Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Reality Check: Measuring Real-World Applicability of State-of-the-Art Audio Deepfake Detectors on Social Media Data

Karla Schäfer[1,2], **Martin Steinebach**[1,2]
[1]Fraunhofer Institute for Secure Information Technology
[2]National Research Center for Applied Cybersecurity
karla.schaefer@sit.fraunhofer.de, martin.steinebach@sit.fraunhofer.de

Abstract

Audio deepfakes are becoming both more realistic and easier to create. At the same time, several audio deepfake detectors have been developed. While some of these have been evaluated using real-world data, there has been no in-depth analysis of their performance in real-world settings. We evaluate five SOTA detectors on two real-world social media datasets. Currently, the equal error rate (EER) is mostly used to evaluate audio deepfake detectors. However, when using the EER, the threshold for classifying whether a recording is genuine or not is calculated based on the prediction scores of the test set. In real-world scenarios, this threshold must be set in advance. We are the first to test the performance of SOTA detectors using varying, beforehand set, thresholds, thereby creating a real-world setting. We found degradations on the ITW test set (e.g. F1: 91.35%- 64.65%) when using other thresholds as set with EER calculation. The SocialDF dataset was found to be especially challenging, with an F1-score of 52.92% achieved using an EER threshold. Using pre-set thresholds resulted in an even lower performance of 50.89%, demonstrating that current detectors are unable to reliably detect real-world audio deepfakes.

Introduction

In social media, deepfakes have emerged as a powerful tool for spreading misinformation and disinformation. Malicious actors can use deepfakes to portray individuals saying or doing things they never actually did. These false narratives are designed to provoke emotional responses, reinforce existing biases, or exploit societal divisions, thereby amplifying the spread of misinformation (Chesney and Citron 2019). Recently, deepfakes have become more realistic. Distinguishing between genuine and manipulated content becomes harder (Gosse and Burkell 2020). Therefore, reliable detectors have to be developed. Several audio deepfake detectors have recently been developed, and they are achieving increasingly accurate results. For example, incorporating self-supervised learning (SSL) as feature extraction method in the detection pipeline led to improvements on unknown data (Tak et al. 2022). Several works (Wang et al. 2023; Müller, Sperl, and Böttinger 2023; Wang and Yamagishi 2022; Xie et al. 2025; Wang and Yamagishi 2024; Pianese et al. 2024) used the in-the-wild dataset (ITW) to test the performance of the detectors on real-world social media data. ITW (Müller et al. 2022) was published 2022 and contains real recordings and audio deepfakes of celebrities, extracted from social media. SocialDF (Batra et al. 2025) is a more recently created dataset (2025). SocialDF contains, besides audio and video deepfakes, also metadata. This allows further possibilities for future research, possibly incorporating multi-modal features for deepfake detection. In this work, we will focus on audio data.

Audio deepfake detectors, normally, provide a score (prediction score) for a given recording. Based on this score, the recording is classified as spoof or bona-fide. For the classification, a threshold has to be set. In recent work, mostly, this threshold is calculated during equal error rate (EER) calculation. While being a good metric for balancing the amount of false acceptance and false rejections, the threshold for classification is calculated during evaluation. For a detector being applied in real-world scenarios, this threshold has to be set beforehand. In this work, we provide an overview of the performance of five state-of-the-art detectors on social media data using four different thresholds. All detectors were trained on the newly created ASVspoof5 dataset (Wang et al. 2024) and, besides the test on social media data, also tested on the ASVspoof5 evaluation split (ASV5 eval). As thresholds, we used (1) the threshold of the EER calculation, as baseline, and (2) the threshold calculated on ASV5 eval using the EER calculation. Furthermore, we transformed the prediction scores (logits) to probabilities using the sigmoid function. Using the probabilities we calculated the detectors' performances using (3) 50% as threshold (logits=0) and (4) the threshold when calculating the maximum F1-score on ASV5 eval.

Two research questions will be analysed in this paper: **RQ1:** How effective are SOTA detectors in identifying up-to-date deepfakes obtained from social media? **RQ2:** To what extent does the choice of threshold influence performance evaluation? Therefore, to what extent do laboratory conditions (using EER as a performance benchmark) reflect real-world scenarios in which threshold values must be set in advance?

