algorithms

MDPI

Review

A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions

Zaynab Almutairi [1,*] and Hebah Elgibreen [1,2]

[1] Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh P.O. Box 145111, Saudi Arabia; hjibreen@ksu.edu.sa
[2] Artificial Intelligence Center of Advanced Studies (Thakaa), King Saud University, Riyadh P.O. Box 145111, Saudi Arabia
* Correspondence: 442202923@student.ksu.edu.sa

Abstract: A number of AI-generated tools are used today to clone human voices, leading to a new technology known as Audio Deepfakes (ADs). Despite being introduced to enhance human lives as audiobooks, ADs have been used to disrupt public safety. ADs have thus recently come to the attention of researchers, with Machine Learning (ML) and Deep Learning (DL) methods being developed to detect them. In this article, a review of existing AD detection methods was conducted, along with a comparative description of the available faked audio datasets. The article introduces types of AD attacks and then outlines and analyzes the detection methods and datasets for imitation- and synthetic-based Deepfakes. To the best of the authors' knowledge, this is the first review targeting imitated and synthetically generated audio detection methods. The similarities and differences of AD detection methods are summarized by providing a quantitative comparison that finds that the method type affects the performance more than the audio features themselves, in which a substantial tradeoff between the accuracy and scalability exists. Moreover, at the end of this article, the potential research directions and challenges of Deepfake detection methods are discussed to discover that, even though AD detection is an active area of research, further research is still needed to address the existing gaps. This article can be a starting point for researchers to understand the current state of the AD literature and investigate more robust detection models that can detect fakeness even if the target audio contains accented voices or real-world noises.

Keywords: Audio Deepfakes (ADs); Machine Learning (ML); Deep Learning (DL); imitated audio

Citation: Almutairi, Z.; Elgibreen, H. A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. *Algorithms* **2022**, *15*, 155. https://doi.org/10.3390/a15050155

Academic Editor: Theodore B. Trafalis

Received: 23 March 2022
Accepted: 1 May 2022
Published: 4 May 2022

1. Introduction

AI-synthesized tools have recently been developed with the ability to generate convincing voices [1]. However, while these tools were introduced to help people, they have also been used to spread disinformation around the world using audio [2], and their malicious use has led to fear of the "Audio Deepfake." Audio Deepfakes, recently called audio manipulations, are becoming widely accessible using simple mobile devices or personal PCs [3]. This has led to worldwide public cybersecurity concerns regarding the side effects of using AD. Regardless of the benefit of this technology, ADs go beyond a simple text message or an email link. People can use it as a logical-access voice spoofing technique [4], where it can be used to manipulate public opinion for propaganda, defamation, or even terrorism. Massive amounts of voice recordings are broadcast daily over the Internet, and detecting fakeness from them is a challenging task [5]. However, AD attackers have targeted not only individuals and organizations but also politicians and governments [6]. In 2019, fraudsters used AI-based software to impersonate a CEO's voice and swindled more than USD 243,000 via a telephone call [7]. For this reason, we need to authenticate any distributed audio recordings to avoid spreading disinformation. This problem has thus been of significant interest to the research community in recent years. Three

types of AD have emerged, increasing the challenge in detection; they are synthetic-based, imitation-based, and replay-based, as will be explained in the following section.

With regard to Deepfakes, many detection methods have been introduced to discern fake audio files from real speech. A number of ML and DL models have been developed that use different strategies to detect fake audio. The following strategies describe the AD detection process in general, as illustrated in Figure 1. First, each audio clip should be preprocessed and transformed into suitable audio features, such as Mel-spectrograms. These features are input into the detection model, which then performs the necessary operations, such as the training process. The output is fed into any fully connected layer with an activation function (for a nonlinearity task) to produce a prediction probability of class 0 as fake or class 1 as real. However, there is a trade-off between accuracy and computational complexity. Further work is therefore required to improve the performance of AD detection and overcome the gaps identified in the literature.

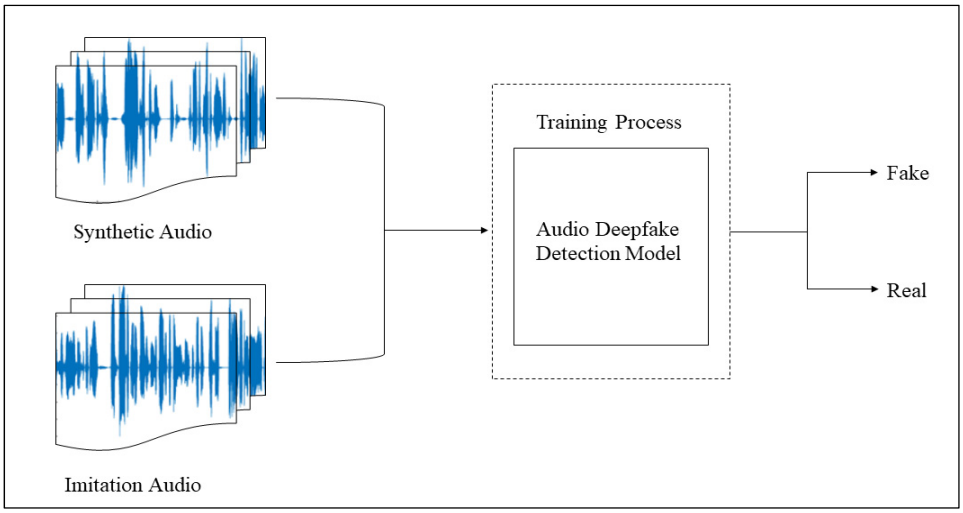

Figure 1. An illustration of the AD detection process.

AD detection has therefore become an active area of research with the development of advanced techniques and DL methods. However, with such advancements, current DL methods are struggling, and further investigation is necessary to understand what area of AD detection needs further development. Moreover, a comparative analysis of current methods is also important, and to the best of the authors' knowledge, a review of imitated and synthetically generated audio detection methods is missing from the literature. Thus, this article introduces the following significant contributions to the literature:

- A review of state-of-the-art AD detection methods that target imitated and synthetically generated voices;
- provision of a brief description of current AD datasets;
- a comparative analysis of existing methods and datasets to highlight the strengths and weaknesses of each AD detection family;
- a quantitative comparison of recent state-of-the-art AD detection methods; and
- a discussion of the challenges and potential future research directions in this area.

The rest of this article is organized as follows. An AD definition and its types are presented in Section 2. Section 3 discusses and summarizes the current methods developed for AD detection. Section 4 presents the generated audio dataset used for AD detection and highlights its characteristics. Section 5 presents a quantitative comparison of recent state-of-the-art AD detection methods. Section 6 presents the challenges involved in detecting AD and discusses potential future research directions for the detection methods. Finally, this article concludes with Section 7, which summarizes our findings.

2. Types of Audio Deepfake Attacks

AD technology is a recent invention that allows users to create audio clips that sound like specific people saying things they did not say [2]. This technology was initially developed for a variety of applications intended to improve human life, such as audiobooks, where it could be used to imitate soothing voices [8]. As defined from the AD literature, there are three main types of audio fakeness: imitation-based, synthetic-based, and replay-based Deepfakes.

Imitation-based Deepfakes are "a way of transforming speech (secret audio) so that it sounds like another speech (target audio) with the primary purpose of protecting the privacy of the secret audio" [3]. Voices can be imitated in different ways, for example, by using humans with similar voices who are able to imitate the original speaker. However, masking algorithms, such as Efficient Wavelet Mask (EWM), have been introduced to imitate audio and Deepfake speech. In particular, an original and target audio will be recorded with similar characteristics. Then, as illustrated in Figure 2, the signal of the original audio Figure 2a will be transformed to say the speech in the target audio in Figure 2b using an imitation generation method that will generate a new speech, shown in Figure 2c, which is the fake one. It is thus difficult for humans to discern between the fake and real audio generated by this method [3].

