Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 219 (2023) 211–219

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2022

The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation

Mvelo Mcuba[a], Avinash Singh[a,]*, Richard Adeyemi Ikuesan[b], Hein Venter[a]

[a]*Department of Computer Science, University of Pretoria, Pretoria, South Africa*
[b]*Computing and Applied Technology, College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates*

Abstract

Voice cloning methods have been used in a range of ways, from customized speech interfaces for marketing to video games. Current voice cloning systems are smart enough to learn speech characteristics from a few samples and produce perceptually unrecognizable speech. These systems pose new protection and privacy risks to voice-driven interfaces. Fake audio has been used for malicious purposes and is difficult to classify what is real and fake during a digital forensic investigation. This paper reviews the issue of deep-fake audio classification and evaluates the current methods of deep-fake audio detection for forensic investigation. Audio file features were extracted and visually presented using MFCC, Mel-spectrum, Chromagram, and spectrogram representations to further study the differences. Harnessing the different deep learning techniques from existing literature were compared using five iterative tests to determine the mean accuracy and the effects thereof. The results showed a Custom Architecture gave better results for the Chromagram, Spectrogram, and Me-Spectrum images and the VGG-16 architecture gave the best results for the MFCC image feature. This paper contributes to further assisting forensic investigators in differentiating between synthetic and real voices.

Keywords: Deepfake audio; digital investigation; CNN, voice cloning.

* Corresponding author.
E-mail address: asingh@cs.up.ac.za

10.1016/j.procs.2023.01.283

1. Introduction

Although the advances of the current internet have revolutionized our daily lives, it still lacks the security to provide every user with secure access and protection from malicious attacks [1]. This also poses challenges for current biometric technology used for authentication [2]. The risk of using biometrics falls into a few categories such as data and network hacking, deep fake audio, spoofed sensors, and sensor inaccuracy [3]. Digital forensic experts need to be familiar with the advances in the latest technology to give them a competitive edge over attackers [4]. A recently revived controversy over the infallibility of certain traditional forensic techniques leads to new requirements in digital forensics [5]. The new voice biometrics system has drawbacks such as the tools used to replicate voices. Detecting such synthetic voices would enable the evidence to be admissible in a court of law, as long as the scientifically sound techniques utilize standard protocols and show their research capacity, precision, and acceptance by the academic community is likely to be accepted to the court of law [5]. There is an inherent lack of testing and techniques to identify voice deep fakes successfully [6]. A deep fake voice could lead to cybercrimes such as scamming and exploitation of confidential data due to the little research on the topic and only a few effective techniques being presented since it is an emerging technology that is constantly fluctuating. The digital forensic discipline, particularly multimedia forensics, is saddled with the responsibility of investigating the authenticity (or otherwise) of a given media file [7]. A major component of digital forensics is the analysis process. For forensic examiners to effectively analyze the authenticity of a given fake audio multimedia file, particularly, deep fakes that leverage advanced machine learning algorithms to generate a fake audio component, a critical analysis of the component of such audio file is required. This study provides the step towards actualizing this process. to the best of the authors knowledge, this is the first study to explore the technical nitty gritty of a deep fake generated audio file for forensic purposes. This paper evaluates existing techniques to detect deep fake audio using deep learning to help digital forensic investigators identify voice cloning or deep fake audio for evidence collection. This paper aims to compare existing deep learning models and utilize various pre-processing techniques to decipher how to aid an investigator with deep fake detection.

2. Background and related works

2.1. Deep fake audio and CNN architecture

An image may depict a thousand words; however, images fall short in comparison to an audio or video clip of an event [8]. Recording audio and video allow people to become first-hand observers of an event, without the need to believe what another person witnessed as it may be subjective to what the person interpreted [8]. Developments in technology will soon bring about the nightmare of misuse. Due to the development of "deep fakes," the visual manipulation of audio or video is highly convincing and impossible to detect. It is easier than ever to show an individual saying or doing what they have never said or done [8]. A recent example of this was observed in the recent Russia and Ukraine war where there was a deep fake video of President Zelenskyy telling soldiers to give up arms and surrender [9].

