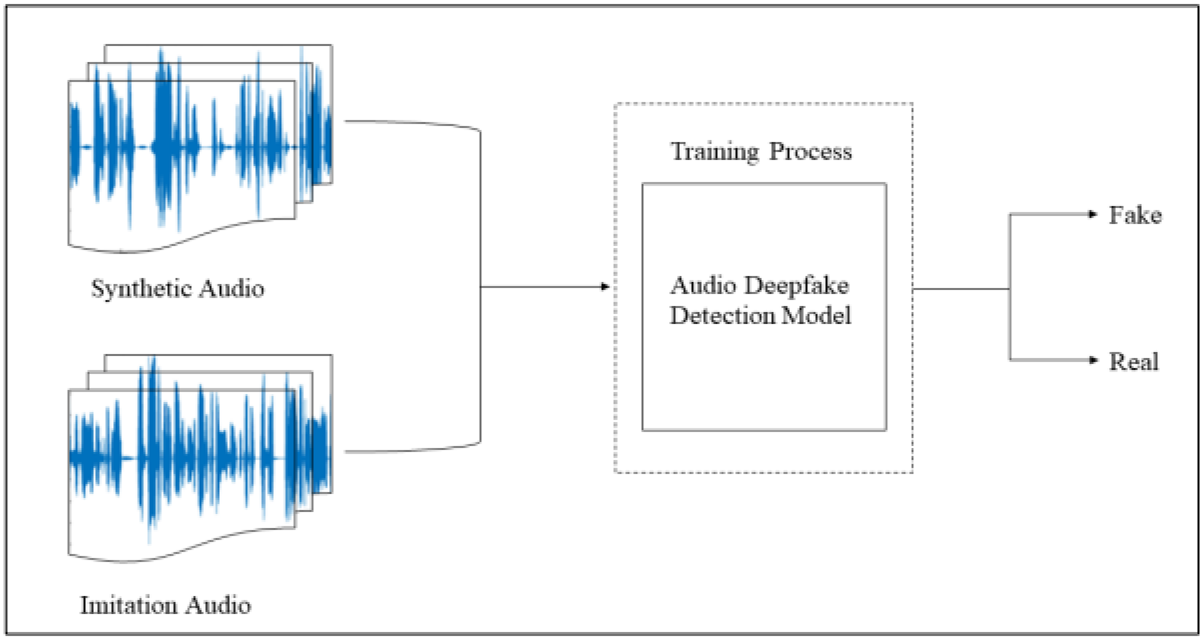
Audio Deepfake

1. Vấn đề

- AI-synthesized tools have recently been developed with the ability to generate convincing voices

2. Recent method to detect AD



- 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.