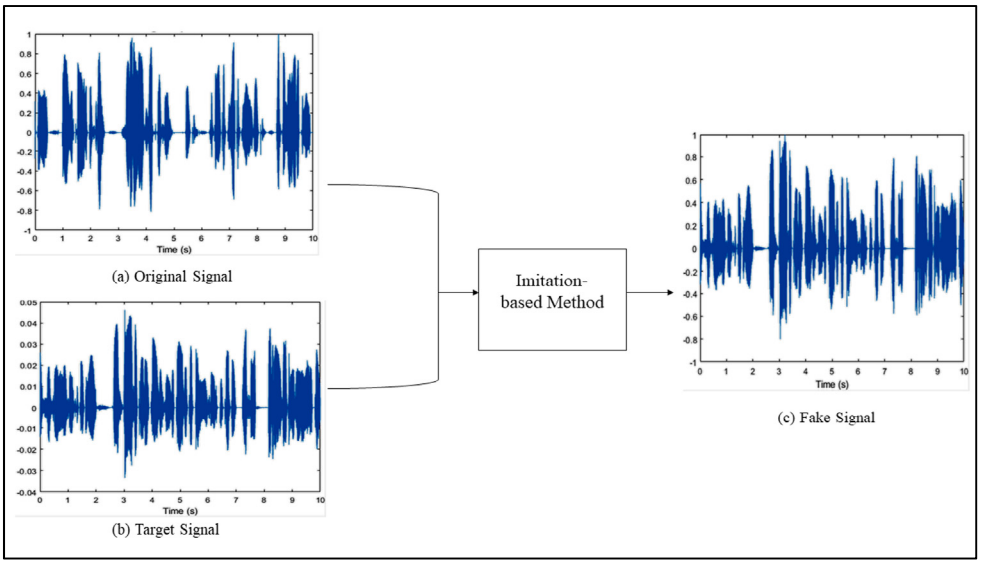

Figure 2. Imitation-based Deepfake.

Synthetic-based or Text-To-Speech (TTS) aims to transform text into acceptable and natural speech in real time [9] and consists of three modules: a text analysis model, an acoustic model, and a vocoder. To generate synthetic Deepfake audio, two crucial steps should be followed. First, clean and structured raw audio should be collected, with a transcript text of the audio speech. Second, the TTS model must be trained using the collected data to build a synthetic audio generation model. Tactoran 2, Deep Voice 3, and FastSpeech 2 are well-known model generation techniques and are able to produce the highest level of natural-sounding audio [10,11]. Tactoran 2 creates Mel-spectrograms with a modified WaveNet vocoder [12]. Deep Voice 3 is a neural text-to-speech model that uses a position-augmented attention mechanism for an attention-based decoder [13]. FastSpeech 2 produces high-quality results with the fastest training time [11]. In the synthetic technique, the transcript text with the voice of the target speaker will be fed into the generation model. The text analysis module then processes the incoming text and converts it into linguistic characteristics. Then, the acoustic module extracts the parameters of the target speaker from the dataset depending on the linguistic features generated from the text analysis module. Last, the vocoder will learn to create speech waveforms based on

the acoustic feature parameters, and the final audio file will be generated, which includes the synthetic fake audio in a waveform format. Figure 3 illustrates the process of synthetic-based voice generation.

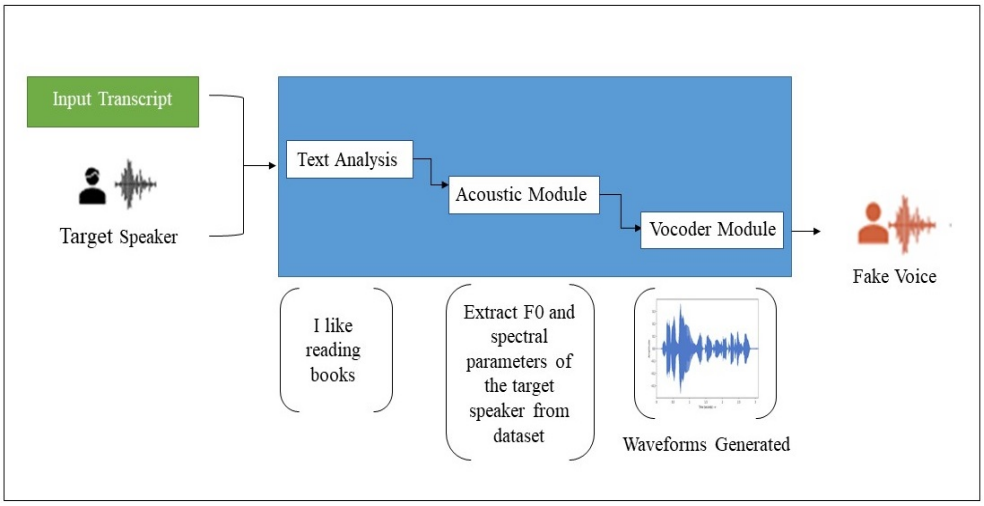

Figure 3. The Synthetic-based Deepfake Process.

Replay-based Deepfakes are a type of malicious work that aims to replay a recording of the target speaker's voice [14]. There are two types: far-field detection and cut-and-paste detection. In far-field detection, a microphone recording of the victim recording is played as a test segment on a telephone handset with a loudspeaker [15]. Meanwhile, cutting and pasting involves faking the sentence required by a text-dependent system [15]. This article will focus on Deepfake methods spoofing real voices rather than approaches that use edited recordings. This review will thus cover the detection methods used to identify synthetic and imitation Deepfakes, and replay-based attacks will be considered out of scope.

3. Fake Audio Detection Methods

The wide range of accessible tools and methods capable of generating fake audio has led to significant recent attention to AD detection with different languages. This section will therefore present the latest work on detecting imitated and synthetically produced voices. In general, the current methods can be divided into two main types: ML and DL methods.

Classical ML models have been widely adopted in AD detection. Rodríguez-Ortega et al. [3] contributed to the literature on detecting fake audio in two aspects. They first developed a fake audio dataset based on the imitation method by extracting the entropy features of real and fake audio. Using the created H-Voice dataset [16], the researchers were able to build an ML model using Logistic Regression (LR) to detect fake audio. The model achieved a 98% success rate in detecting tasks, but the data needed to be pre-processed manually to extract the relevant features.

Kumar-Singh and Singh [17] proposed a Quadratic Support Vector Machine (Q-SVM) model to distinguish synthetic audio from natural human voices. When adopting the model for binary classification, the authors divided the audio into two classes, human and AI-generated. This model was compared to other ML methods, such as Linear Discriminant, Quadratic Discriminant, Linear SVM, weighted K-Nearest Neighbors (KNN), boosted tree ensemble, and LR. As a result, they found that Q-SVM outperformed other classical methods by 97.56%, with a misclassification rate of 2.43%. Moreover, Borrelli et al. [18] developed an SVM model with Random Forest (RF) to predict synthetic voices based on a new audio feature called Short-Term Long-Term (STLT). The models were trained using the Automatic Speaker Verification (ASV) spoof challenge 2019 [19] dataset.

Experiments found that the performance of SVM was higher than that of RF by 71%. Liu et al. [20] compared the robustness of SVM with the DL method called Convolutional Neural Network (CNN) to detect the faked stereo audio from the real ones. From that comparison, it was found that CNN is more robust than SVM even though both achieved a high accuracy of 99% in the detection. However, SVM suffered from what the LR model had faced in the feature extraction process.

According to the works discussed thus far, the features in the ML models need to be manually extracted, and intensive preprocessing is needed before training to ensure good performance. However, this is time-consuming and can lead to inconsistencies, which has led the research community to develop high-level DL methods.. To address this, the CNN model used by Subramani and Rao [21] created a novel approach for detecting synthetic audio based on two CNN models, EfficientCNN and RES-EfficientCNN. As a result, RES-EfficientCNN achieved a higher F1-score of 97.61 than EfficientCNN (94.14 F1-score) when tested over the ASV spoof challenge 2019 dataset [19]. M. Ballesteros et al. [5] developed a classification model named Deep4SNet that visualized the audio dataset based on a 2D CNN model (histogram) to classify imitation and synthetic audio. Deep4SNet showed an accuracy of 98.5% in detecting imitation and synthetic audio. However, Deep4SNet's performance was not scalable and was affected by the data transformation process. E.R. Bartusiak and E.J. Delp [22] compared the performance of the CNN model against the random method in detecting synthetic audio signals. Although the CNN achieved an accuracy 85.99% higher than that of the baseline classifier, it suffered from an overfitting problem. The Lataifeh et al. [23] experimental study compared CNN and Bidirectional Long Short-Term Memory (BiLSTM) performance with ML models. The proposed method targeted the imitation-based fakeness of the Quranic audio clips dataset named Arabic Diversified Audio (AR-DAD) [24]. They tested the ability of CNN and BiLSTM to distinguish real voices from imitators. In addition, ML methods such as SVM, SVM-Linear, Radial Basis Function (SVMRBF), LR, Decision Tree (DT), RF, and Gradient Boosting (XGBoost) were also tested. Ultimately, the study found that SVM had the highest accuracy with 99%, while the lowest was DT with 73.33%. Meanwhile, CNN achieved a detection rate higher than BiLSTM with 94.33%. Although the accuracy of the CNN method was lower than that of the ML models, it was better in capturing spurious correlations. It was also effective in extracting features that could be achieved automatically with generalization abilities. However, the main limitation of the CNN models that are used thus far for AD is that they can only handle images as input, and thus the audio needs to be preprocessed and transformed to a spectrogram or 2D figure to be able to provide it as input to the network.