Audio deep fakes, scientifically known as logical-access audio spoofing methods, have become an enhanced challenge to voice interfaces due to rapid breakthroughs in speech recognition and voice transfer technology [10]. As new forms of speech synthesis and voice conversion technologies are evolving, the potential to generalize countermeasures is becoming an increasingly critical challenge [10]. While the manipulation of visual and auditory information is as old as the internet on its own, the recent introduction of deep fakes marked a turning point in the development of fake material [11]. Deep fakes deliver automated procedures to create false information that is more difficult for human analysts to spot [11]. An indication of a significant deep fake event was a video of former President Obama in 2019, where he was cursing during a public service announcement [11].

2.2. Developments of Deep fakes

The developments of deep fakes are based on two machine learning progress: neural networks and [12] generative adversarial networks (GANs). Neural networks represent how the human brain works, and they demonstrate how the brain processes information. The faster and more precisely the brain can replicate the more the human brain is introduced to representations of something, such as how to catch a hit cricket ball or harmony for a new song. The more occurrences introduced in the network, the more accurately a new example will be created from scratch [12]. This means that, as more video or audio data is sent to the neural networks, the new, false, audio, or video would be more reliable and credible. Deep fakes would not have been as credible without GANs as GANs consist of two networks: the generator, and discriminator network [13]. By trying to reproduce the dataset that is being fed as input, the first neural network, referred to as the generator, is programmed to generate a new, false video or audio [12], [13]. A second neural network, known as the discriminator, is then fed into all the original datasets and the newly generated deep fake [12], [13]. The discriminator's role is to determine which videos or audio are legitimate in the data set (which now contains a deep fake). Recent work by Jemine et al [14] has implemented a three-stage pipeline that enables a voice to be cloned from a few seconds of reference speech during training, without retraining the model.

2.3. Deep Convolutional Neural Networks

CNN's architecture was influenced by the human brain's visual cortex structure [15]. Essentially, a deep learning algorithm that takes images or spectrograms and assigns different weights to each image for distinction, and performs any given task, such as image classification. In contrast to hard-coded basic techniques, it is possible to train them to know the necessary filters for proper preparation [15]. Throughout the years, a variety of architectural developments have been made to resolve concerns related to computer efficiency, error rate, and further changes in the domain [15]. LeNet-5 was the first "popular" CNN architecture developed by LeCun et al. [16] for the recognition of handwritten numbers. For ten years, LeCun and his research team have been working on CNN models to develop an effective architecture. Krizhevsky et al. [17] suggested a network called AlexNet that is like LeNet-5 but deeper and wider with more hidden layers. AlexNet consists of eight layers, five fully convolutional layers with 11 × 11 receptive filters, and three fully connected layers with 60 million parameters. The high degree of parametrization, and thus representational ability, makes the network susceptible to over-fitting in the conventional context of machine learning. Rectified Linear Units (ReLu) is another innovative solution to this work used as an activation function for non-linearity. ReLu produces a dataset 6 times faster than CNN using the tanh activation function [18]. After the success of this work, Zeiler et al. [18] developed a related architecture with smaller receptive fields, known as ZFNet. The ZFNet architecture was an enhanced version of AlexNet which was popular as it was followed by a deeper understanding of how CNNs function internally [18]. ZFNet also incorporates a new approach to visualizing the representation of features within the network, a technique that has been established by a deconvolutional network.

For several other CNN architectures based on the same principles, such as GoogLeNet [19], Inception networks have paved the way for many design changes. Instead of fully linked layers, GoogLeNet uses global average pooling, with 1x1 convolution at the core of the network [19]. Beating GoogLeNet with results of up to 92.7% accuracy, the next advancement in the fields of deep learning and computer vision was VGG-16 [20]. Around the time, it was the top-performing model and sparked further work into deep CNNs. The significance of network depth on classification accuracy was analyzed by Simonyan and Zisserman [20] by piling fully convolutional layers with small 3x3 receptive fields with a slide of 1. Not only does it improve non-linearity with the use of small receptive filters, but it also decreases the total number of network parameters.

