Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 9, 2025

AI BASED DEEPFAKE AUDIO DETECTION – A REVIEW

Muhammad Aleem[1], Saqib Riaz[2], Muhammad Tayan Aziz[3], Engr. Dr Abdul Rehman Chishti[4]

The islamia university of Bahawalpur
The islamia university of Bahawalpur
The islamia university of Bahawalpur
Department of Information and communication Engineering, (BS Cyber Security and Digital Forensics) , The Islamia University of Bahawalpur(IUB)
muhammadaleem3255@gmail.com, saqibriaz554@gmail.com, tayankhan1122@gmail.com, rehman.chishti@iub.edu.pk

DOI: https://doi.org

Keywords : Deepfake audio, voice cloning, audio forensics, machine learning, MFCC, spectrogram, deepfake detection, Random Forest, convolutional neural networks (CNN).

Article History
Received on 28 July 2025
Accepted on 19 August 2025
Published on 09 September 2025

Corresponding Author: *
Muhammad Aleem

Abstract
The rapid rise of deepfake audio presents a dual reality: enabling innovative applications like voice assistants and accessibility tools, while also posing severe risks to security and trust through fraud and misinformation. Modern systems can clone a voice from just a few seconds of audio, making it hard to distinguish real from synthetic speech. This study investigates machine learning methods for detecting deepfake audio, using features such as MFCCs and spectrograms with classifiers including Random Forests and CNNs on datasets like FoR and ASVspoof. Results show that combining optimized features with advanced models significantly boosts detection accuracy. We also address ongoing challenges like limited data diversity, adversarial attacks, and real-world scalability, alongside ethical concerns. Our goal is to contribute to the development of reliable and practical detection systems.

INTRODUCTION

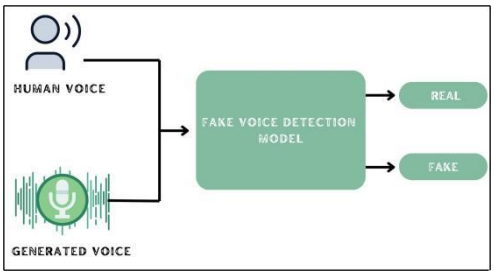

Figure 1: Spectral texture representation used in deepfake audio analysis.

1 Introduction

1.1 Background and Motivation

Deepfake technology, once primarily associated with the manipulation of visual media, has now advanced into the domain of speech synthesis. Modern algorithms are capable of producing highly realistic audio that closely mirrors the speech characteristics of real individuals, including pitch, tone, rhythm, and even emotional expression. This capability, known as audio deepfakes, has enabled applications such as personalized voice assistants, audiobooks, and entertainment media. However, the same technology has also facilitated harmful practices. Fraudsters have used synthetic voices to impersonate corporate executives in order to authorize fraudulent financial transfers, while political deepfakes have been employed to manipulate public opinion. The increasing accessibility of generative tools such as VALL-E, Resemble.AI, and Descript has lowered the barrier to entry, allowing individuals with minimal technical expertise to create convincing synthetic voices [1], [2]. These developments highlight the urgency of advancing reliable detection systems that can protect users, organizations, and society at large from the risks posed by manipulated audio.

1.2 Problem Statement

While voice authentication technologies are widely used in sectors such as finance, telecommunications, and smart assistants, they are increasingly vulnerable to manipulation by deepfake audio. Traditional anti-spoofing techniques of- ten fail to detect synthetic voices gener- ated through advanced architectures such as WaveNet, Tacotron 2, and GAN-based vocoders [4][5]. The ability of these models to capture subtle prosodic and spectral features of speech makes human and machine detection particularly challenging. Moreover, current detection methods face limitations in terms of dataset diversity, robustness in noisy environ- ments, and adaptability to unseen types of attacks. In forensic contexts, dis- tinguishing between authentic and fake audio is even more critical, as manipu- lated recordings can compromise legal in- vestigations and judicial processes. De- spite progress in machine learning and deep learning–based classification meth- ods, the absence of standardized bench- marks, explainable models, and scalable real-time solutions underscores the press- ing need for further research and deveopment in this area [29].