Zhenchun Lei et al. [25] proposed a 1-D CNN and Siamese CNN to detect fake audio. In the case of the 1-D CNN, the input to the model was the speech log-probabilities, while the Siamese CNN was based on two trained GMM models. The Siamese CNN contained two identical CNNs that were the same as the 1-D CNN but concatenated them using a fully connected layer with a softmax output layer. The two models were tested over the ASVspoof 2019 dataset to find that the proposed Siamese CNN outperformed the GMM and 1-D CNN by improving the min-tDCF and Equal Error Rate (EER) (EER is the error rate where the false-negative rate and the false-positive rate are equal [26]) by ~55% when using the LFCC features. However, the performance was slightly lower when using the CQCC features. It was also found that the model is not sufficiently robust and works with a specific type of feature.

Another CNN model was proposed in [27], where the audio was transferred to scatter plot images of neighboring samples before giving it as input to the CNN model. The developed model was trained over a dataset called the Fake or Real (FoR) dataset [28] to evaluate the model, and the model accuracy reached 88.9%. Although the proposed model addressed the generalization problem of DL-based models by training with data from different generation algorithms, its performance was not as good as the others in the literature. The accuracy (88%) and EER (11%) were worse than those of the other DL models

tested in the experiment. Hence, the model needs further improvement, and more data transformers need to be included.

On the other hand, Yu et al. [29] proposed a new scoring method named Human Log-Likelihoods (HLLs) based on the Deep Neural Network (DNN) classifier to enhance the detection rate. They compared this with a classical scoring method called the Log-Likelihood Ratios (LLRs) that depends on the Gaussian Mixture Model (GMM). DNN-HLLs and GMM-LLRs have been tested with the ASV spoof challenge 2015 dataset [30] and extracted features automatically. These tests confirmed that DNN-HLLs produced better detection results than GMM-LLRs since they achieved an EER of 12.24.

Wang et al. [31] therefore developed a DNN model named Deep-Sonar that captured the neuron behaviors of speaker recognition (SR) systems against AI-synthesized fake audio. Their model depends on Layer-wise neuron behaviors in the classification task. The proposed model achieved a detection rate of 98.1% with an EER of approximately 2% on the voices of English speakers from the FoR dataset [28]. However, DeepSonar's performance was highly affected by real-world noise. Wijethunga et al.'s [32] research used DNNs to differentiate synthetic and real voices and combined two DL models, CNNs and Recurrent Neural Network (RNN). This is because CNN is efficient at extracting features, while RNN is effective at detecting long-term dependencies in time variances. Interestingly, this combination achieved a 94% success rate in detecting audio generated by AI synthesizers. Nevertheless, the DNN model does not carry much artifact information from the feature representation perspective.

Chintha et al. [33] developed two novel models that depend on a convolution RNN for audio Deepfake classification. First, the Convolution Recurrent Neural Network Spoof (CRNN-Spoof) model contains five layers of extracted audio signals that are fed into a bidirectional LSTM network for predicting fake audio. Second, the Wide Inception Residual Network Spoof (WIRE-Net-Spoof) model has a different training process and uses a function named weighted negative log-likelihood. The CRNN-Spoof method obtained higher results than the WIRE-Net-Spoof approach by 0.132% of the Tandem Decision Cost Function (t-DCF) (t-DCF is a single scalar that measure the reliability of decisions made by the systems [34]) with a 4.27% EER in the ASV spoof challenge 2019 dataset [19]. One limitation of this study is that it used many layers and convolutional networks, which caused it to suffer from management complexities. To address this limitation, Shan and Tsai [35] proposed an alignment technique based on the classification models: Long Short-Term Memory (LSTM), bidirectional LSTM, and transformer architectures. The technique classifies each audio frame as matching or nonmatching from 50 recordings. The results reported that bidirectional LSTM outperforms the other models with a 99.7% accuracy and 0.43% EER. However, the training process took a long time, and the dataset used in the study was small, which led to overfitting.

In regard to transfer learning and unimodal methods, P. RahulT et al. [36] proposed a new framework based on transfer learning and the ResNet-34 method for detecting faked English-speaking voices. The transfer learning model was pretrained on the CNN network. The Rest-34 method was used for solving the vanishing gradient problem that always occurs in any DL model. The results showed that the proposed framework achieved the best results measured by the EER and t-DCF metrics with results of 5.32% and 0.1514%, respectively. Although ResNet-34 solves the vanishing gradient issue, training takes a long time because of its deep architecture. Similarly, Khochare et al. [37] investigated feature-based and image-based approaches for classifying faked audio generated synthetically. New DL models called the Temporal Convolutional Network (TCN) and Spatial Transformer Network (STN) were used in this work. TCN achieved promising outcomes in distinguishing between fake and real audio with 92% accuracy, while STN obtained an accuracy of 80%. Although the TCN works well with sequential data, it does not work with inputs converted to Short-Time Fourier Transform (STFT) and Mel Frequency Cepstral Coefficients (MFCC) features.

Khalid et al. [38] contributed a new Deepfake dataset named FakeAVCeleb [39]. The authors investigated unimodal methods that contain five classifiers to evaluate their efficiency in detection; the classifiers were MesoInception-4, Meso-4, Xception, EfficientNet-B0, and VGG16. The Xception classifier was found to achieve the highest performance with a result of 76%, while EfficientNet-B0 had the worst performance with a result of 50%. They concluded that none of the unimodal classifiers were effective for detecting fake audio. Alzantot et al. [40] highlighted the need to develop a system for AD detection based on residual CNN. The main idea of this system is to extract three crucial features from the input, MFCC, constant Q cepstral coefficients (CQCC), and STFT, to determine the Counter Major (CM) score of the faked audio. A high CM score proves that the audio is real speech, while a low CM score suggests that it is fake. The proposed system showed promising results, improving the CM rate by 71% and 75% in two matrices of t-DCF (0.1569) and EER (6.02), respectively. However, further investigation is still needed due to the generalization errors in the proposed system.

T. Arif et al. [41] developed a new audio feature descriptor called ELTP-LFCC based on a Local Ternary Pattern (ELTP) and Linear Frequency Cepstral Coefficients (LFCC). This descriptor was used with a Deep Bidirectional Long Short-Term Memory (DBiLSTM) network to increase the robustness of the model and to detect fake audio in diverse indoor and outdoor environmental conditions. The model created was tested over the ASVspoof 2019 dataset with synthetic and imitated-based fake audio. From the experiment, it was found that the model performed better over the audio synthetic dataset (with 0.74% EER) but not as well with imitated-based samples (with 33.28% EER).

An anti-Spoofing with Squeeze-Excitation and Residual neTworks (ASSERT) method was proposed in [42] based on variants of the Squeeze-Excitation Network (SENet) and ResNet. This method uses log power magnitude spectra (logspec) and CQCC acoustic features to train the DNN. The model was tested with the ASVspoof 2019 dataset to find that ASSERT obtained more than a 17% relative improvements in synthetic audio. However, the model had zero t-DCF cost and zero EER with a logical access scenario during the test, which indicates that the model is highly overfitting.

Based on the literature discussed thus far, we can say that although DL methods avoid manual feature extraction and excessive training, they still require special transformations for audio data. Consequently, self-supervised DL methods have recently been introduced into the AD detection literature. In particular, Jiang et al. [43] proposed a self-supervised spoofing audio detection (SSAD) model inspired by an existing self-supervised DL method named PASE+. The proposed model depends on multilayer convolutional blocks to extract context features from the audio stream. It was tested over the dataset with a 5.31% EER. While the SSAD did well in terms of efficiency and scalability, its performance was not as good as other DL methods. Future research could thus focus on the advantages of self-supervised learning and improving its performance.

Ultimately, the literature discussed thus far is summarized in Table 1, which shows that the method type affects the performance more than the feature used. It is very clear that ML methods are more accurate than DL methods regardless of the features used. However, due to excessive training and manual feature extraction, the scalability of the ML methods is not confirmed, especially with large numbers of audio files. On the other hand, when DL algorithms were used, specific transformations were required on the audio files to ensure that the algorithms could manage them. In conclusion, although AD detection is an active area of study, further research is still needed to address the existing gaps. These challenges and potential future research directions will be highlighted in Section 6.

Table 1. Summary of AD detection methods studies surveyed.