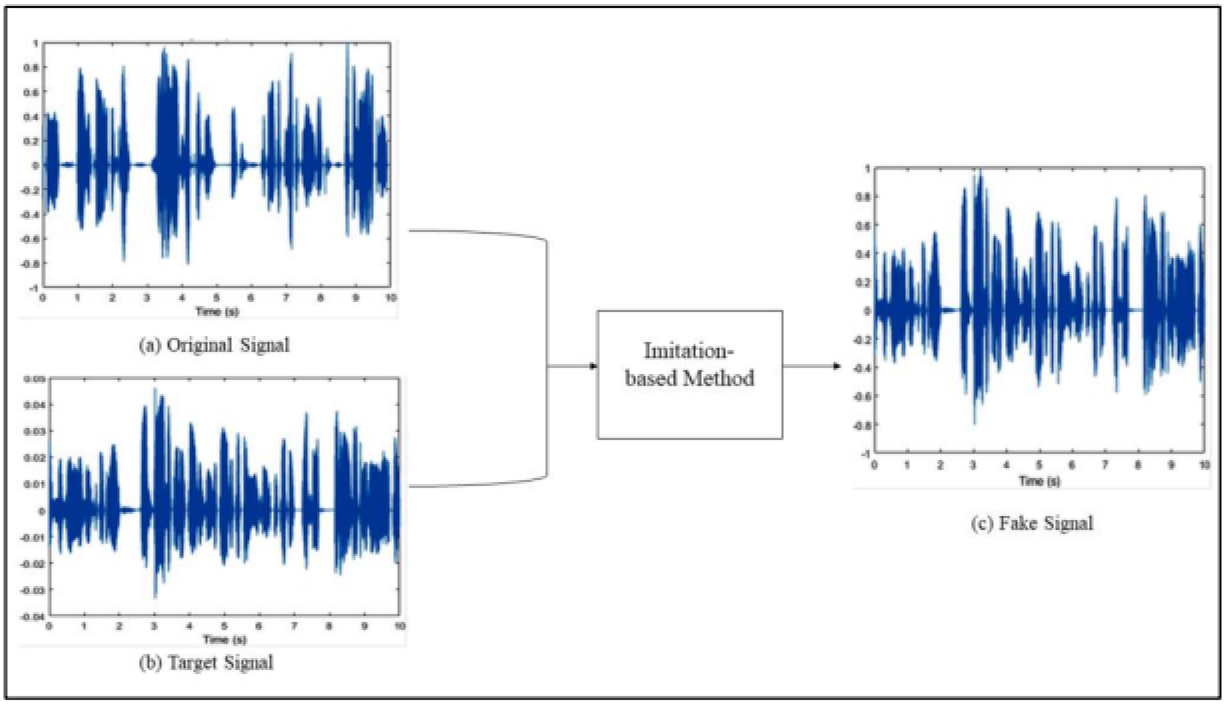
3. Type of Audio Deepfake

- imitation-based
- synthetic-based
- replay based Deepfakes

3.1 Imitation based

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

- Voices can be imitated in different ways, for example, by
  using humans with similar voices who are able to imitate the original speaker
- Using algorithm: Efficient Wavelet Mask (EWM)



(a) Original Signal

(b) Target Signal

(c) Fake Signal

- 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

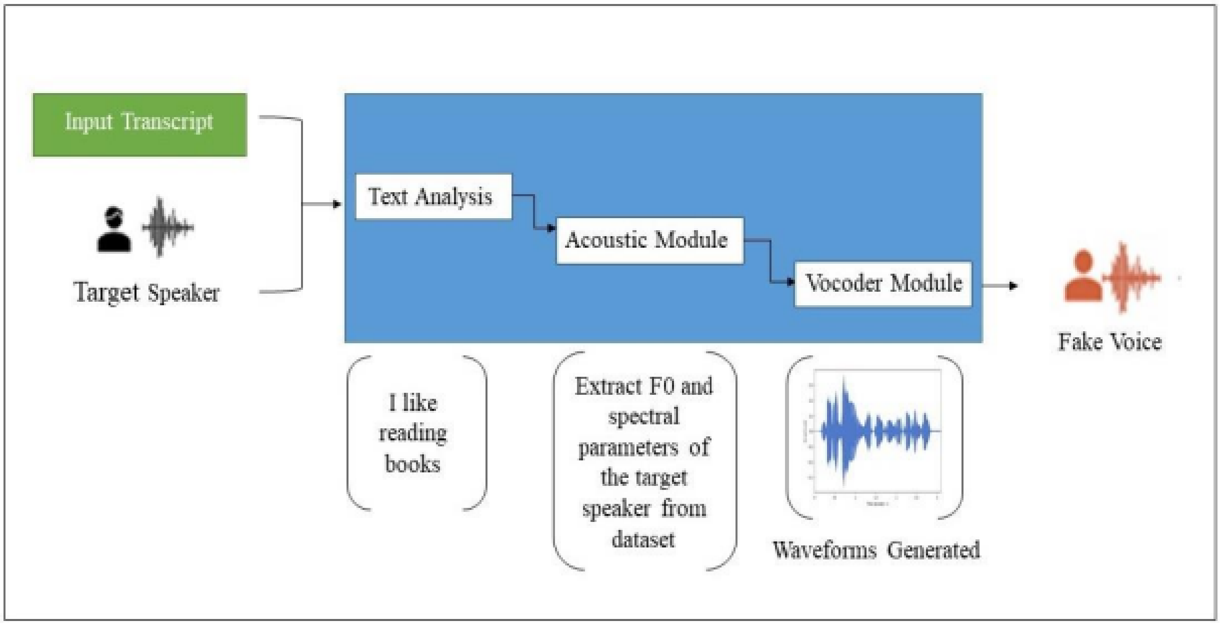
3.2 Synthetic base or Text-to-Speech

Aims to transform text into acceptable and
natural speech in real time and consists of three modules:

- text analysis model
- acoustic model
- vocoder

To generate synthetic Deepfake audio, 2 crucial steps should be followed

- First, clean and structured raw audio should be collected, with a transcript text of the audio speech
- Using model to train such as: Tactoran 2, Deep Voice 3, FastSpeech 2



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 accoustic feature parameters, and the final audio file will be generated, which includes the synthetic
fake audio in a waveform format

3.3 Replay based

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-andpaste 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 cons

Focus on 3.1 and 3.2

4. Fake audio detection methods

Types: ML and DL methods

4.1 ML

- Xây dựng fake audio dataset based on imitation method bằng cách extract entropy features của real và fake audio. Sử dụng H-Voice dataset + Model Logistic Regression, Các model khác như Q-SVM, KNN, STLT, ….

⇒ ML cần extract feature thủ công và tiền xử lý chuyên sâu ⇒ Mất tg

4.2 DL

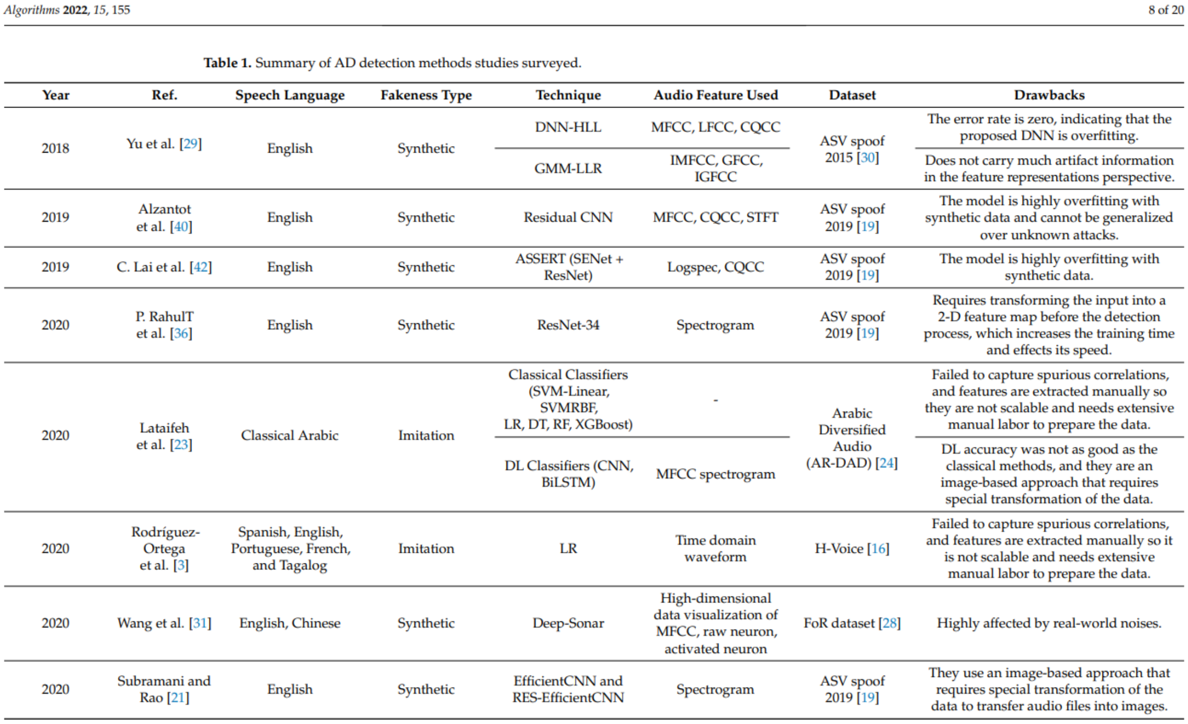


Table 1. Summary of AD detection methods studies surveyed.

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.