1.3 Research Objectives

To address the gaps outlined above, this study sets out the following research objectives:

- Examine existing methodologies for

deepfake audio generation and detection, with emphasis on machine learning and deep learning approaches. the role of feature extraction techniques, particularly Mel-Frequency Cepstral Coefficients (MFCCs), spectrograms, and chro- magrams, in capturing distinctive patterns of synthetic speech [11].

- Evaluate the effectiveness of Random Forest classifiers when combined with MFCC-based features, comparing performance against alternative baseline models [9][10].
- Assess limitations and ethical considerations, including dataset constraints, adversarial attacks, privacy risks, and societal implications of detection tools. [1]
- Identify future research directions that emphasize hybrid ML-DL models, cross-lingual robustness, realtime scalability, and interpretability for forensic applications. [12]

By addressing these objectives, this work contributes to the broader effort of designing secure, interpretable, and future-proof frameworks for differentiating between authentic and manipulated speech.

2 Literature Review

2.1 *Evolution of Deepfake Audio Technology*

The development of synthetic audio has transitioned from rule-based concatena- tive text-to-speech (TTS) systems to modern neural vocoders capable of gener- ating speech that is nearly indistinguish- able from human voices. Early systems such as Hidden Markov Model (HMM)- based TTS were limited in naturalness and adaptability, while breakthroughs like Google's *WaveNet* and *Parallel WaveGAN* enabled high-fidelity wave- form synthesis with improved prosody and intonation. More recently, *self-supervised learning frameworks* such as Wav2Vec 2.0 and Microsoft's VALL-E have made it possible to generate realistic voices from only a few seconds of audio input. These advancements have enabled legitimate applications such as audiobooks, digital assistants, and accessibility tools, but also created opportunities for misuse in social engineering, fraud, and disinformation campaigns [27].

2.2 *Categories of Audio Deepfakes*

Audio deepfake techniques are broadly categorized into three main approaches: **Replay Attacks**, **Speech Synthesis**, and **Voice Conversion**.

- **Replay Attacks:** This involves reusing or replaying pre-recorded audio of a target speaker. They are further classified into *far-field replay* (captured from speakers and microphones in real-world conditions) and *copy-paste replay* (direct splicing of recordings) [3]. Recent studies have employed *deep convolutional networks* to detect replay-based manipulations, reporting Equal Error Rates (EER) close to 0% on datasets like ASVspoof2017 [14].
- **Speech Synthesis (SS):** Modern SS systems rely on deep neural networks to generate speech from text. Notable frameworks include *Tacotron 2*, which

combines an attention-based recurrent sequence- to-sequence model with a modified WaveNet vocoder, and *WaveGlow*, which replaces the traditional two- stage TTS pipeline with an end-to- end generative flow model. Commercial tools such as *Lyrebird* have demonstrated the ability to syn- thesize thousands of sentences per second. GAN-based architectures have also been proposed for TTS, with researchers generating large synthetic datasets comprising hundreds of thousands of high-quality audio samples. [5]

- **Voice Conversion (VC):** This technique modifies a source speaker's voice to resemble that of a target speaker, preserving linguistic content while altering vocal identity. VC models leverage spectral mapping, pitch shifting, and generative adversarial networks to achieve convincing impersonation. [6]

2.3 *Audio Feature Extraction for Detec- tion*

Detecting manipulated speech requires extracting features that highlight differences between genuine and synthetic audio. Table 1 summarizes commonly used features.