Year	Ref.	Speech Language	Fakeness Type	Technique	Audio Feature Used	Dataset	Drawbacks
2018	Yu et al. [29]	English	Synthetic	DNN-HLL	MFCC, LFCC, CQCC	ASV spoof 2015 [30]	The error rate is zero, indicating that the proposed DNN is overfitting.
				GMM-LLR	IMFCC, GFCC, IGFCC		Does not carry much artifact information in the feature representations perspective.
2019	Alzantot et al. [40]	English	Synthetic	Residual CNN	MFCC, CQCC, STFT	ASV spoof 2019 [19]	The model is highly overfitting with synthetic data and cannot be generalized over unknown attacks.
2019	C. Lai et al. [42]	English	Synthetic	ASSERT (SENet + ResNet)	Logspec, CQCC	ASV spoof 2019 [19]	The model is highly overfitting with synthetic data.
2020	P. RahulT et al. [36]	English	Synthetic	ResNet-34	Spectrogram	ASV spoof 2019 [19]	Requires transforming the input into a 2-D feature map before the detection process, which increases the training time and effects its speed.
2020	Lataifeh et al. [23]	Classical Arabic	Imitation	Classical Classifiers (SVM-Linear, SVMRBF, LR, DT, RF, XGBoost)	-	Arabic Diversified Audio (AR-DAD) [24]	Failed to capture spurious correlations, and features are extracted manually so they are not scalable and needs extensive manual labor to prepare the data.
				DL Classifiers (CNN, BiLSTM)	MFCC spectrogram		DL accuracy was not as good as the classical methods, and they are an image-based approach that requires special transformation of the data.
2020	Rodríguez-Ortega et al. [3]	Spanish, English, Portuguese, French, and Tagalog	Imitation	LR	Time domain waveform	H-Voice [16]	Failed to capture spurious correlations, and features are extracted manually so it is not scalable and needs extensive manual labor to prepare the data.
2020	Wang et al. [31]	English, Chinese	Synthetic	Deep-Sonar	High-dimensional data visualization of MFCC, raw neuron, activated neuron	FoR dataset [28]	Highly affected by real-world noises.
2020	Subramani and Rao [21]	English	Synthetic	EfficientCNN and RES-EfficientCNN	Spectrogram	ASV spoof 2019 [19]	They use an image-based approach that requires special transformation of the data to transfer audio files into images.
2020	Shan and Tsai [35]	English	Synthetic	Bidirectional LSTM	MFCC	--	The method did not perform well over long 5 s edits.
2020	Wijethunga et al. [32]	English	Synthetic	DNN	MFCC, Mel-spectrogram, STFT	Urban-Sound8K, Conversational, AMI-Corpus, and FoR	The proposed model does not carry much artifact information from the feature representations perspective.
2020	Jiang et al. [43]	English	Synthetic	SSAD	LPS, LFCC, CQCC	ASV spoof 2019 [19]	It needs extensive computing processing since it uses a temporal convolutional network (TCN) to capture the context features and another three regression workers and one binary worker to predict the target features.
2020	Chintha et al. [33]	English	Synthetic	CRNN-Spoof	CQCC	ASV spoof 2019 [19]	The model proposed is complex and contains many layers and convolutional networks, so it needs an extensive computing process. Did not perform well compared to WIRE-Net-Spoof.

Year	Ref.	Speech Language	Fakeness Type	Technique	Audio Feature Used	Dataset	Drawbacks
				WIRE- Net-Spoof	MFCC		Did not perform well compared to CRNN-Spoof.
2020	Kumar-Singh and Singh [17]	English	Synthetic	Q-SVM	MFCC, Mel-spectrogram	--	Features are extracted manually so it is not scalable and needs extensive manual labor to prepare the data.
2020	Zhenchun Lei et al. [25]	English	Synthetic	CNN and Siamese CNN	CQCC, LFCC	ASV spoof 2019 [19]	The models are not robust to different features and work best with LFCC only.
2021	M. Ballesteros et al. [5]	Spanish, English, Portuguese, French, and Tagalog	Synthetic Imitation	Deep4SNet	Histogram, Spectrogram, Time domain waveform	H-Voice [16]	The model was not scalable and was affected by the data transformation process.
2021	E.R. Bartusiak and E.J. Delp [22]	English	Synthetic	CNN	Spectrogram	ASV spoof 2019 [19]	They used an image-based approach, which required a special transformation of the data, and the authors found that the model proposed failed to correctly classify new audio signals indicating that the model is not general enough.
2021	Borrelli et al. [18]	English	Synthetic	RF, SVM	STLT	ASV spoof 2019 [19]	Features extracted manually so they are not scalable and needs extensive manual labor to prepare the data.
2021	Khalid et al. [38]	English	Synthetic	MesoInception-4, Meso-4, Xception, EfficientNet-B0, VGG16	Three-channel image of MFCC	FakeAVCeleb [39]	It was observed from the experiment that Meso-4 overfits the real class and MesoInception-4 overfits the fake class, and none of the methods provided a satisfactory performance indicating that they are not suitable for fake audio detection.
2021	Khochare et al. [37]	English	Synthetic	Feature-based (SVM, RF, KNN, XGBoost, and LGBM)	Vector of 37 features of audio	FoR dataset [28]	Features extracted manually so they are not scalable and needs extensive manual labor to prepare the data.
				Image-based (CNN, TCN, STN)	Melspectrogram		It uses an image-based approach and could not work with inputs converted to STFT and MFCC features.
2021	Liu et al. [20]	Chinese	Synthetic	SVM	MFCC	--	Features extracted manually so it is not scalable and needs extensive manual labor to prepare the data.
				CNN	--		The error rate is zero indicating that the proposed CNN is overfitting.
2021	S. Camacho et al. [27]	English	Synthetic	CNN	Scatter plots	FoR dataset [28]	It did not perform as well as the traditional DL methods, and the model needed more training.
2021	T. Arif et al. [41]	English	Synthetic imitated	DBiLSTM	ELTP-LFCC	ASV spoof 2019 [19]	Does not perform well over an imitated-based dataset.

4. Fake Audio Detection Datasets

The previous section discussed various methods used to distinguish fake and real voices. The detection methods discussed use models that must first be trained on a sample of data. Different datasets were published in the literature along with the detection method, while other studies focused on explaining their dataset and its characteristics. This section describes eight recent datasets created for use in fake detection methods, which are summarized in Table 2.

More datasets have been reported for AD detection and were published separately. A German audio dataset named M-AILABS Speech was published for use with speech recognition and synthetic audio. It is freely accessible and contains 9265 real audio samples along with 806 fake samples. Each sample differs in length from 1 to 20 s, and the set has a total length of 18.7 h. The company Baidu published another public dataset called the Baidu Silicon Valley AI Lab cloned audio dataset, which was generated by a neural voice cloning tool. This dataset contains 6 h of high-quality and multi-speaker audio clips, each 2 s long. In 2019, the Fake or Real (FoR) dataset was released, which included eight synthetically generated English-accented voices by the Deep Voice 3 and Google-WavNet generation models. It is available for public access, and its most crucial feature is that it includes samples in two types of formats, MP3 and WAV. The total dataset includes 198,000 files divided into 111,000 real samples and 87,000 faked samples, with each sample being 2 s long. There was also a faked audio dataset of Arabic speakers called the Ar-DAD Arabic Diversified Audio gathered from the Holy Quran audio portal. It contains the original and imitated voices of Quran reciters, while the audio speech is for 30 male Arabic reciters and 12 imitators. Specifically, the reciters are Arabic people from Saudi Arabia, Kuwait, Egypt, Yemen, Sudan, and the UAE. The data consist of 379 fake and 15,810 real samples, each being 10 s long. The dataset language is named Classical Arabic (CA) since it is in the Quranic language.

Furthermore, the H-Voice dataset was generated recently based on imitation and synthetic voices speaking in different languages, namely, Spanish, English, Portuguese, French, and Tagalog. It contains samples saved as a histogram, which is in the PNG format. This dataset contains 6672 samples and has many folders, as illustrated in Figure 4, which also depicts the number of imitated and synthetically created samples in each folder. However, Table 2 combines the number of samples as 3332 real and 3264 fake imitated samples, as well as 4 real and 72 fake synthetic samples. It is public access, and the model generation for the synthetic-based files is Deep Voice 3.