Experimental Methodology

Five audio deepfake detectors were tested on three test sets. For training, we used the official provided repositories. We

examined RawNet3 (Jung et al. 2022b) and AASIST (Jung et al. 2022a) as often used baseline in audio deepfake detection (ADD). Furthermore, AASIST with Wav2Vec2 (Tak et al. 2022) and the Whisper-based detector (Kawa et al. 2023) were selected because they have shown to generalize better, probably due to their SSL-based front-end. Lastly, Nes2Net (Liu et al. 2025) was incorporated due to its recent emergence and its promising preliminary results. For Nes2Net a pretrained checkpoint of the model trained on ASVspoof5 is available online[1], which we used. The other four detectors were trained by us.

For the Whisper-based detector, Kawa et al. (2023) tested different combinations of feature input and classifier. We used the best performing combination, being a combination of Whisper (fine-tuned) and MFCC with its delta and delta delta and MesoNet as classifier. For AASIST with Wav2Vec2 (Tak et al. 2022) we used, as in the original paper, the model Wav2Vec2.0 XLSR 300m (Babu et al. 2022) for feature extraction. For RawNet3, AASIST with Wav2Vec2, and the Whisper-based detector we set the learning rate to 10^{-6}. For AASIST we used a learning rate of 10^{-4}. Following the original implementation, AASIST, AASIST with Wav2Vec2 and Nes2Net were trained with cross entropy loss. The Whisper-based detector and RawNet3 were trained with BCE with logits loss. We trained all detectors for 100 epochs with a batch size of 24. The training dataset ASVspoof5 is highly unbalanced. Therefore, we weighted the underrepresented class (bona-fide) higher during training (1:9). After each epoch, the models were tested on the ASVspoof5 development set, the model with the best EER on the development set was used for final testing. The audio recordings were pre-processed using padding, resulting in an input of 64600 samples (~ 4 seconds).

Datasets For training and in-domain testing we used the ASVspoof5 (Wang et al. 2024) dataset with its original splits, containing SOTA audio deepfakes from a variety of generation methods. For testing the performance of the detectors for its real-world performance on social media data, we used the widely applied ITW dataset and the newly released SocialDF dataset.

SocialDF (Batra et al. 2025) was introduced as benchmark dataset for deepfake content on social media platforms. The dataset contains mainly content from Instagram, and 6 YouTube links in the bona-fide part. Only links are given, the content has to be scraped by oneself. Overall, 506 links to videos labelled as real and 1,066 links to deepfakes were available. While scraping, we encountered problems because some videos had been deleted, while others were only available after log in. For privacy and legal reasons, these videos were not included in the data set used here for evaluation. Consequently, our SocialDF dataset consists of 967 deepfakes (instead of 1,055) and 497 real Instagram videos and 6 real YouTube videos. SocialDF contains video deepfakes. We extracted the audio and split it in the speech and accompaniment part using spleeter (Hennequin et al. 2020). The speech part was used in the following for ADD. The speech data consists of recordings with mean lengths of 16.48 seconds (spoof) and 51.85 seconds (bona-fide). Overall, the recordings of SocialDF are longer than the input used to train the models, being 4 seconds. Therefore, in further tests, we tested the performance of the detectors on SocialDF when (1) using the first 4 seconds, (2) the whole recording being split in 4 seconds and calculating the mean over the prediction scores of the splits, (3) using the whole recording being split in 4 sec. with a 50% overlap and again calculation the mean over the prediction scores and (4/5) using the majority vote of the prediction scores of the approaches (3) and (4).

The ITW dataset (Müller et al. 2022) is known in ADD as test set for evaluating detectors on real-world data. The dataset contains real and deepfakes of celebrities, also extracted from social media. The recordings have a mean length of 5.21/3.74 seconds (deepfake/real), fitting perfectly in the trained input length of 4 seconds. Therefore, no further tests were carried out.