In 2015, following Simonyan and Zisserman's work [20], He et al. [21] suggested a basic but efficient network called ResNet. This approach to the architecture greatly decreases the number of parameters needed for a deep network and exceeds the previous state-of-the-art in terms of accuracy [21]. ResNet has introduced shortcut connections and identity mapping that skips one or more layers. The suggestion is that instead of piling additional layers onto the network, alternatively they are added as residual blocks (with identity mappings). To achieve results in this scheme, the authors of the work tweaked the underlying mapping and have the non-linear layers learn mapping instead of conventional mapping [21]. ResNet quickly became one of the most popular architectures in various computer vision tasks due to its compelling results [18], [21]. ResNets has grown, its architecture has been heavily studied, and researchers have come up with different variants of the work originally proposed, such as

ResNeXt and DenseNet [18]. More researchers are finding ways to develop the current architecture, which will increase the efficiency and accuracy of the models.

3. Deepfake detection

To assist digital forensic investigators with deep fake detection and the effects of existing detection mechanisms a criterion was built and developed to compare different types of CNN architectures used by several authors for voice recognition, image classification, and synthetic speech detection. This criterion consists of eleven parts that describe the layers used for each architecture in each article. This criterion includes the input size, the convolution layers, the filters and strides, the pooling, the pooling step, the architecture receptive fields, the activation function, the padding, the drop-out regularization, the accuracy, and the type of architecture.

Usually, the CNN hidden layers consist of a sequence of convolutional layers converging with a dot product or matrix multiplication. The activation function is normally a softmax or a RELU layer and is eventually accompanied by subsequent convolutions, such as normalization layers, fully connected layers, and pooling layers referred to as hidden layers. Although the layers are referred to colloquially as convolutions, this is by convention only. It is technically a moving dot product or cross-correlation, mathematically. For the indices in the matrix, this has importance, in that it influences how weight is calculated at a given index point. In a CNN the input layer is represented by an input image whereby a computer sees an input image as an array of pixel values, which are defined as a volume matrix (height x width x depth). Depending on the resolution and size of the image, for instance, it will see a 64 x 64 x 3 array of numbers.

A cloned audio dataset from Baidu Silicon Valley AI Lab is used for performance assessment. This dataset can be downloaded from https:/audiodemos.github.io. The data collection consists of 10 audio samples of the original audio, 120 cloned speech, and four altered speech recordings. Then section three accounts for the conceptual design and implementation of the proposition. A list of CNN architectures relevant to the task at hand listed in the literature has been compiled and consists of the work by Malik et al [22], Wu et al [23], Chugh et al [24], Thai et al [25], and Reimao et al [26]. The performance and the parameters used in each of these studies are represented in Table 1.

Table 1. Performance of CNNs for voice cloning.

Parameters	Malik et al [22]	Wu et al [23]	Chugh et al [24]	Thai et al [25]	Reimao et al [26]
Input Size (pixels)	625x469x3	863x256	-	-	-
Convolutional Layers	4	10	9	7	16
Filters / Stride	3x3	4x4	3x3	-	-
Pooling	Max (2x2)	Max	Max	Max	Max
Stride For Pooling	2	2	1x1 & 2x2	2	2
Activation Function	Softmax	Leaky ReLu & BatchNorm	BatchNorm	BatchNorm & log softmax	Softmax & ReLu
Types of Layers	Fully Connected	Fully Connected	Convolutional	Fully Connected with LSTM	Fully Connected
Padding	No	Yes	No	No	No
Dropout	Yes	No	Yes	Yes	No
Accuracy (%)	100%	95.93%	50%	91.91%	99.94%
Architecture	Custom	FG-LCNN	ResNet	CRNN/CNN	VGG16/19

From the performance highlighted in Table 1, the worst performing network was Chugh et al [24] which proposed a bimodal deepfake detection approach based on the modality dissonance score which captures the similarity between audio and visual streams for real and fake videos thereby facilitating separability. Chugh et al [24] based the audio stream architecture on convolutional neural networks designed for image recognition.