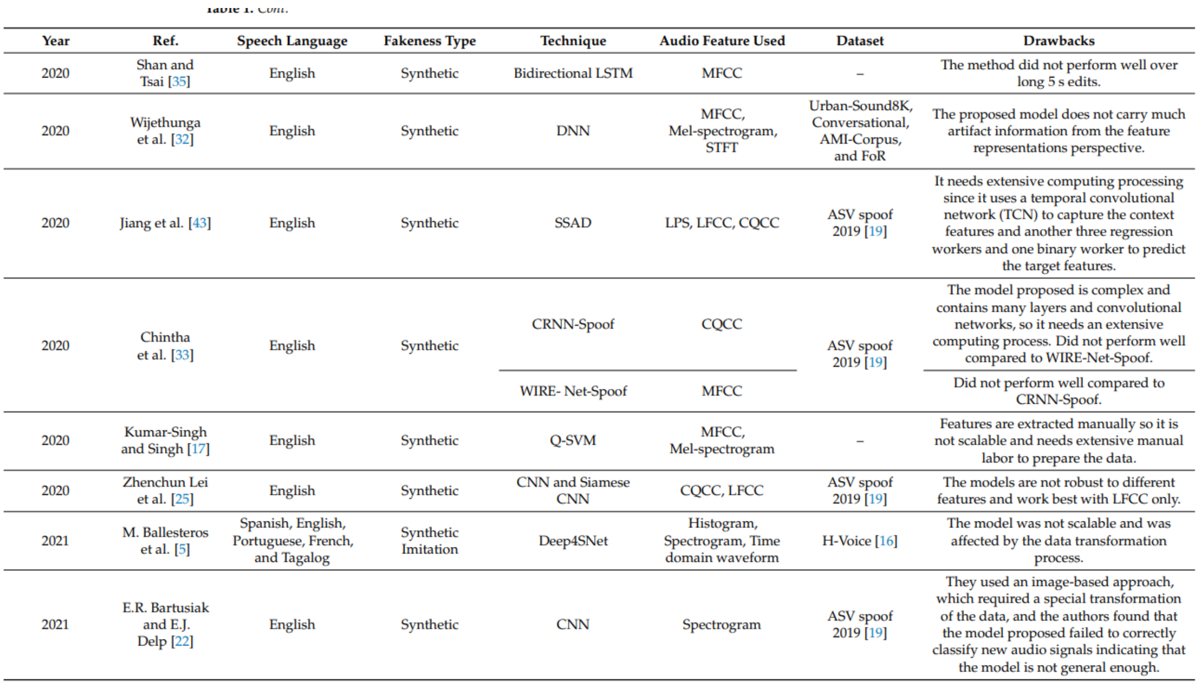
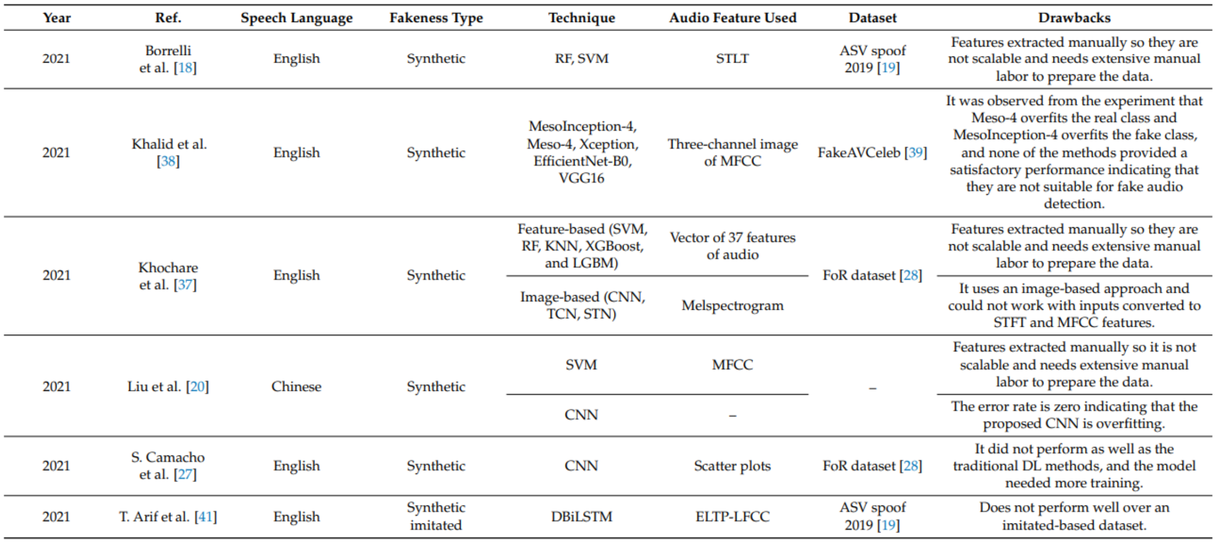


Table 1. Cont.