Results and Discussion

We evaluated the performance of five detectors using the EER, Accuracy, and F1-Score. Due to place restrictions we only provide EERs and F1-scores. The evaluation results are given as the mean with its standard deviation over three test runs. For real-world usability not only the performance, which will be the main focus in this work, but also the resources needed are important. Therefore, we included information about the size (# parameters) and inference time of the detectors in Table 1. The inference time was calculated on an NVIDIA A100 GPU. Overall, AASIST with Wav2Vec2 (W2V2) and Nes2Net have the highest number of parameters and also needed the longest time for inference. Both models have SSL-based front-ends, which increases their resource requirements. RawNet3 (15.5M) is the smallest model viewed, followed by AASIST (298K parameters).

Viewing the performance in terms of EER, W2V2 performed in three of five settings the best, achieving an EER of 7.35% on the in-domain ASV5 eval set, 10.63% on ITW and 38.12/37.86% on SocialDF (all/all_withOverlap). Nes2Net was superior to W2V2 on ASV5 eval with an EER of 5.93% and AASIST was superior to W2V2 on SocialDF (4 sec. input) with an EER of 47.86%. RawNet3 and the Whisper-based detector were outperformed by the other detectors. Interestingly, for RawNet3 the highest deviations over the three test runs were calculated. For the Whisper-based detector no deviations were calculated, making Whisper the most stable and RawNet3 the most varying detector.

Varying thresholds In Table 2 the F1-Scores on the test sets using different thresholds are given. On ITW, all F1-scores using other thresholds as the threshold of the EER are worse (marked red). On ASV5 eval and SocialDF, using the probability of 50% as threshold worsened the F1-scores using all detectors. Using the threshold of the maximized F1-Score (calculated on ASV5 eval) the F1-scores, unsurprisingly, improved. The same can be viewed on SocialDF, with the exception of RawNet3 and W2V2 as detector. Viewing **RQ2**, overall, the threshold used did impact the performance

[1]https://github.com/Liu-Tianchi/Nes2Net_ASVspoof_ITW/tree/asvspoof5

	Detector	RawNet3	AASIST	AASIST with Wav2Vec2 (W2V2)	Whisper-based	Nes2Net
	Year	2022	2022	2022	2023	2025
	# Parameters	15.5M	298K	318M	7.7M	316M
	Inference time (min)	78	91	169	133	156
Test sets (EER)	ASV5 eval	$38.79_{\pm 0.02}$	$25.32_{\pm 0.04}$	$7.35_{\pm 0.03}$	$25.90_{\pm 0}$	$\mathbf{5.93_{\pm 0.01}}$
	ITW	$62.11_{\pm 3.10}$	$26.08_{\pm 0.07}$	$\mathbf{10.63_{\pm 0.04}}$	$22.85_{\pm 0}$	$13.69_{\pm 0.08}$
	SocialDF (4 sec. input)	$60.58_{\pm 1.39}$	$\mathbf{47.86_{\pm 0.41}}$	$49.62_{\pm 1.86}$	$52.97_{\pm 0}$	$48.62_{\pm 0.49}$
	SocialDF (all)	$49.80_{\pm 0.68}$	$38.88_{\pm 0}$	$\mathbf{38.12_{\pm 0}}$	$50.59_{\pm 0}$	$42.93_{\pm 0}$
	SocialDF (all_withOverlap)	$50.24_{\pm 0.82}$	$38.93_{\pm 0}$	$\mathbf{37.86_{\pm 0}}$	$49.29_{\pm 0}$	$43.84_{\pm 0}$

Table 1: First evaluation of the detectors (EER; %). Inference time is given on the ASV5 eval test set. Best result per test set are highlighted in bold, second best are underlined.