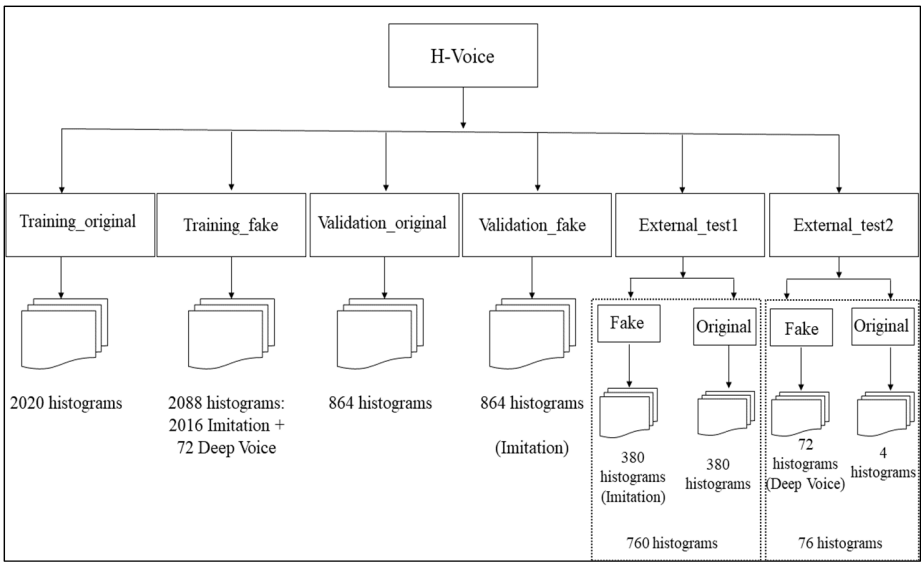

Figure 4. The Structure of the H-Voice Dataset.

Table 2. Summary of AD datasets.

Year	Dataset	Total Size	Real Sample Size	Fake Sample Size	Sample Length (s	Fakeness Type	Format	Speech Language	Accessibility	Dataset URL
2018	The M-AILABS Speech [44]	18,7 h	9265	806	1–20	Synthetic	WAV	German	Public	https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/ (accessed 3 March 2022)
2018	Baidu Silicon Valley AI Lab cloned audio [45]	6 h	10	120	2	Synthetic	Mp3	English	Public	https://audiodemos.github.io/ (accessed 3 March 2022)
2019	Fake oR Real (FoR) [28]	198,000 Files	111,000	87,000	2	Synthetic	Mp3, WAV	English	Public	https://bil.eecs.yorku.ca/datasets/(accessed 20 November 2021)
2020	AR-DAD: Arabic Diversified Audio [24]	16,209 Files	15,810	397	10	Imitation	WAV	Classical Arabic	Public	https://data.mendeley.com/datasets/3kndp5vs6b/3(accessed 20 November 2021)
2020	H-Voice [16]	6672 Files	Imitation 3332 Synthetic 4	Imitation 3264 Synthetic 72	2–10	Imitation Synthetic	PNG	Spanish, English, Portuguese, French, and Tagalog	Public	https://data.mendeley.com/datasets/k47yd3m28w/4 (accessed 20 November 2021)
2021	ASV spoof 2021 Challenge	-	-	-	2	Synthetic	Mp3	English	Only older versions available thus far	https://datashare.ed.ac.uk/handle/10283/3336(accessed 20 November 2021)
2021	FakeAVCeleb [39]	20,490 Files	490	20,000	7	Synthetic	Mp3	English	Restricted	https://sites.google.com/view/fakeavcelebdash-lab/(accessed 20 November 2021)
2022	ADD [46]	85 h	LF:300 PF:0	LF:700 PF:1052	2–10	Synthetic	WAV	Chinese	Public	https://sites.google.com/view/fakeavcelebdash-lab/(accessed 3 May 2022)

Moreover, the FakeAVCeleb dataset is a new restricted dataset of English speakers that has been synthetically generated by the SV2TTS tool. It contains a total of 20,490 samples divided between 490 real samples and 20,000 fakes, each being 7 s long in MP3 format. Last, the ASV spoof 2021 challenge dataset also consists of two fake scenarios, a logical and a physical scenario. The logical scenario contains fake audio made using synthetic software, while the physical scenario is fake audio made by reproducing prerecorded audio using parts of real speaker data. While this dataset has yet to be published, older versions are available to the public (2015 [30], 2017 [47], and 2019 [19]).

However, the ASV spoof challenge has one limitation that has not been considered a crucial factor in the AD area, which is noise. A new synthetic-based dataset was therefore developed in the current year to fill this gap called the Audio Deep synthesis Detection challenge (ADD). This dataset consists of three tracks, a low-quality fake audio detection (LF), a partially fake audio detection (PF), and a fake audio game (FG), which is outside the scope of the current article. LF contains 300 real voices and 700 fully faked spoken words with real-world noises, while PF has 1052 partially fake audio samples. The language of the ADD dataset is Chinese, and it is publicly available.

From Table 2, it can be concluded that most datasets have been developed for English. While one dataset was found for Classical Arabic (CA) language, it covered only imitation fakeness, while other types of the Arabic language were not covered. There is thus still a need to generate a new dataset based on the syntactic fakeness of the Arabic language. The developed dataset can be used to complement the developed AD detection model to detect both imitation and synthetic Deepfakes with minimal preprocessing and training delays.

5. Discussion

From the literature, it was clear that the methods proposed thus far require special data processing to perform well, where classical ML methods require extensive amounts of manual labor to prepare the data, while the DL-based methods use an image-based approach to understand the audio features. The preprocessing approach used can affect the performance of the method, and thus new research is recommended to develop new solutions that allow the models to understand the audio data as it is. Nevertheless, it was crucial to analyze the statutes of the current AD detection methods based on previous work experiments. Thus, from the experimental results of the cited studies, a quantitative comparison was conducted based on three criteria (EER, t-DCF, and accuracy), as illustrated in Table 3.

Table 3. A quantitative comparison between AD detection methods.

Measures	Dataset	Detection Method	Results (The Result Is Approximate from the Evaluation Test Published in the Study)
EER	ASV spoof 2015 challenge	DNN-HLLs [29]	12.24%
		GMM-LLR [29]	42.5%
	ASV spoof 2019 challenge	Residual CNN [40]	6.02%
		SENet-34 [42]	6.70%
		CRNN-Spoof [33]	4.27%
		ResNet-34 [36]	5.32%
		Siamese CNN [25]	8.75%
		CNN [25]	9.61%
		DBiLSTM [41] (Synthetic Audio)	0.74%
		DBiLSTM [41] (Imitation-based)	33.30%
		SSAD [43]	5.31%
	-	Bidirectional LSTM [35]	0.43%

Measures	Dataset	Detection Method	Results (The Result Is Approximate from the Evaluation Test Published in the Study)
	FoR	CNN [27]	11.00%
		Deep-Sonar [31]	2.10%
t-DCF	ASV spoof 2019 challenge	Residual CNN [40]	0.1569
		SENet-34 [42]	0.155
		CRNN-Spoof [33]	0.132
		ResNet-34 [36]	0.1514
		Siamese CNN [25]	0.211
		CNN [25]	0.217
		DBiLSTM [41] (Synthetic Audio)	0.008
		DBiLSTM [41] (Imitation-based)	0.39
Accuracy	ASV spoof 2019 challenge	CNN [22]	85.99%
		SVM [18]	71.00%
	AR-DAD	CNN [23]	94.33%
		BiLSTM [23]	91.00%
		SVM [23]	99.00%
		DT [23]	73.33%
		RF [23]	93.67%
		LR [23]	98.00%
		XGBoost [23]	97.67%
		SVMRBF [23]	99.00%
		SVM-LINEAR [23]	99.00%
	FoR	DNN [32]	94.00%
		Deep-Sonar [31]	98.10%
		STN [37]	80.00%
		TCN [37]	92.00%
		SVM [37]	67%
		RF [37]	62%
		KNN [37]	62%
		XGBoost [37]	59%
		LGBM [37]	60%
		CNN [27]	88.00%
	FakeAVCeleb	EfficientNet-B0 [38]	50.00%
		Xception [38]	76.00%
		MesoInception-4 [38]	53.96%
		Meso-4 [38]	50.36%
		VGG16 [38]	67.14%
	H-Voice	LR [3]	98%
		Deep4SNet [5]	98.5%
	-	Q-SVM [17]	97.56%
	-	CNN [20]	99%
	-	SVM [20]	99%

Starting with the EER and t-DCF, as shown in Figure 5, it can be concluded that there is no clear pattern in performance with respect to the approach or dataset used. Each method performs differently depending on the technique used. For instance, the Bidirectional LSTM method provides the best EER and t-DCF compared to the other methods, but the dataset information was not clarified in the study, and overfitting was a concern.

Another example is GMM-LLR, which provides the worst EER even though it used the same features and trained on the same dataset as DNN-HLLs. In regard to the CNN methods highlighted in the orange box, regardless of the dataset used, all versions have a similar performance with respect to the EER and t-DCF. However, one interesting observation that can be highlighted is the fact that the type of fakeness can have an effect on the performance of the method. For instance, DBiLSTM provides a very low EER and t-DCF compared to the other methods when applied with synthetic AD, while it is one of the worst when applied to the imitation-based datasets.