Year	Ref.	Speech Language	Fakeness Type	Technique	Audio Feature Used	Dataset	Drawbacks
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.
				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.

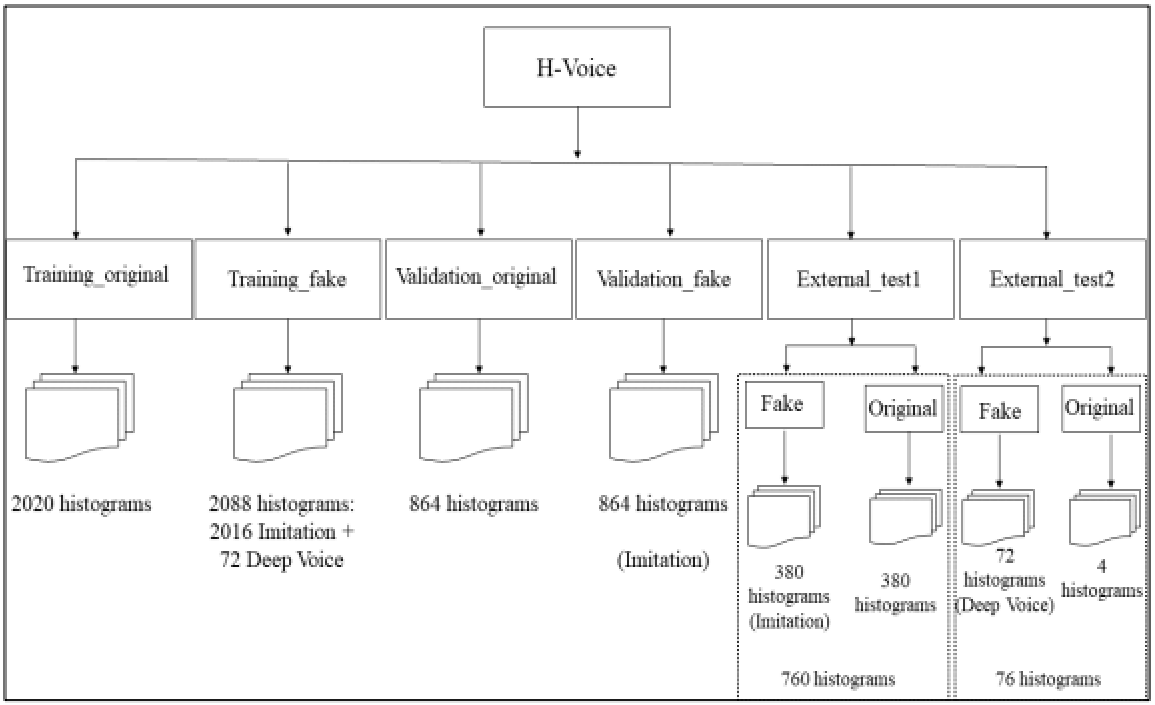
Table 1. Cont.



Year	Ref.	Speech Language	Fakeness Type	Technique	Audio Feature Used	Dataset	Drawbacks
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.

5. Fake Audio Detection Datasets

H-Voice: base on imitation vs synthetic voices speaking English, Spanish, Portuguese, French, Tagalog