		F1 Score (%)				
Test set	Threshold	RawNet3	AASIST	AASIST with Wav2Vec2 (W2V2)	Whisper-based	Nes2Net
ASV5 eval	EER	$39.13_{\pm 0.02}$	$54.59_{\pm 0.05}$	$83.71_{\pm 0.05}$	$53.83_{\pm 0}$	$\mathbf{86.61_{\pm 0.02}}$
	Prob. (50%)	$36.34_{\pm 0.02}$	$25.02_{\pm 0.01}$	$\mathbf{80.40_{\pm 0.02}}$	$24.48_{\pm 0}$	$68.53_{\pm 0.01}$
	Prob. maxF1	$41.04_{\pm 0.02}$	$55.02_{\pm 0.05}$	$\mathbf{94.38_{\pm 3.34}}$	$53.84_{\pm 0}$	$87.40_{\pm 0.01}$
ITW	EER	$43.37_{\pm 3.23}$	$78.07_{\pm 0.06}$	$\mathbf{91.35_{\pm 0.03}}$	$80.92_{\pm 0}$	$88.79_{\pm 0.06}$
	ASV5	$42.73_{\pm 2.04}$	$48.59_{\pm 0.04}$	$77.56_{\pm 0.05}$	$68.89_{\pm 0}$	$\mathbf{84.77_{\pm 0.16}}$
	Prob. (50%)	$39.64_{\pm 0.10}$	$\mathbf{66.99_{\pm 0.06}}$	$64.65_{\pm 0.14}$	$45.87_{\pm 0}$	$64.75_{\pm 0.07}$
	Prob. maxF1 (ASV5)	$42.91_{\pm 0.73}$	$44.04_{\pm 0.53}$	$\mathbf{88.05_{\pm 0.11}}$	$68.69_{\pm 0}$	$85.85_{\pm 0.11}$
SocialDF (4 sec. input)	EER	$30.79_{\pm 1.25}$	$\mathbf{42.71_{\pm 0.42}}$	$41.02_{\pm 1.80}$	$37.83_{\pm 0}$	$41.96_{\pm 0.46}$
	ASV5	$26.76_{\pm 0.32}$	$\mathbf{48.51_{\pm 0.83}}$	$47.99_{\pm 1.54}$	$42.62_{\pm 0}$	$46.29_{\pm 1.45}$
	Prob. (50%)	$25.65_{\pm 0.18}$	$27.21_{\pm 0.27}$	$\mathbf{30.64_{\pm 0.54}}$	$25.49_{\pm 0}$	$25.88_{\pm 0.28}$
	Prob. maxF1 (ASV5)	$25.97_{\pm 0.06}$	$\mathbf{45.98_{\pm 0.31}}$	$39.68_{\pm 0}$	$42.91_{\pm 0}$	$42.78_{\pm 0.38}$

Table 2: F1-Score (%) with various thresholds. Red/Green: worse/better than using EER threshold. Best result per test set and threshold used are highlighted in bold, second best are underlined. Grey: real-world applicable.

scores, partially, heavily. For example, using the Whisper-based detector on ITW, from an F1-Score of 80.92% using the EER threshold to 45.87% using the 50% threshold, being a degradation of 35.05%.

W2V2 performed over all test sets rather good, being in five settings the best, and in three the second best (of 11 settings tested). See Figure 1 for the distribution of the prediction scores (logits) of W2V2 on all three test sets. On ASV5 eval, using the logits, an F1-score of 83.71% was calculated. Viewing Figure 1 (a) one can see that the logits can be divided comparable good, with several recordings of spoofs (label 0) having rater low scores (hill around -3) while bona-fide recordings being attributed to positive scores (around 3). This good separability is also reflected in the rather good F1-scores. Viewing the performance on ITW in Figure 1 (b) the separability is worse. Most recordings have prediction scores around 2 and 3. Calculating the EER on this test set, a threshold of 3.11 was determined. On ASV5 eval a threshold of 2.51 was calculated, and used as pre-set threshold. This resulted in a degradation of the F1-score of 13.79% (91.35% - 77.56%), but an increase in accuracy (74.12% - 81.62%), possibly due to the higher amount of bona-fide in ITW (see also orange at score 3) which are now correctly identified as bona-fide, whereby also several spoofs are now, wrongly, classified as bona-fide. For SocialDF one can see in Figure 1 (c) that a clear separation cannot be made. Spoofs (label 0) and bona-fide recordings are, mainly, predicted with scores higher 0 and over the whole range. This is also reflected in poor F1-scores of 41.02%/47.99% and accuracies of 35.92%/58.71% (EER threshold/ASV5 threshold).