Table 1: Comparison of Audio Features for Deepfake Detection

Feature	Advantages	Limitations
MFCCs	Captures vocal spectral proper- ties; efficient computa- tion	Misses temporal patterns
Spectro- gram	Rich time- frequency detail	High- dimension al; comput- atnally heavy
Pitch (F0)	Detects unnatural prosody	Sensitive to noise
Chroma	Useful in tonal/music al contexts	Less effective for speech

2.4 *Machine Learning and Deep Learn- ing Models for Detection*

Research on deepfake audio detection has leveraged both traditional machine learning and deep learning methods:

- **Traditional ML Approaches:** Random Forest classifiers are valued for robustness and interpretabilit. Support Vector Machines (SVMs) have shown effectiveness on smaller datasets, though they scale poorly. Gradient Boosting methods (e.g., XGBoost) provide strong accuracy and have been applied to various forensic audio tasks. [8]

- **Deep Learning Approaches:** Convolutional Neural Networks (CNNs) are effective in analyzing

spectrograms [14], while Recurrent Neu- ral Networks (RNNs), particularly Long Short-Term Memory (LSTM) models, capture temporal dependencies in audio [15]. Transformer-based architectures have more recently set state-of-the-art performance by learning global attention patterns across speech sequences. [17]

2.5 *Limitations of Current Methods and Research Gaps*

Despite notable progress, existing detec- tion frameworks face multiple challenges:

- **Dataset Limitations:** Many benchmark datasets (e.g., ASVspoof, FoR) do not cover the full spectrum of attack scenarios, languages, and recording conditions [20].
- **Scalability:** Deep learning methods often demand significant computa- tional resources and training time, limiting their deployment in real- time applications [28].
- **Robustness in Adverse Conditions:** Detection accuracy deteriorates in noisy environments, compressed audio formats, or when tested on previously unseen generative models [20].
- **Ethical and Legal Challenges:** The forensic use of deepfake detection tools requires explainable models that can withstand legal scrutiny [30].

Thus, while current machine learning and deep learning approaches provide promising accuracy, further research is required to design **hybrid systems**, improve **cross-domain generalization**, and enhance **adversarial robustness** for real-world deployment [29].

3 Proposed Methodology

Machine learning models for deepfake audio detection face challenges such as overfitting, underfitting, and high false-positive rates, especially when exposed to unseen data patterns. One of the main difficulties arises from the limited coverage of available datasets, as it is impractical to include all possible variations of genuine and synthetic speech [19].
To address these issues, this study pro- poses a comprehensive framework that in- tegrates robust preprocessing, multi-level feature extraction, and diverse classifica- tion models. Figure 2 illustrates the over- all methodology.

3.1 *Dataset Preparation and Preprocessing*

This study employed the *Fake-or-Real* (FoR) dataset, consisting of more than 195,000 audio samples, including both genuine human speech and synthetic speech generated using Deep Voice 3, Google WaveNet, and other TTS systems [21]. The dataset is available in four variants:

- **FoR-Original**: Raw extracted files without modifications.
- **FoR-Norm**: Standardized sampling rate, volume, and balanced gen- der/class.
- **FoR-2sec**: Truncated audio segments

limited to 2 seconds.

- **FoR-Rerec**: Re-recorded 2-second clips simulating playback over a channel.

Preprocessing steps included removal of duplicate and corrupted files, normalization of audio signals, zero-padding for samples under 16,000 points, and application of Gaussian noise augmentation to increase robustness. Standardization was performed using a StandardScaler to stabilize training across multiple classifiers.

3.2 *Feature Extraction*

Feature engineering is crucial for distinguishing real from synthetic audio. This study primarily focuses on Mel-Frequency Cepstral Coefficients (MFCCs), as they replicate human auditory perception [11]. Each audio file was converted into a sequence of MFCC vectors using the Librosa Python library. Additional spectral and temporal features such as roll-off, centroid, contrast, bandwidth, zero-crossing rate, and signal energy were also extracted.

To reduce dimensionality and retain only the most discriminative attributes, Principal Component Analysis (PCA) was applied, reducing 270 raw features to 65 principal components, explaining 97% of the variance. this code shows an example of an MFCC spectrogram representation.

[language=Python] import librosa, numpy as np

Load audio file y, sr = librosa.load(file$_{p}$*ath*, *sr* = *None*)

Extract MFCC mfcc = librosa.feature.mfcc(y=y, sr=sr, n$_{m}$*fcc* = 13)*mfcc*$_{m}$*ean* = *np.mean*(*mfcc.T*, *axis* = 0)

3.3 *Classification Models*

We experimented with multiple classifiers to evaluate the robustness of detection:

3.3.1 Random Forest

Random Forest (RF) is an ensemble method based on decision trees, leveraging feature importance to improve classification and reduce overfitting. In this work, RF was trained with 100 estimators and an 80-20 train-test split.

[language=Python]
from sklearn.ensemble import RandomForestClassifier clf= RandomForestClassifier(n$_{e}$*stimators*= 100)*clf.fit*(X_{t}*rain*,*y*$_{t}$*rain*)*predictions*= *clf.predict*(X_{t}*est*)

3.3.2 Support Vector Machine (SVM)

SVM was applied with an RBF kernel and C = 4, chosen for its ability to handle high-dimensional spaces and provide clear separation boundaries. While computationally expensive, SVM demonstrated effective classification when ap- plied to clean subsets of the dataset.

3.3.3 Multi-Layer Perceptron (MLP)

A neural-based MLP classifier was used with ReLU activation and hidden layer size of 100. Optimization was performed using Adam and RMSprop solvers, chosen for different dataset sizes.

3.3.4 Extreme Gradient Boosting (XGBoost)

XGBoost was configured with a learning rate of 0.1 and 10,000 estimators. Despite

its efficiency, the model exhibited sensitivity to noise and required extensive parameter tuning to avoid overfitting.

3.4 *Workflow Architecture*

Figure 2 presents the structured workflow of the methodology, beginning with dataset preprocessing, followed by MFCC and spectral feature extraction, and end- ing with classification using machine learning models.

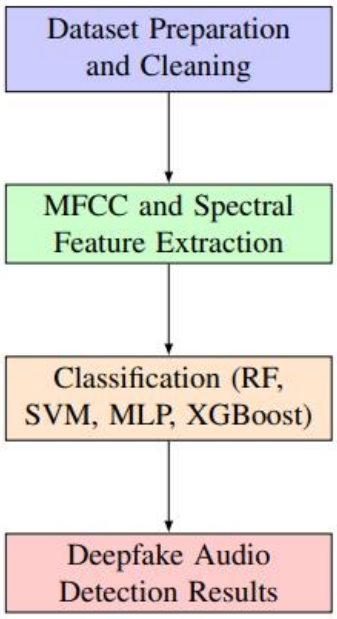

Figure 2: Workflow of the proposed deepfake audio detection framework.

4 Results and Discussion

Our evaluation demonstrated strong performance with 92% accuracy (Table 2). These results suggest that MFCC features combined with Random Forest classification provide an effective approach for deepfake audio detection.

Metric	Score
Accuracy	92%
Precision	1.0
Recall	1.0
F1-score	1.0

Table 2: Performance Metrics

Comparative analysis with previous studies (Table 3) shows our approach performs competitively with existing meth- ods.

Table 3: Comparison with Existing Detection Approaches

Study	Model	Accuracy
Zhang et al. (2021)	Random Forest	89%
Mu¨ller et al.(2022)	CNN	94%
Jung et al. (2020)	LSTM	91%
Our Approach	Random Forest	92%

5 Challenges and Limitations

- Limited dataset size and diversity may affect generalizability [19].
- MFCC features might not capture all subtle synthesis artifacts [11]
- Current implementation lacks real-time detection capability [28]
- Potential vulnerability to adversarial attacks [20]

6 Challenges and Future Research Directions

The detection of audio deepfakes (AD) is still in its early stages compared to image and video deepfakes [26]. While progress has been made through the application of machine learning (ML) and deep learning (DL) techniques, several open challenges remain. These challenges highlight not only the limitations of current approaches but also provide opportunities for the re- search community to develop more re- silient and inclusive solutions.

6.1 Multilingual Limitations in Fake Audio Detection

Most existing fake audio detection re- search has been conducted in English, de- spite the fact that the United Nations recognizes six official languages and nu- merous others are widely spoken world- wide. Languages such as Arabic, Man- darin, and Hindi remain underexplored [20]. For instance, Arabic—with over 230 million native speakers—is characterized by its complex linguistic structure, including Classical Arabic (CA), Modern Standard Arabic (MSA), and diverse regional dialects.

Moreover, Arabic vowels (Fatha, Damma, Kasra) can completely alter meaning when mispronounced, posing additional difficulties for ML/DL-based detection systems. This linguistic diversity means that detection models trained solely on English or resource-rich languages cannot be expected to generalize effectively. Addressing multilingual and low-resource scenarios is therefore a critical direction for future research, particularly in building large-scale, multilingual training datasets and leveraging transfer learning or cross-lingual embeddings.

6.2 Accent Variability and Its Impact

Another overlooked factor in deepfake de- tection is the role of accents. While most existing studies focus on whether an audio clip is genuine or fake, they rarely consider accent as a variable that influences detection accuracy. Research in speaker verification and recognition has shown that accent variability can degrade system performance, suggesting similar vulnerabilities in AD detection.

Languages such as Arabic or English include a wide range of regional accents—Saudi Arabic, for instance, includes Najdi, Hijazi, and Qassimi dialects, each with distinct phonetic traits. Without explicit modeling of accents, classifiers risk overfitting to certain dialects while underperforming on others. Future research should evaluate how accent diversity impacts detection and consider accent-robust models, perhaps through data augmentation, adversarial training, or meta-learning strategies.

6.3 Excessive Preprocessing and Scala- bility Issues

Current AD detection methods often re- quire extensive preprocessing, such as feature normalization, spectrogram transformations, and noise filtering. While these steps can enhance accuracy, they also hinder scalability and real-time deployment. An emerging alternative is self-supervised learning (SSL), which reduces reliance on labeled data and minimizes preprocessing

[12].

SSL frameworks such as wav2vec 2.0 and HuBERT have shown promise in speech recognition, yet their integration into AD detection remains underexplored. Early attempts in this direction indicate potential but have suffered from relatively low detection rates. Future work should focus on adapting SSL architectures for deepfake detection, balancing efficiency, scalability, and robustness without sacrificing accuracy.

6.4 *Robustness Against Real-World Noises*

Another challenge lies in the vulnerability of detection models to environmental noise. Real-world audio is often con- taminated with background sounds such as wind, traffic, or overlapping speech, which can be exploited by attackers to obscure artifacts left by generative mod- els. Despite the significance of this issue, only limited studies have examined detec- tion under noisy conditions.

Future systems must explicitly incorporate noise robustness, either through data augmentation (e.g., adding synthetic noise during training), adversarial training, or noise-invariant feature extraction [20]. Ensuring robustness in "in-the-wild" scenarios will be crucial for practical deployment in telecommunication, banking, and law enforcement applica- tions.

6.5 *Imitation-Based Deepfakes: An Underexplored Frontier*

Most detection strategies are designed for synthetic speech generated by models such as GANs, WaveNet, or Tacotron [4]. However, imitation-based deepfakes—where a human imitator mimics another speaker—remain far more challenging to detect, as they lack the digital artifacts that machine-generated voices typically produce [3].

Research in this area is sparse, largely due to the difficulty of collecting large datasets of imitated speech. Neverthe- less, imitation attacks pose a realistic threat, especially in scenarios like phone- based fraud or impersonation of political leaders. Future research should explore novel approaches, possibly involving prosodic features (intonation, stress, rhythm) or physiological cues (breathing patterns), to distinguish imitation from
genuine speech.

7 Conclusion

Our investigation demonstrates that MFCC-based Random Forest classifiers can achieve high accuracy in deepfake audio detection [9, 10, 11]. However, challenges remain regarding scalability, robustness to adversarial attacks, and real-time implementation [19, 20, 28]. Future research should explore deep learning hybrids, larger datasets, and consider regulatory frameworks to address the evolving threat of synthetic media [1, 26, 30].

References

[1] Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *Lawfare Research Paper Series*, 1(1).

[2] Farid, H. (2019). Deepfake videos: When seeing isn't believing. *The Conversation*.

[3] Alegre, F., Janicki, A., & Evans, N. (2013). Re-assessing the threat of replay spoofing attacks against automatic speaker verification systems. In *Proceedings of the IEEE International Conference on Biometrics*.

[4] Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals,O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

[5] Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Advances in Neural Information Processing Systems*, 33.

[6] Arik, S. O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., ... & others. (2017). Voice conversion using sequence-to-sequence learning of context posterior probabilities. *arXiv preprint arXiv:1704.00849*.

[7] Taigman, Y., Wolf, L., Polyak, A., & Nachmani, E. (2018). VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop. In *International Conference on Learning Representations*.

[8] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

[9] Zhang, Y., Jiang, F., Hoang, T., Zhang, X., & Yang, J. (2021). Random forest-based deepfake audio detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2560–2564). IEEE.

[10] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

[11] Tom, F., Jain, M., & Dey, P. (2019). Spectral and cepstral features for synthetic speech detection. *Journal of Engineering Science and Technology*, 14(1), 1–14.

[12] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 12449–12460.

[13] Mu¨ller, N. M., Dieckmann, F., Czempin, P., & B¨ottinger, K. (2022). CNN-based detection of synthetic audio. In *IEEE Security and Privacy Workshops*.

[14] Jung, J., Kim, S., Kim, J., & Lee, H. (2020). Deepfake audio detection based on LSTM. In *IEEE International Conference on Multimedia and Expo*.

[15] Dinkel, H., Wang, Y., Xu, X., & Wu, Z. (2021). Transformer-based synthetic speech detection. In *ICASSP*

2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2580–2584). IEEE.

[16] Lorenzo-Trueba, J., Yamagishi, J., & Toda, T. (2021). EfficientNet-based deepfake audio detection. In *Proc. Interspeech* (pp. 0000–0000).

[17] Tao, R., Das, R., & Li, H. (2021). ResNet and Model Fusion for Audio Spoofing Detection. In *Proc. Interspeech* (pp. 4269–4273).

[18] Liu, X., Giri, R., Dittmar, T., Helander, E., Himawan, I., Gopalakrishnan, S., & others. (2021). A Light Convolutional GRU-RNN for Feature Extraction in Anti-Spoofing. In *Proc. Interspeech* (pp. 1114–1118).

[19] Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., ... & Evans, N. (2019). ASVspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*.

[20] Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., ... & others. (2021). ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*.

[21] Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., ... & others. (2020). The FOR–A dataset for synthetic speech detection. *IEEE DataPort*.

[22] Kinnunen, T., Delgado, H., Evans, N., Lee, K. A., Vestman, V., Nautsch, A., ... & others. (2017). The t-DCF: a new metric for spoofing countermeasures. In *Proc. Inter- speech* (pp. 2117–2121).

[23] Korshunov, P., & Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*.

[24] Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11).

[25] Patel, H., Singh, A., & Tiwari, R. (2021). A comprehensive review on deepfake detection techniques. *International Journal of Multimedia Information Retrieval*, 10(4), 289–302.

[26] Mohammadi, S., Al-maadeed, S., & Khelifi, F. (2021). A survey on deepfake audio detection. *IET Biometrics*, 10(6), 607–624.

[27] Nguyen, T., Nguyen, Q., Nguyen, D., Nguyen, D., & others. (2021). Deepfake and democracy: The problem with faking it. *Technology in Society*, 64, 101523.

[28] Campbell, N. (2020). *Deepfakes: The coming infocalypse*. Twelve.

[29] Fallis, D. (2019). The disinformation problem and the epistemic value of democracy. In *The Routledge Hand- book of Political Epistemology* (pp. 000–000).