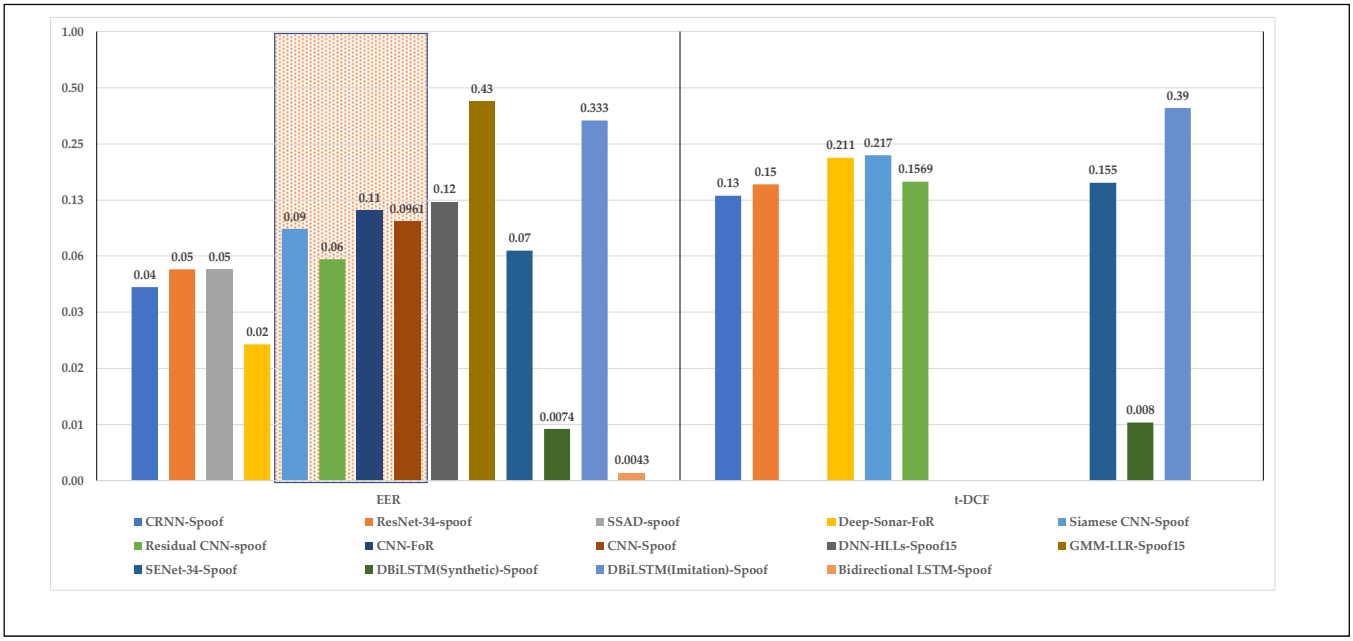

Figure 5. Quantitative comparison of AD detection methods measured by EER and t-DCF.

When considering the accuracy, illustrated in Figure 6, a pattern can be identified with respect to the dataset and fakeness type. In particular, regardless of the method used, the accuracy of the methods applied on the FakeAVCeleb dataset did not perform well compared to the other datasets. This can be attributed to the fact that this dataset is a combination of video-audio Deepfakes. Moreover, ML methods perform better with imitation-based fakeness compared to the synthetic-based datasets. For instance, in SVM, RF, and XGBoost, highlighted with green, gray, and purple boxes, respectively, it is clear that they perform almost perfectly when applied to the imitation-based datasets (AR-DAD and H-voice) while preforming poorly when applied over the synthetic-based datasets (ASV spoof and FoR). Moreover, it is interesting to note that DL-based methods such as CNN are more stable than the ML methods with respect to the fakeness type. For instance, comparing CNN versions (under the orange box) with SVM versions (under the gray box), it is clear that CNN is stable and has a similar performance regardless of the dataset, while SVM is unstable and performs differently depending on the data and fakeness type. Thus, it can be concluded that regardless of how ML methods perform well, for more stability and consistency in performance, DL-based methods are better options. However, further improvements are still needed to allow the addressing of audio data directly and to surpass the extensive preprocessing and data transformation needed in the current literature.

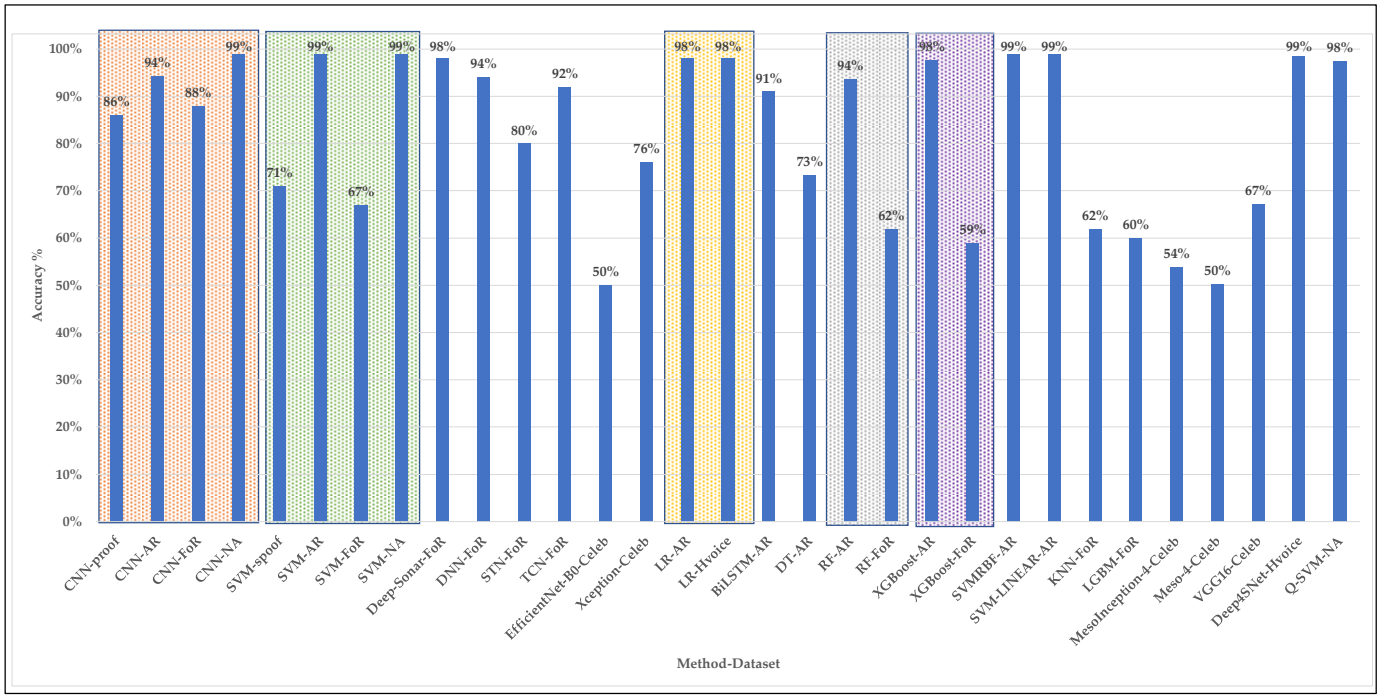

Figure 6. Quantitative comparison between recent AD detection methods measured by accuracy on multiple datasets.

6. Challenges and Future Research Directions

This section will highlight the most important challenges and opportunities facing researchers in the AD field. It will examine the most crucial challenges with fake audio detection methods.

6.1. Limited AD Detection Methods with Respect to Non-English Languages

Almost all existing studies focus on developing detection methods to detect fake voices speaking English, although six official languages are included in the United Nations' list of official languages [48]. For example, the authors are aware of no existing studies focusing on Arabic. Indeed, Arabic is the world's fourth most widely spoken language behind Chinese, Spanish, and English, with over 230 million native speakers [49]. It consists of three core types: Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialect Arabic (DA) [50]. CA is the official language of the Quran; MSA is the official language of the present era; and DA is the spoken language of everyday life that differs between regions [50]. The reason for highlighting the Arabic language in this section is because the Arabic language has a unique challenge in alphabet pronunciation, which the traditional techniques of audio processing and ML learning models cannot deal with [51]. It contains three crucial long vowels named Fatha, Damma, and Khasra [52], which if pronounced incorrectly will change the meaning of the sentence [51]. The authors [53] therefore pointed out that the performance of any model in a specific language will not be the same in other languages, especially in languages that have limited available data, such as Arabic and Chinese. There was only one attempt by [24], where the authors collected CA data based on imitation fakes. For this reason, we can directly understand the lack of detection methods for non-English AD. We therefore encourage the research community to meet this research gap by proposing a new detection method to detect other languages, such as Arabic.

6.2. Lack of Accent Assessment in Existing AD Detection Methods

The majority of detection methods rely on detecting the type of fake itself without considering other factors that could affect the accuracy of the detection. One such factor

is accents, which are defined as the way a specific group of people typically speak, particularly the citizens or natives of a particular country [54]. Research on this subject is still missing from the AD literature, and it is presently unclear whether accents can affect detection accuracy. In other audio fields, such as speaker recognition, accents affected the performance of the methods proposed [55]. Thus, it is expected that accents can be a challenge in the AD area. To address this challenge, further study is needed on languages that use many different accents, such as Arabic. One country will often contain speakers using many different accents, and the Saudi language is no exception, as it contains Najdi, Hijazi, Qaseemi, and many other accents. Further research is necessary because when the number of accents increases, the chance of the classifier learning a more generalized model for the detection task will increase [28]. We therefore suggest that future research focus on covering the AD detection area and measuring the effectiveness of accents, especially Saudi accents.

6.3. Excessive Preprocessing to Build Deepfake Detection Models

As discussed in the literature of this article, and even though AD detection methods are able to provide high detection accuracy, these methods are currently trading off efficiency with scalability. For this reason, new self-supervised methods should be developed to avoid the excessive preprocessing of ML methods and the extra transformation of DL methods. To date, the Self-Supervised Learning (SSL) approach has not been fully considered in the AD area and can be a valuable solution to overcome these challenges. As confirmed in [56], the most crucial aspect of SSL effectiveness is dealing with unlabeled data to work effectively in detection tasks. Only one attempt [43] has been seen in the literature addressing the SSL method, and although it was efficient and scalable in solving the issues of supervised algorithms, the detection rate was very low. Thus, it is encouraged to develop new SSL methods with better accuracy that can overcome the ML and DL challenges while introducing better performances.

6.4. Limited Assessment of Noisy Audio in Existing AD Detection Methods

Noise in general is defined as "arbitrary, unwanted electrical energy that enters the communications system through the communicating medium and obstructs with the conveyed message" [57]. Noises can also be generated from natural sources, such as rain, wind, cars, or voices. Voices that have been recorded indoors or outdoors can be affected by real-world noises, such as laughter and rain [31]. However, attackers can easily deceive detectors by introducing real-world noises, so robustness is crucial for fake voice detectors. Unfortunately, only one attempt has been made to study this issue in the AD area, and this failed to tackle the effects of real-world noises using the proposed detection method, which is in work [31]. This direction could thus be a starting point for researchers looking to develop a robust fake audio detection method that works even with noisy data in the wild.

6.5. Limited AD Imitation-Based Detection Methods

From the literature discussed, most of the related works have been focused on synthetic-based detection methods, whereas imitation-based methods have been limited. The reason for that is confirmed by M. Ballesteros et al. [5], where detecting imitated voices is not a trivial process since a faked voice sounds more similar to the original. Thus, to fill this limitation, we encourage the research community to take on this limitation in the future.

7. Conclusions

This review article has discussed the field of AD, carefully surveying a number of studies exploring detection methods with respect to current datasets. It began by presenting a broad overview of AD, along with their definitions and types. Then, it reviewed the

relevant articles that have addressed the subject over the last four years and examined the limitations covered in the literature on classical ML and DL detection methods. Following this, the available faked audio datasets were summarized, and the discussed methods were also compared. Moreover, a quantitative comparison of recent state-of-the-art AD detection methods was also provided. Finally, the research challenges and opportunities of the field were discussed. From this analysis, it can be concluded that further advancements are still needed in the literature of fake audio detection to develop a method that can detect fakeness with different accents or real-world noises. Moreover, the SSL approach can be one future research direction to help solve the current issues affecting the existing AD methods. Imitation-based AD detection is an important part of the AD field that also needs further development in comparison to the synthesis-based methods.

Author Contributions: Conceptualization, Z.A.; formal analysis, Z.A.; resources, Z.A.; writing—original draft preparation, Z.A.; writing—review and editing, H.E.; visualization, Z.A.; supervision, H.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lyu, S. Deepfake detection: Current challenges and next steps. *IEEE Comput. Soc.* **2020**, 1–6. doi: 10.1109/ICMEW46912.2020.9105991.
2. Diakopoulos, N.; Johnson, D. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media Soc.* **2021**, *23*, 2072–2098. https://doi.org/10.1177/1461444820925811.
3. Rodríguez-Ortega, Y.; Ballesteros, D.M.; Renza, D. A machine learning model to detect fake voice. In *Applied Informatics*; Florez, H., Misra, S., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 3–13.
4. Chen, T.; Kumar, A.; Nagarsheth, P.; Sivaraman, G.; Khoury, E. Generalization of audio deepfake detection. In Proceedings of the Odyssey 2020 The Speaker and Language Recognition Workshop, Tokyo, Japan, 1–5 November 2020; pp. 132–137.
5. Ballesteros, D.M.; Rodriguez-Ortega, Y.; Renza, D.; Arce, G. Deep4SNet: Deep learning for fake speech classification. *Expert Syst. Appl.* **2021**, *184*, 115465. https://doi.org/10.1016/j.eswa.2021.115465.
6. Suwajanakorn, S.; Seitz, S.M.; Kemelmacher-Shlizerman, I. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph. ToG* **2017**, *36*, 1–13.
7. Catherine Stupp Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. Available online: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 (accessed on 29 January 2022).
8. Chadha, A.; Kumar, V.; Kashyap, S.; Gupta, M. Deepfake: An overview. In *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security*; Singh, P.K., Wierzchoń, S.T., Tanwar, S., Ganzha, M., Rodrigues, J.J.P.C., Eds.; Springer: Singapore, 2021; pp. 557–566.
9. Tan, X.; Qin, T.; Soong, F.; Liu, T.-Y. A survey on neural speech synthesis. *arXiv* **2021**, arXiv:2106.15561.
10. Ning, Y.; He, S.; Wu, Z.; Xing, C.; Zhang, L.-J. A Review of Deep Learning Based Speech Synthesis. *Appl. Sci.* **2019**, *9*. https://doi.org/10.3390/app9194050.
11. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.-Y. Fastspeech 2: Fast and High-Quality End-to-End Text to Speech. *arXiv* **2020**, arXiv:2006.04558.
12. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R. *Natural Tts Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions*; IEEE: Piscataway, NJ, USA, 2018; pp. 4779–4783.
13. Ping, W.; Peng, K.; Gibiansky, A.; Arik, S.O.; Kannan, A.; Narang, S.; Raiman, J.; Miller, J. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv* **2017**, arXiv:1710.07654.
14. Khanjani, Z.; Watson, G.; Janeja, V.P. How deep are the fakes? Focusing on audio deepfake: A survey. *arXiv* **2021**, arXiv:2111.14203.
15. Pradhan, S.; Sun, W.; Baig, G.; Qiu, L. Combating replay attacks against voice assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2019**, *3*, 1–26.
16. Ballesteros, D.M.; Rodriguez, Y.; Renza, D. A dataset of histograms of original and fake voice recordings (H-voice). *Data Brief* **2020**, *29*, 105331. https://doi.org/10.1016/j.dib.2020.105331.