Reimao et al [26] also had a similar idea on how to differentiate between synthetic and real speech. The deep learning models consist of extracted and converted audio features from each audio file using CQT, short-time Fourier transform (stft), MFCC, and Mel-Spectrograms. The resulting images were then used for training selected architectures of pre-trained deep learning, such as VGG16 and VGG19. The results from the deep learning analysis showed that the VGG16 and VGG19 models using the STFT audio representation presented the highest validation accuracy of 99.96% and 99.94% respectively [26]. The difference between Chugh et al [24] and Reimao et al [26] models is that Reimao et al [26] added a ReLu activation function for all the convolutional layers than a softmax activation function for the last convolutional layer. Instead of the use of dropout regularization, Reimao et al [26] used three dense layers after the fully convoluted layer to feed all outputs from the previous layer to all its neurons.

Thai et al [25]and Wu et al [23] papers did not show accuracy but displayed the Equal Error Rate (EER) and the Tandem detection cost function (t-DCF). For both papers the EER ranged between 0.000 – 6.02% and t-DCF ranged between 0.000 – 0.3% [25]. Thai et al [25] experimented with two models: a convolution-recurrent neural network and a fully convolutional neural network. Each model is evaluated on the effectiveness of detecting spoofed audio from a genuine human speech as a standalone system using an equal error rate. The distinct feature which Thai et al [25] used include a wide block in the models and the use of bidirectional LSTM. A Wide Block contains multiple paths, each with a different convolution kernel size, to learn different ranges of temporal dependencies. Wu et al [23] proposed a genuinization transformer that consists of two functionalities: encoding and decoding. During the encoding phase, the input signal is compressed through five stridden convolutional layers, and then the convolution result is obtained by leaky ReLU. In the decoding phase, the encoding process is reversed by deconvolution, and then by ReLU [23]. Even though feature genuinization is based on an LCNN system, the dropout regularization was not used but max-pooling (2x2) and padding were included in the model.

Malik et al [22] designed and developed a nonparametric and fully supervised CNN model to perform speech classification. For the convolutional layer a 3×3 kernel, with a kernel count of 32 is used [22]. An asymmetric window of size 2×2 is used for the pooling layer. The audio dataset had 248 images each of size 625×469×3 pixels. A stack of four layers is used to reduce each input image to a fully connected layer of size 512. A softmax activation function was applied on a fully connected layer to get the probability distribution vector of class labels over the categorical cross-entropy loss function. The proposed method achieved 100% accuracy for the predictions on the test dataset [22]. Out of all the models proposed by these different studies, Malik et al [22] had an accuracy of 100% surpassing the other research papers, as highlighted in Table 1. However, the obtained accuracy was based on a custom architecture that may not apply to a different context. Deepfakes can be designed with different contexts and processes such that one deepfake algorithm may be architecturally different from another. Thus, whilst the need to develop a customized architecture may be important, there is a greater need to develop/use a detection algorithm based on standardized architecture for experimental repeatability and usability. Furthermore, depending on such a generic architecture provides a forensic justification when such techniques are used in audio forensics. This study conducted a comparative analysis of these existing architecture (including the custom architecture) on a public dataset to further evaluate their applicability in a forensic context. Through the experimental process, a uniform platform was provided to evaluate the reliability of the different architecture on same dataset. Detail of this process is presented in the next section.

4. Experimentation

4.1. Dataset

The Baidu Silicon Valley AI Lab dataset is arranged according to its source in folders, and the folders are split into two: the original audio and the deep fake audio. For the training of the first dataset, 84 VCTK speakers (48 kHz sampling rate) were used; voice cloning was conducted on other VCTK speakers (48 kHz sampling rate). For the second dataset, voice audio was trained on LibriSpeech speakers (16 kHz sampling rate), and voice cloning was performed on VCTK speakers (downsampled to 16 kHz sampling rate). The average length of the cloning sample for the entire dataset is 3.7 seconds. As is standard practice in machine-learning research, both pre-processed versions of the dataset have been divided into 80/20 split for training and validation.

