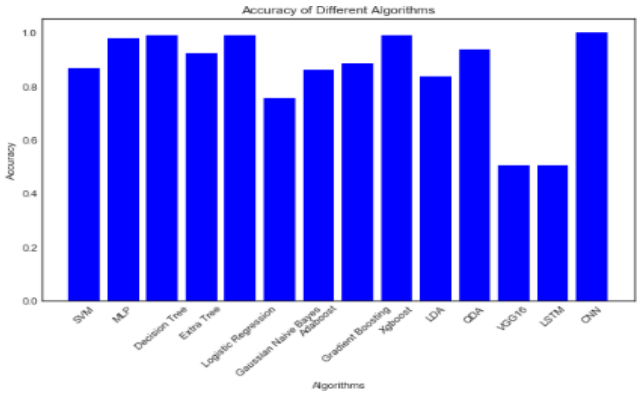

Figure 6: Accuracy graph for FOR-NORM dataset

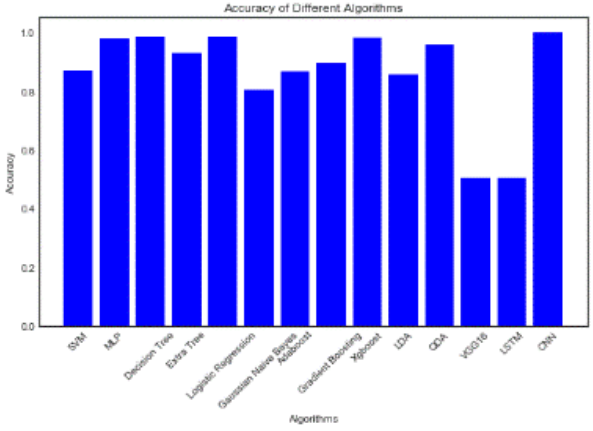

Figure 7: Accuracy graph for FOR-ORIGINAL dataset

5 Conclusion

In conclusion, this study demonstrates the effectiveness of machine learning (ML) and deep learning-based approaches in detecting deepfake audio, a critical aspect in combating the challenges posed by fake content. By employing the Mel-frequency cepstral coefficients (MFCCs) technique for feature extraction, this research utilized the fake or genuine dataset, which serves as a new benchmark dataset encompassing varying audio lengths and bit rates. The experimental results underscore the superior performance of the Convolutional Neural Network (CNN) model across multiple sub-datasets, particularly excelling in the for-rece and for-2-sec categories. This achievement highlights the robustness of CNNs in identifying subtle patterns within audio data, surpassing other ML models in terms of precision. Notably, the Gradient Boosting model demonstrated impressive results in the for-norm dataset, illustrating the adaptability of ensemble methods in addressing specific characteristics of fake audio.

6 Future Scope

Future research will focus on refining and expanding the proposed deepfake detection framework to encompass emerging audio manipulation techniques. Investigating the integration of multi-modal approaches, combining audio and visual cues, could enhance overall detection accuracy. Additionally, exploring real-time implementation and scalability for large datasets will be crucial for practical deployment. Collaborations with industry stakeholders can facilitate the development of robust, real-world solutions. As the deepfake landscape evolves, continuous adaptation and improvement of detection models are imperative, requiring ongoing research efforts to address novel challenges and ensure the resilience of AI-driven methods against increasingly sophisticated synthetic content.

References

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

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

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

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

[5] A. R. Javed, F. Shahzad, S. U. Rehman, Y. B. Zikria, I. Razzak, Z. Jalil, and G. Xu.

(2022) ‘‘Future smart cities:Requirements, emerging technologies, applications, challenges, and future aspects,’’ Cities, vol. 129, Art. no. 103794.

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

[7] S. Anwar, M. O. Beg, K. Saleem, Z. Ahmed, A. R. Javed, and U. Tariq. (2022) ‘‘Social relationship analysis usingstate-of-the-art embeddings,’’ ACM Trans. Asian Low-Resource Lang. Inf. Process.

[8] C. Stupp. (2022) ‘‘Fraudsters used Ai to mimic CEO’s voice in unusual cybercrime case,’’ Wall Street J., vol. 30, no.8, pp. 1–2.

[9] 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. (2019) ‘‘Deep learning fordeepfakes creation and detection: A survey,’’ arXiv:1909.11573.