17. Singh, A.K. and Singh, P. Detection of ai-synthesized speech using cepstral & bispectral statistics. In Proceedings of the 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), Tokyo, Japan, 8–10 September 2021; pp. 412–417.
18. Borrelli, C.; Bestagini, P.; Antonacci, F.; Sarti, A.; Tubaro, S. Synthetic speech detection through short-term and long-term prediction traces. *EURASIP J. Inf. Secur.* **2021**, *2021*, 2. https://doi.org/10.1186/s13635-021-00116-3.
19. Todisco, M.; Wang, X.; Vestman, V.; Sahidullah, M.; Delgado, H.; Nautsch, A.; Yamagishi, J.; Evans, N.; Kinnunen, T.; Lee, K.A. ASVspoof 2019: Future horizons in spoofed and fake audio detection *arXiv* **2019**, arXiv:1904.05441
20. Liu, T.; Yan, D.; Wang, R.; Yan, N.; Chen, G. Identification of fake stereo audio using SVM and CNN. *Information* **2021**, *12*, 263. https://doi.org/10.3390/info12070263.
21. Subramani, N.; Rao, D. Learning efficient representations for fake speech detection. In Proceedings of the The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, 7–12 February 2020; pp. 5859–5866.
22. Bartusiak, E.R.; Delp, E.J. Frequency domain-based detection of generated audio. In Proceedings of the Electronic Imaging; Society for Imaging Science and Technology, New York, USA, 11–15 January 2021; Volume 2021, pp. 273–281.
23. Lataifeh, M.; Elnagar, A.; Shahin, I.; Nassif, A.B. Arabic audio clips: Identification and discrimination of authentic cantillations from imitations. *Neurocomputing* **2020**, *418*, 162–177. https://doi.org/10.1016/j.neucom.2020.07.099.
24. Lataifeh, M.; Elnagar, A. Ar-DAD: Arabic diversified audio dataset. *Data Brief* **2020**, *33*, 106503. https://doi.org/10.1016/j.dib.2020.106503.
25. Lei, Z.; Yang, Y.; Liu, C.; Ye, J. Siamese convolutional neural network using gaussian probability feature for spoofing speech detection. In Proceedings of the INTERSPEECH, Shanghai, China, 25–29 October 2020; pp. 1116–1120.
26. Hofbauer, H.; Uhl, A. Calculating a boundary for the significance from the equal-error rate. In Proceedings of the 2016 International Conference on Biometrics (ICB), Halmstad, Sweden, 13 June 2016; pp. 1–4.
27. Camacho, S.; Ballesteros, D.M.; Renza, D. Fake speech recognition using deep learning. In *Applied Computer Sciences in Engineering*; Figueroa-García, J.C., Díaz-Gutierrez, Y., Gaona-García, E.E., Orjuela-Cañón, A.D., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 38–48.
28. Reimao, R.; Tzerpos, V. For: A dataset for synthetic speech detection. In Proceedings of the 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timisoara, Romania, 10 October 2019; pp. 1–10.
29. Yu, H.; Tan, Z.-H.; Ma, Z.; Martin, R.; Guo, J. Guo spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 4633–4644. https://doi.org/10.1109/TNNLS.2017.2771947.
30. Wu, Z.; Kinnunen, T.; Evans, N.; Yamagishi, J.; Hanilçi, C.; Sahidullah, M.; Sizov, A. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In Proceedings of the Interspeech 2015, Dresden, Germany, 6–10 September 2015; p. 5.
31. Wang, R.; Juefei-Xu, F.; Huang, Y.; Guo, Q.; Xie, X.; Ma, L.; Liu, Y. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. In Proceedings of the the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1207–1216.
32. Wijethunga, R.L.M.A.P.C.; Matheesha, D.M.K.; Al Noman, A.; De Silva, K.H.V.T.A.; Tissera, M.; Rupasinghe, L. Rupasinghe deepfake audio detection: A deep learning based solution for group conversations. In Proceedings of the 2020 2nd International Conference on Advancements in Computing (ICAC), Malabe, Sri Lanka, 10–11 December 2020; Volume 1, pp. 192–197.
33. Chintha, A.; Thai, B.; Sohrawardi, S.J.; Bhatt, K.M.; Hickerson, A.; Wright, M.; Ptucha, R. Ptucha recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE J. Sel. Top. Signal. Process.* **2020**, *14*, 1024–1037. https://doi.org/10.1109/JSTSP.2020.2999185.
34. Kinnunen, T.; Lee, K.A.; Delgado, H.; Evans, N.; Todisco, M.; Sahidullah, M.; Yamagishi, J.; Reynolds, D.A. T-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification. *arXiv* **2018**, arXiv:1804.09618.
35. Shan, M.; Tsai, T. A cross-verification approach for protecting world leaders from fake and tampered audio. *arXiv* **2020**, arXiv:2010.12173.
36. Aravind, P.R.; Nechiyil, U.; Paramparambath, N. Audio spoofing verification using deep convolutional neural networks by transfer learning. *arXiv* **2020**, arXiv:2008.03464.
37. Khochare, J.; Joshi, C.; Yenarkar, B.; Suratkar, S.; Kazi, F. A deep learning framework for audio deepfake detection. *Arab. J. Sci. Eng.* **2021**, *47*, 3447–3458. https://doi.org/10.1007/s13369-021-06297-w.
38. Khalid, H.; Kim, M.; Tariq, S.; Woo, S.S. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In Proceedings of the the 1st Workshop on Synthetic Multimedia, ; ACM Association for Computing Machinery: New York, NY, USA,20 October 2021; pp. 7–15.
39. Khalid, H.; Tariq, S.; Kim, M.; Woo, S.S. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks, Virtual, 6–14 December 2021; p. 14.
40. Alzantot, M.; Wang, Z.; Srivastava, M.B. Deep residual neural networks for audio spoofing detection. *arXiv CoRR* **2019**, arXiv:1907.00501.

41. Arif, T.; Javed, A.; Alhameed, M.; Jeribi, F.; Tahir, A. Voice spoofing countermeasure for logical access attacks detection. *IEEE Access* **2021**, *9*, 162857–162868. https://doi.org/10.1109/ACCESS.2021.3133134.
42. Lai, C.-I.; Chen, N.; Villalba, J.; Dehak, N. ASSERT: Anti-spoofing with squeeze-excitation and residual networks. *arXiv* **2019**, arXiv:1904.01120.
43. Jiang, Z.; Zhu, H.; Peng, L.; Ding, W.; Ren, Y. Self-supervised spoofing audio detection scheme. In Proceedings of the INTERSPEECH 2020, Shanghai, China, 25–29 October 2020; pp. 4223–4227.
44. Imdat Solak The M-AILABS Speech Dataset. Available online: https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/(accessed on 10 March 2022).
45. Arik, S.O.; Chen, J.; Peng, K.; Ping, W.; Zhou, Y. Neural voice cloning with a few samples. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018); Montréal, QC, Canada, 2–8 December 2018; p. 11.
46. Yi, J.; Fu, R.; Tao, J.; Nie, S.; Ma, H.; Wang, C.; Wang, T.; Tian, Z.; Bai, Y.; Fan, C. Add 2022: The first audio deep synthesis detection challenge. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing; Singapore, 23–27 May 2022; p. 5.
47. Kinnunen, T.; Sahidullah, M.; Delgado, H.; Todisco, M.; Evans, N.; Yamagishi, J.; Lee, K.A. The 2nd Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2017) Database, Version 2. Available online: https://datashare.ed.ac.uk/handle/10283/3055 (accessed on 5 November 2021).
48. Nations, U. Official Languages. Available online: https://www.un.org/en/our-work/official-languages (accessed on 5 March 2022).
49. Almeman, K.; Lee, M. A comparison of arabic speech recognition for multi-dialect vs. specific dialects. In Proceedings of the Seventh International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013), Cluj-Napoca, Romania; 16–19 October 2013; pp. 16–19.
50. Elgibreen, H.; Faisal, M.; Al Sulaiman, M.; Abdou, S.; Mekhtiche, M.A.; Moussa, A.M.; Alohali, Y.A.; Abdul, W.; Muhammad, G.; Rashwan, M.; et al. An Incremental Approach to Corpus Design and Construction: Application to a Large Contemporary Saudi Corpus. *IEEE Access* **2021**, *9*, 88405–88428. https://doi.org/10.1109/ACCESS.2021.3089924.
51. Asif, A.; Mukhtar, H.; Alqadheeb, F.; Ahmad, H.F.; Alhumam, A. An approach for pronunciation classification of classical arabic phonemes using deep learning. *Appl. Sci.* **2022**, *12*, 238. https://doi.org/10.3390/app12010238.
52. Ibrahim, A.B.; Seddiq, Y.M.; Meftah, A.H.; Alghamdi, M.; Selouani, S.-A.; Qamhan, M.A.; Alotaibi, Y.A.; Alshebeili, S.A. Optimizing Arabic Speech Distinctive Phonetic Features and Phoneme Recognition Using Genetic Algorithm. *IEEE Access* **2020**, *8*, 200395–200411. https://doi.org/10.1109/ACCESS.2020.3034762.
53. Maw, M.; Balakrishnan, V.; Rana, O.; Ravana, S.D. Trends and patterns of text classification techniques: A systematic mapping study. *Malays. J. Comput. Sci.* **2020**, *33*, 102–117.
54. Rizwan, M.; Odelowo, B.O.; Anderson, D.V. Word based dialect classification using extreme learning machines. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24 July 2016; pp. 2625–2629.
55. Najafian, M. Modeling accents for automatic speech recognition. In Proceedings of the 23rd European Signal Proceedings (EUSIPCO), Nice, France, 31 August–4 September 2015; University of Birmingham: Birmingham, UK, 2013; Volume 1568, p. 1.
56. Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; Tang, J. Self-supervised learning: Generative or contrastive. *IEEE Trans. Knowl. Data Eng.* **2021**. doi:10.1109/TKDE.2021.3090866.
57. Jain, D.; Beniwal, D.P. Review paper on noise cancellation using adaptive filters. *Int. J. Eng. Res. Technol.* **2022**, *11*, 241–244.