Figure 2 visualizes the distributions of the logits for AASIST (a) and Nes2Net (b) calculated on the ITW test set. For **Nes2Net**, the thresholds of the EER calculation on the ITW dataset and using the threshold of ASV5 are rather close (4.11/5.09). Still, the best F1-score is calculated with the EER threshold, being 88.79%. Using the threshold of ASV5, the F1-score degrades to 84.77%. The accuracy increased again using the ASV5 threshold. Again, due to the smaller number used as threshold (4.11) more bona-fide recordings are labelled as such, with an increase in spoofs not being detected, resulting in higher accuracies. Using Nes2Net most bona-fide recordings were labelled with a positive score. Whereby, for spoofs, the predictions are distributed over scores from -4 to 6. Similar to W2V2 (see Figure 1 (b)) the bona-fide recordings are identified more clearly, with higher scores, whereby spoofs are more spread out. **AASIST** behaves differently. Two distribution hills of spoofs and bona-fide recordings are built, but, overlapping each other. Also, the threshold calculated on the dataset (using EER) and ASV5 differ more widely (1.94/5.61). leading to a degradation of 29.48% in the F1-Scores (78.07%- 48.59%).

We examined the initial results for answering **RQ1**: How effective are SOTA detectors in identifying up-to-date deepfakes obtained from social media? On ITW we calculated satisfactory results with the best F1-score, using pre-set thresholds, being 88.05% (W2V2; ASV5 MaxF1 probabil-

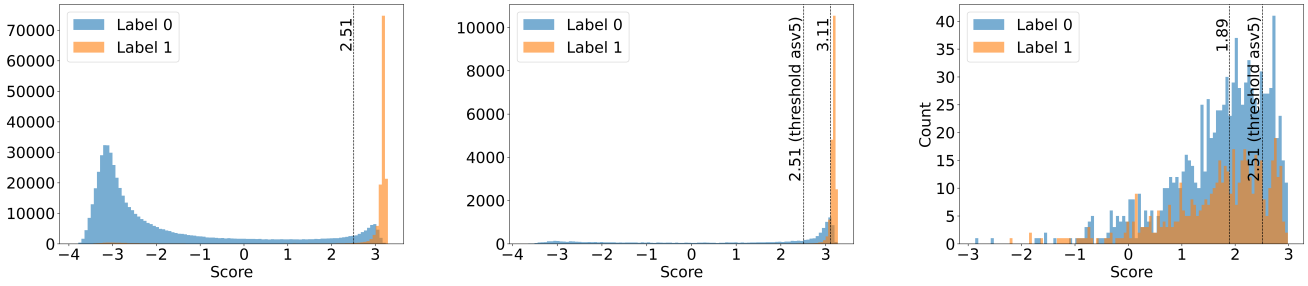

Figure 1: Logits distribution of AASIST with Wav2Vec2 on (a) ASV5 eval (b) ITW (c) SocialDF

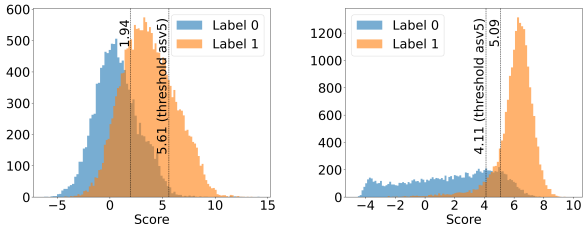

Figure 2: Logits distribution of AASIST (a) and Nes2Net (b) on ITW.

ity). For SocialDF the results are, overall, rather poor. The best F1-score calculated being 48.51% (using AASIST and the ASV5 threshold). Until now, we only used the first 4 seconds of each recording. As the SocialDF recordings are rather long, we will analyse the whole recordings in the following, potentially improving the results.

SocialDF: using whole recordings In Table 3 the F1-Scores on SocialDF using the whole recordings and different combination strategies are given. Interestingly, using the whole recording and setting the threshold at probability 50%, all recordings were classified as bona-fide (marked yellow), resulting in an F1-Score of 25.49%. The performance improved using the whole sample size, but being only slightly over 50% (bold). The best F1-score using a pre-set threshold was 53.11% (using W2V2, maxF1 ASV5 threshold; mean over the predictions of 4 seconds splits of the whole recording). Using the EER threshold the results are similar, with the best F1-Score being 52.92% (W2V2; all_withOverlap). With this, the results on SocialDF are rather poor. SocialDF consists, potentially, of more up-to date audio deepfakes which could have led to this rather poor performance, compared to ITW. Another explanation of these results could be post-processing techniques, such as transmission and compression commonly applied on social media recordings. Although some post-processing steps are part of the ASVspoof5 dataset, on which the detectors were trained, there are constant improvements and additional methods that may not have been included in the training set. Future work should examine SocialDF more closely, as it highlights potential new difficulties for detecting audio deepfakes in social media.