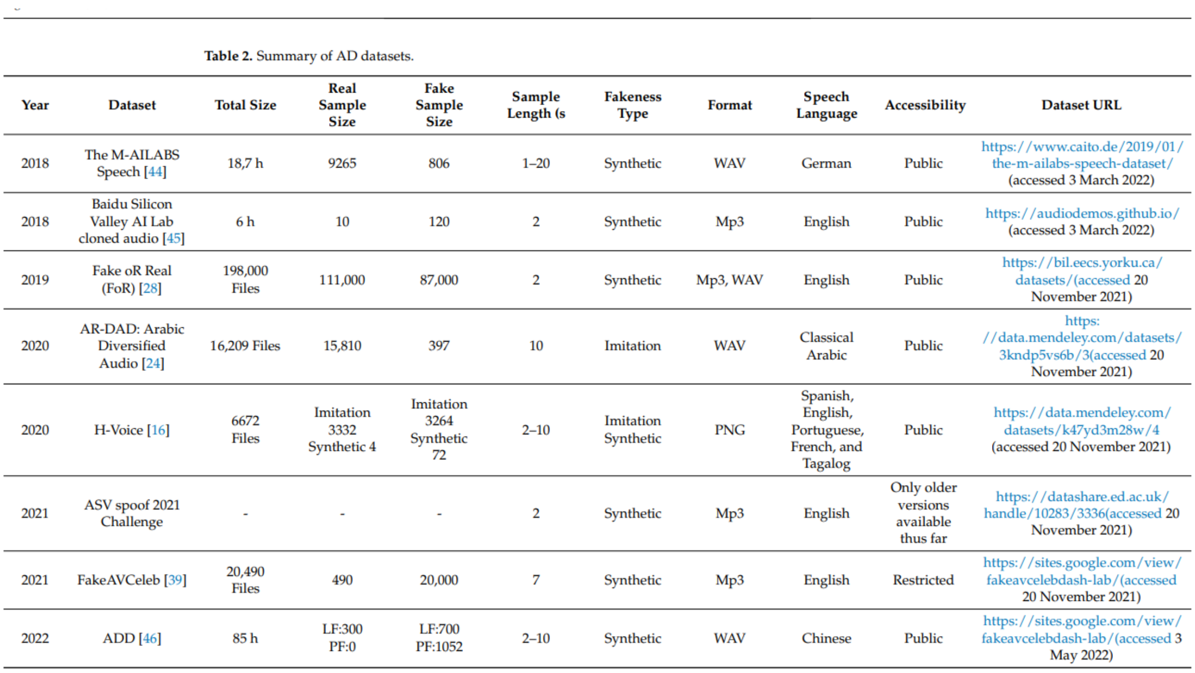


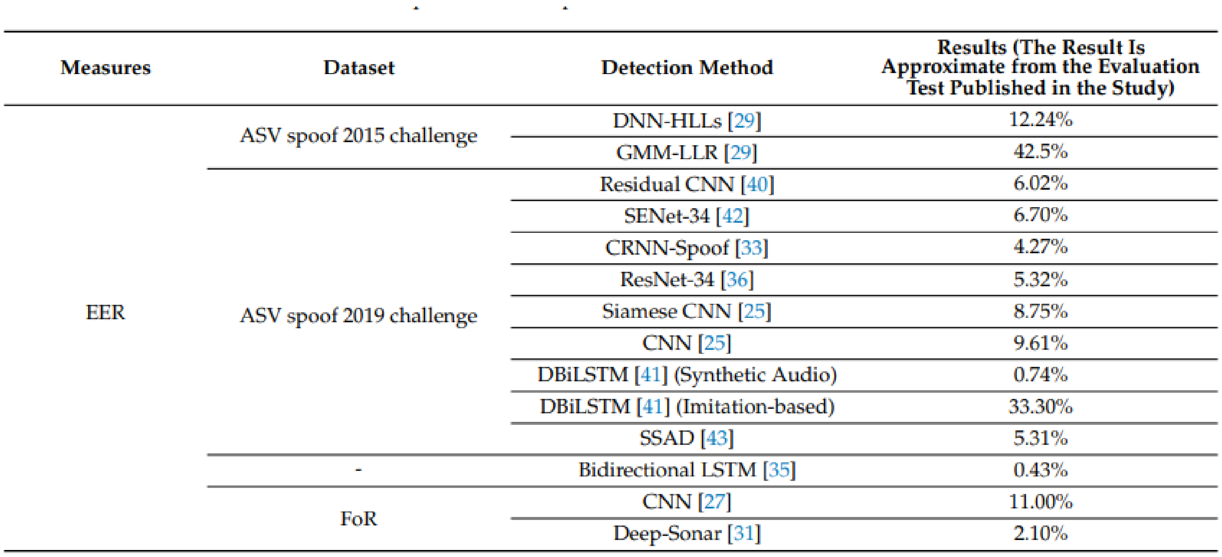
Table 2. Summary of AD datasets.