4.2. Preprocessing

Now that the audio samples have been grouped, the next step is to distinguish the effects of the different representations of audio samples to images to see which architecture and machine learning algorithm performs the best on all types of information. Various models were developed based on the criteria in Table 1. As seen in Fig. 1, the deep learning models consist of extracted audio features using the Libros Python library to extract the Mel-Spectrograms, MFCC, Spectrogram, and Chromagram for each audio file, then transformed them into PNG format images from each audio file. The resulting images were then used to train the chosen deep learning models, such as FG-LCNN [23], ResNet [24], VGG-16 [26], and a Custom Architecture [22] model proposed by Malik et al [22]. In the training and testing stages, different optimizers were used such as Adam (learning rate of 0.001), SGD (learning rate of 0.01), and Adadelta (learning rate of 0.001) to see which, one gives the best learning rate and the best accuracy for the models. An iteration of five testing trials was used to give the average results of the accuracy since the spiting of the training and validation set were randomized every time the models were tested. The testing system that was used comprised of an AMD Ryzen 7 3700U with Radeon Vega GFX 2.3Ghz processor and 32GB of RAM.

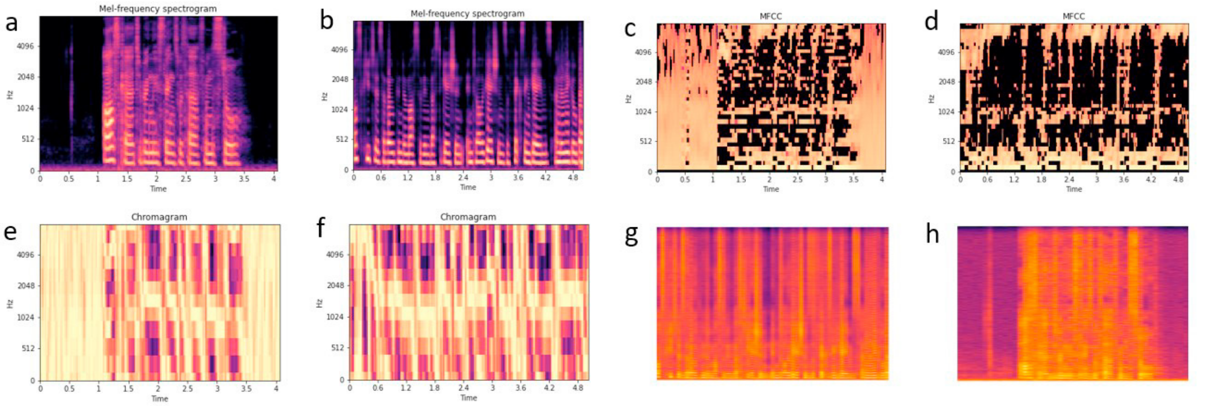

Fig. 1. (a) original Mel-freq audio, (b) cloned Mel-freq audio, (c) original MFCC audio, (d) cloned MFCC audio, (e) original chromagram audio, (f) cloned chromagram audio, (g) original spectrogram audio, (h) cloned spectrogram audio.

4.3. Results and Discussion

In Tables 2-4 the Custom Architecture by Malik et al [22] generates a relatively higher accuracy for the SGD Optimizer. In Table 2, the performance of the custom architecture demonstrated a higher performance of chromagram audio distinction. Using the SDG for instance, the performance of the custom architect demonstrates a relatively significant performance. However, the lowest performance as observed in Table 2 can be attributed to the Adadelta optimizer for the chromagram images. The learning rate used for the SGD was 0.01 and for both Adam and Adadelta optimizer learning rate of 0.001. All these learning rates gave better accuracy for the models with the data that was used. The custom architecture gave the best results for all the audio feature extractors simply because of how the model is built. The model is lightweight and composed of four stacks of convolution, pooling, and activation layers of fundamental CNN building blocks. For convolution operations, each layer implements a 2D convolution function, a max-pooling method for data set size reduction, and a rectified linear unit (ReLu) function for non-linear activation. At the top, the stack of four layers is bound by a completely connected layer of predefined thickness. The soft-max activation feature and dropout with a rate of 0.1 is used at the end of the fully connected layer. This model has been able to produce the best outcomes compared to other ones.

Furthermore, in Table 3, the VGG-16 model with an accuracy of 68.636% makes it the second model to give a relatively higher accuracy by using the Adadelta optimizer with a learning rate of 0.001. As it does not use a diverse range of hyperparameters, the VGG-16 architecture can be referred to as a basic model. It often uses 3 x 3 filters with one step in the convolution layer and uses SAME padding with two steps in the 2 x 2 pooling layer. This model is also composed of 16 stacks of convolutions with a max-pooling method every after 2 stacks of convolution layers for the first 2 sets then, the layer stacks were incremented by another layer making it every after 3 layers the max pool method is implemented. The VGG-16 model was the runner-up for Tables 2-4 for all the audio feature extractors (MFCC, Mel-spectrum, and spectogram). Similar observations can be deduced from Table 4 about the custom and the VGG-16 architecture. However, in Table 5 the VGG-16 presents the best-performed architecture for distinguishing an original and cloned MFCC audio feature with an accuracy of 86.906%, outperforming the custom architecture by Malik et al [22], and other architectures.

Table 2. Model performance from chromagram images.

Architecture	Adam (%)	SDG (%)	Adadelta (%)
FG-LCNN	44.096	55.454	54.542
RestNet	51.364	47.272	47.272
VGG-16	54.552	44.090	43.182
Custom	67.726	83.636	56.366

Table 3. Model performance from spectrogram images.

Architecture	Adam (%)	SDG (%)	Adadelta (%)
FG-LCNN	65.000	54.088	60.764
RestNet	49.400	53.634	49.090
VGG-16	61.364	60.908	68.636
Custom	72.270	50.000	37.274

Table 4. Model performance from Mel-Spectrum images.

Architecture	Adam (%)	SDG (%)	Adadelta (%)
FG-LCNN	59.546	58.636	40.912
RestNet	51.364	49.544	44.546
VGG-16	42.728	61.368	53.182
Custom	71.362	72.044	48.186

Table 5. Model performance from MFCC images.

Architecture	Adam (%)	SDG (%)	Adadelta (%)
FG-LCNN	64.544	51.364	69.544
RestNet	52.272	53.636	56.302
VGG-16	42.728	71.362	85.906
Custom	46.814	45.000	39.082

Given the performance of VGG-16, RestNet, and FG-LCNN from the MFCC images in Table 5, it is safe to assert that a custom architecture is context-dependent. A similar observation can be deduced from Table 3 on the Adadelta optimizer with a learning rate of 0.001. Consequently, diverse custom architectures might be required for a different context. This is important from a forensic point of view. Deepfakes are a major challenge in forensic investigations. However, to scale through forensic scrutiny, some fundamental forensic requirements should be satisfied. Although to further state, the current performance is below the forensic standard. However, they present a substratum for the development of a more forensically suitable and efficiently reliable architecture for deepfake audio detection and analysis. In tandem with the trends in multimedia forensics and investigation, the growing sophistication of the Deepfake technique requires a directed effort toward an effective forensic analysis approach. Such an approach will be required to satisfy the technical, legal, and ethical constraints often associated with forensic analysis. The current study, therefore, is the right step in this direction.

5. Conclusion

This paper proposed a metric of comparisons for various deep learning approaches that detect deep fake audio and for the implementation the model architectures were taken from papers by Malik et al [22], Wu et al [23], Chugh et al [24], Thai et al [25], and Reimao et al [26]. From the experiment carried out, an average of the precision of five test iterations was conducted using four visual representations of the frequency spectrum (Mel-Spectrograms, MFCC, Spectrogram, and Chromagram). Although the data set used to test the deep learning models was not sufficient, in the future the authors intend to explore approaches to improve the accuracy by collecting more datasets to test the models created. Furthermore, aspect of the deep learning model that requires tweaking will be considered, as the default parameter might not be effective for all contexts. For the instance, the deployment of the softmax activation function as opposed to Max and ArgMax is considered in this current study. However, the exploration of both Max and ArgMax remains unexplored. Similarly, fuzzification process will be explored to provide a comprehensive exploratory process for fake multimedia forensics. As speech synthesis improves, there is also a need for an up-to-date synthetic speech dataset that can be used in synthetic speech recognition experiments. An improvement would be critical, as synthetic voice recognition is a massive issue as TTS systems gain human rawness that can be used for imitation. From this research it was observed that Mel-Spectrum frequency for visualization provides the most accurate results from all the optimizers and architectures explored. While the results provide a good indication of which architectures are the best as well as the optimizer, more work will be carried out to generating more robust and featureful datasets, such that more effects can be observed, which can further validate the algorithms. The result presented in this study can be leveraged for multimedia forensic analysis, to guide forensic analysts in choosing the appropriate approach for deepfake audio analysis. This approach can also be applied to triage deepfake video during video analysis, during video forensics.