SocialDF using...	Threshold	AASIST	W2V2	Nes2Net
first 4 seconds	EER	$42.71_{\pm 0.42}$	$41.02_{\pm 1.80}$	$41.96_{\pm 0.46}$
	ASV5	$48.51_{\pm 0.83}$	$47.99_{\pm 1.54}$	$46.29_{\pm 1.45}$
	Prob. (50%)	$27.21_{\pm 0.27}$	$30.64_{\pm 0.54}$	$25.88_{\pm 0.28}$
	maxF1 (ASV5)	$45.98_{\pm 0.31}$	$39.68_{\pm 0}$	$42.78_{\pm 0.38}$
all	EER	$\mathbf{51.85_{\pm 0}}$	$\mathbf{52.66_{\pm 0}}$	$47.64_{\pm 0}$
mean	ASV5	$42.22_{\pm 0.17}$	$30.02_{\pm 0.06}$	$49.06_{\pm 0.24}$
	Prob. (50%)	$25.49_{\pm 0}$	$25.49_{\pm 0}$	$25.49_{\pm 0}$
	maxF1 (ASV5)	$49.86_{\pm 0.84}$	$\mathbf{53.11_{\pm 0.03}}$	$40.67_{\pm 0.68}$
all	ASV5	$40.29_{\pm 0.10}$	$28.37_{\pm 0.06}$	$\mathbf{51.85_{\pm 0.10}}$
majority vote	Prob. (50%)	$25.49_{\pm 0}$	$25.49_{\pm 0}$	$25.49_{\pm 0}$
	maxF1 (ASV5)	$47.07_{\pm 1.02}$	$43.79_{\pm 0.05}$	$41.06_{\pm 0.92}$
all_withOverlap	EER	$\mathbf{51.77_{\pm 0}}$	$\mathbf{52.92_{\pm 0}}$	$46.74_{\pm 0}$
(50%) mean	ASV5	$42.47_{\pm 0.10}$	$28.30_{\pm 0}$	$46.98_{\pm 0.11}$
	Prob. (50%)	$25.49_{\pm 0}$	$25.49_{\pm 0}$	$25.49_{\pm 0}0$
	maxF1 (ASV5)	$48.57_{\pm 0.79}$	$\mathbf{50.89_{\pm 0}}$	$40.65_{\pm 0.35}$
all_withOverlap	ASV5	$40.49_{\pm 0.15}$	$27.63_{\pm 0}$	$48.21_{\pm 0.03}$
(50%)	Prob. (50%)	$25.49_{\pm 0}$	$25.49_{\pm 0}$	$25.49_{\pm 0}$
majority vote	maxF1 (ASV5)	$46.81_{\pm 0.74}$	$42.96_{\pm 0.05}$	$40.62_{\pm 0.45}$

Table 3: F1-Scores (%) on SocialDF. Bold: better than 50%. Yellow: all samples were classified as bona-fide.

Conclusion

We analysed five detectors on their performance in detecting audio deepfakes in social media. W2V2 and Nes2Net performed overall the best. On ITW, satisfactory results were calculated, even with pre-set thresholds. Still, performance degradations were visible when using other thresholds as the one calculated with the EER, leading to further questions about the real detection performance of other audio deepfake detectors in research, as mostly only EERs are given as performance measures in related work. It is evident that there is a discrepancy between the performance levels observed in laboratory settings and those demonstrated in real-world scenarios. Therefore, we recommend using evaluation metrics other than just the EER when testing audio deepfake detectors. We also recommend conducting experiments with different thresholds and reporting these results whenever a new detector is introduced. This will ensure that the detectors are truly usable and can deliver good results on new, unseen data. SocialDF was identified as particularly challenging dataset, showing that even SOTA detectors still cannot reliably detect audio deepfakes in the real-world.

Acknowledgments

This research work was supported by the National Research Center for Applied Cybersecurity ATHENE in the project DREAM.