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)

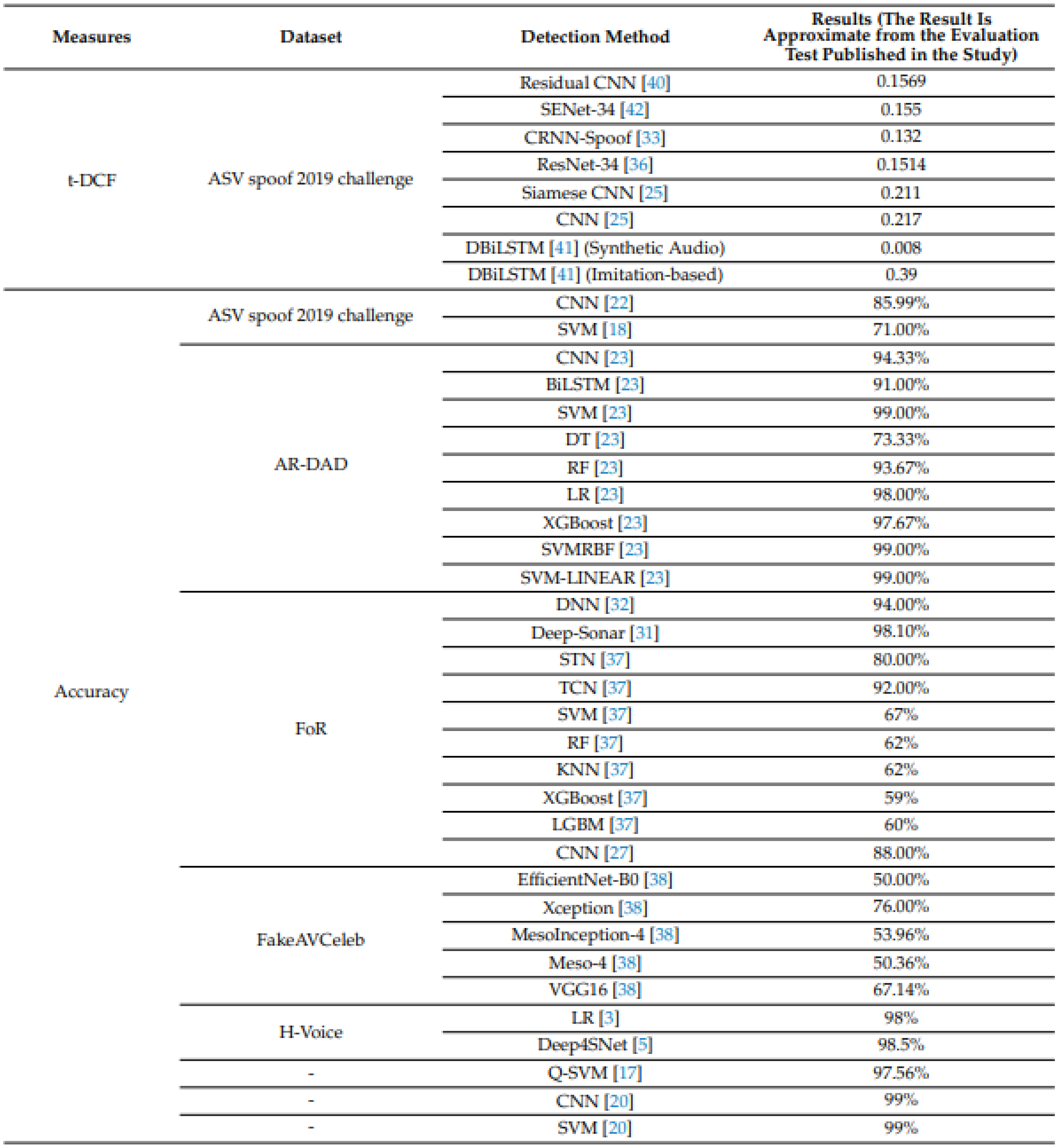
Hvoice: https://data.mendeley.com/datasets/k47yd3m28w/4

Fake oR Real(FoR): https://bil.eecs.yorku.ca/datasets

6. Summary



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%
	FoR	CNN [27]	11.00%
		Deep-Sonar [31]	2.10%



Measures	Dataset	Detection Method	Results (The Result Is Approximate from the Evaluation Test Published in the Study)
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%

7. Tóm tắt

3 loại AD nhưng tập trung vào 2 loại là:

- imitation (voice + voice → voice)
- synthetic (text + voice → voice)

Dạng đầu khá giống Neural style transfer → Xử lý data 1 loại

Dạng 2 thì kết hợp cả text-to-speech → Xử lý data 2 loại

Phương pháp để detect:

- Đơn giản nhất là dùng ML- Logistic
- DL- CNN bth..

Có nhiều loại dataset nhưng chủ yếu là English, các dataset chủ yếu là dạng synthetic fake, có tập H-voice là có cả synthetic lẫn imitation