References

[1] Z. A. Baig *et al.*, “Future challenges for smart cities: Cyber-security and digital forensics,” *Digit. Investig.*, vol. 22, pp. 3–13, Sep. 2017.

[2] H. Zimmerman, “The data of you: Regulating private industry’s collection of biometric information,” *U. Kan. L. Rev.*, vol. 66, p. 637, 2017.

[3] A. K. Jain and A. Kumar, “Biometric recognition: an overview,” in *Second generation biometrics: The ethical, legal and social context*, Springer, 2012, pp. 49–79.

[4] D. Lillis, B. A. Becker, T. O. Sullivan, and M. Scanlon, “Current Challenges and Future Research Areas for Digital Forensic Investigation INVESTIGATION,” no. c, 2016.

[5] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, “Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework,” in *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*, 2006, pp. 1–8.

[6] A. Saleema and S. M. Thampi, “Voice Biometrics: The Promising Future of Authentication in the Internet of Things,” in *Handbook of Research on Cloud and Fog Computing Infrastructures for Data Science*, IGI Global, 2018, pp. 360–389.

[7] B. Zawali, R. A. Ikuesan, V. R. Kebande, S. Furnell, and A. A-Dhaqm, “Realising a Push Button Modality for Video-Based Forensics,” *Infrastructures*, vol. 6, no. 4, p. 54, 2021.

[8] R. Chesney and D. Citron, “Deepfakes and the new disinformation war: The coming age of post-truth geopolitics,” *Foreign Aff.*, vol. 98, p. 147, 2019.

[9] T. Simonite, “A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be,” *Wired*, 2022. .

[10] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, “Generalization of Audio Deepfake Detection,” in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 132–137.

[11] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, “Deepfakes: Trick or treat?,” *Bus. Horiz.*, vol. 63, no. 2, pp. 135–146, 2020.

[12] S. Dack, “Deep fakes, fake news, and what comes next.” Retrieved from The Henry M. Jackson School of International Studies: https~…, 2019.

[13] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” *arXiv Prepr. arXiv1710.10196*, 2017.

[14] C. Jemine and others, “Master thesis: Automatic Multispeaker Voice Cloning,” 2019.

[15] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, “A survey of deep neural network architectures and their applications,” *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[16] G. Wei, G. Li, J. Zhao, and A. He, “Development of a LeNet-5 gas identification CNN structure for electronic noses,” *Sensors*, vol. 19, no. 1, p. 217, 2019.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[18] S. Akcay, M. E. Kundegorski, C. G. Willcocks, and T. P. Breckon, “Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery,” *IEEE Trans. Inf. forensics Secur.*, vol. 13, no. 9, pp. 2203–2215, 2018.

[19] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[20] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[22] H. Malik and R. Changalvala, “Fighting AI with AI: Fake Speech Detection Using Deep Learning,” in *Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics*, 2019.

[23] Z. Wu, R. K. Das, J. Yang, and H. Li, “Light convolutional neural network with feature genuinization for detection of synthetic speech attacks,” *arXiv Prepr. arXiv2009.09637*, 2020.

[24] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, “Not made for each other-Audio-Visual Dissonance-based Deepfake Detection and Localization,” *arXiv Prepr. arXiv2005.14405*, 2020.

[25] B. Thai, “Deepfake detection and low-resource language speech recognition using deep learning,” 2019.

[26] R. Reimao and V. Tzerpos, “FoR: A Dataset for Synthetic Speech Detection,” in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2019, pp. 1–10.