References

Babu, A.; Wang, C.; Tjandra, A.; Lakhotia, K.; Xu, Q.; Goyal, N.; Singh, K.; Von Platen, P.; Saraf, Y.; Pino, J.; et al. 2022. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *Interspeech*.

Batra, A.; Khemani, J.; Gumber, A.; Kumar, A.; Jain, A.; and Gupta, S. 2025. SocialDF: Benchmark Dataset and Detection Model for Mitigating Harmful Deepfake Content on Social Media Platforms. In *Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation*, 81–89.

Chesney, B.; and Citron, D. 2019. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107: 1753.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gosse, C.; and Burkell, J. 2020. Politics and porn: how news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication*, 37(5): 497–511.

Hennequin, R.; Khlif, A.; Voituret, F.; and Moussallam, M. 2020. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50): 2154. Deezer Research.

Jung, J.-w.; Heo, H.-S.; Tak, H.; Shim, H.-j.; Chung, J. S.; Lee, B.-J.; Yu, H.-J.; and Evans, N. 2022a. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 6367–6371. IEEE.

Jung, J.-w.; Kim, Y. J.; Heo, H.-S.; Lee, B.-J.; Kwon, Y.; and Chung, J. S. 2022b. Pushing the limits of raw waveform speaker recognition. *Proc. Interspeech*.

Kawa, P.; Plata, M.; Czuba, M.; Szymanski, P.; and Syga, P. 2023. Improved DeepFake Detection Using Whisper Features. *Interspeech*.

Liu, T.; Truong, D.-T.; Kumar Das, R.; Aik Lee, K.; and Li, H. 2025. Nes2Net: A Lightweight Nested Architecture for Foundation Model Driven Speech Anti-Spoofing. *IEEE Transactions on Information Forensics and Security*, 20: 12005–12018.

Müller, N. M.; Czempin, P.; Dieckmann, F.; Froghyar, A.; and Böttinger, K. 2022. Does Audio Deepfake Detection Generalize? *Interspeech*.

Müller, N. M.; Sperl, P.; and Böttinger, K. 2023. Complex-valued neural networks for voice anti-spoofing.

Pianese, A.; Cozzolino, D.; Poggi, G.; and Verdoliva, L. 2024. Training-Free Deepfake Voice Recognition by Leveraging Large-Scale Pre-Trained Models. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, IHMMSec '24, 289–294. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706370.

Tak, H.; Todisco, M.; Wang, X.; Jung, J.-w.; Yamagishi, J.; and Evans, N. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop*.

Wang, C.; Yi, J.; Tao, J.; Zhang, C.; Zhang, S.; and Chen, X. 2023. Detection of Cross-Dataset Fake Audio Based on Prosodic and Pronunciation Features. *Interspeech*.

Wang, X.; Delgado, H.; Tak, H.; Jung, J.-w.; Shim, H.-j.; Todisco, M.; Kukanov, I.; Liu, X.; Sahidullah, M.; Kinnunen, T.; et al. 2024. ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. *Proc. The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*.

Wang, X.; and Yamagishi, J. 2022. Spoofed Training Data for Speech Spoofing Countermeasure Can Be Efficiently Created Using Neural Vocoders. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Wang, X.; and Yamagishi, J. 2024. Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end? In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10311–10315. IEEE.

Xie, Y.; Lu, Y.; Fu, R.; Wen, Z.; Wang, Z.; Tao, J.; Qi, X.; Wang, X.; Liu, Y.; Cheng, H.; et al. 2025. The codecfake dataset and countermeasures for the universally detection of deepfake audio. *IEEE Transactions on Audio, Speech and Language Processing*.

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, because it does not mention any specific identifiable information

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? No

   (e) Did you describe the limitations of your work? Yes

   (f) Did you discuss any potential negative societal impacts of your work? Yes

   (g) Did you discuss any potential misuse of your work? No

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? No, the work rather highlights problems with existing works.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No, but the training is based on the scripts used by the creators of the detectors. The corresponding links to the GitHub repositories have been omitted for space reasons, but can be found in the papers cited.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, or partially. The inference time of the various detectors (including hardware) was specified.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? Yes, see papers.

(c) Did you include any new assets in the supplemental material or as a URL? No

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, I used pre-existing datasets.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? No, no new dataset was created.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? No, no new dataset was created.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA