iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: XII Month of publication: December 2025

DOI: https://doi.org/10.22214/ijraset.2025.74521

A Survey on Anomaly Detection of Deepfake Audio Using GAN-Based Model

Vyshnavi M[1], Ananya Malali[2], H Shireesha[3], Lakshmi B A[4.]
[1]*Assistant Professor, Department of Computer Science and Engineering, Sapthagiri College of Engineering, Karnataka, India.*
[2, 3, 4]*BE Student, Department of Computer Science and Engineering, Sapthagiri College of Engineering, Karnataka, India*

Abstract: Deepfake audio has emerged as one of the most concerning challenges in digital media authenticity, enabled by advances in deep learning and generative modelling. It has both positive applications in assistive technologies and dangerous implications for misinformation, impersonation, and fraud. Traditional supervised classification approaches often fail to generalize against new synthesis techniques. Anomaly detection methods, particularly those leveraging Generative Adversarial Networks (GANs), have shown promise in identifying deepfake audio by modeling authentic speech distributions. This paper presents a comprehensive survey of anomaly detection techniques applied to deepfake audio, with focus on GAN-based frameworks such as GANomaly and f-AnoGAN, and comparison with CNN and Autoencoder-based methods.

Keywords: Deepfake audio, anomaly detection, GAN working with Mel-Spectrogram, audio forensics.

I. INTRODUCTION

Deepfake technology has gained rapid popularity with the ability to synthesize highly realistic human speech. While beneficial in applications such as virtual assistants, dubbing, and accessibility, it poses severe risks by enabling identity theft, voice-based fraud, and spread of misinformation. Detecting manipulated audio has thus become an urgent research challenge. Conventional supervised learning approaches often require labelled datasets, which are difficult to maintain as synthesis methods evolve. Anomaly detection, particularly GAN-based methods, addresses this limitation by learning normal speech characteristics and identifying deviations as anomalies. The misuse of synthetic audio can lead to identity theft, voice-based fraud, impersonation attacks, and the spread of misinformation, particularly in political or legal contexts. For example, attackers can generate fake audio clips that mimic public figures to spread false statements, manipulate financial systems using cloned voices for authorization, or conduct social engineering attacks in enterprise environments. Detecting manipulated audio has thus become an urgent and complex research challenge. Conventional supervised learning-based approaches, which involve training classifiers to distinguish between real and fake audio, rely heavily on labelled datasets.

II. LITERATURE SURVEY

A. GANomaly-Based Detection

GANomaly is a semi-supervised anomaly detection framework that integrates adversarial learning with an encoder-decoder-encoder architecture. It is trained solely on genuine speech data to learn the underlying distribution of real audio. The model computes an anomaly score using both reconstruction error and latent space discrepancy, flagging inputs with high deviation as potential deepfakes. Although GANomaly has proven effective in identifying fake audio, its performance deteriorates when applied to datasets with high variability, such as those containing multiple speakers, accents, or background noise. This sensitivity limits its robustness in diverse real-world environments.

B. f-AnoGAN and Improved Reconstruction

To overcome the inference inefficiencies of traditional GAN-based methods, f-AnoGAN introduces a pre-trained encoder that directly maps input audio into the latent space of the generator. This significantly speeds up the detection process and enhances reconstruction quality, making it suitable for practical applications. f-AnoGAN also uses feature-matching loss to improve the alignment between original and reconstructed data. However, its improved performance comes at the cost of increased computational requirements and added model complexity, necessitating careful training and resource allocation.

C. CNN-Based Classification Models

Convolutional Neural Networks (CNNs) are widely used in deepfake audio detection by transforming audio into spectrograms or Mel-frequency cepstral coefficients (MFCCs) and treating them as 2D images. CNNs are effective at capturing local patterns and artifacts introduced during audio synthesis. These models often achieve high accuracy when detecting deepfakes generated using known techniques. However, their supervised nature makes them vulnerable to generalization issues; their performance drops significantly on deepfakes created using novel or unseen synthesis methods, limiting their applicability in dynamic threat environments.

D. Autoencoder and VAE Approaches

Autoencoder-based models are commonly applied for anomaly detection in deepfake audio due to their ability to reconstruct input features and measure deviations. Traditional autoencoders minimize reconstruction loss, while Denoising Autoencoders improve robustness by learning to reconstruct clean signals from noisy inputs. Variational Autoencoders (VAEs) introduce probabilistic modeling in the latent space, allowing for better generalization across diverse data. Despite their advantages, autoencoders can struggle with subtle manipulations and are prone to overfitting, especially when trained on limited or homogeneous datasets.

E. Hybrid and Ensemble Techniques

To improve generalization and robustness, recent research has explored **hybrid approaches** that combine multiple model architectures.

Examples include fusing GANs with Transformer-based models to capture long-range dependencies in audio or using ensembles of CNNs, VAEs, and GANs to enhance detection performance. Some studies also integrate audio with visual information for multimodal detection, identifying inconsistencies in audio-visual synchronization. While these hybrid systems demonstrate improved accuracy and resilience against adversarial attacks, they also introduce greater computational overhead and model complexity, which may hinder real-time deployment.

F. Multimodal Detection Approaches

Multimodal approaches combine audio with other data types—such as video or text—to improve deepfake detection. For example, detecting inconsistencies between lip movements and audio (audio-visual mismatch) can help identify fake content. These methods enhance robustness against advanced attacks but require synchronized multimodal data and more complex processing pipelines.

G. Transformer-Based Models

Transformer-based architectures, originally developed for NLP, have recently been applied to deepfake audio detection due to their ability to model long-range dependencies in sequential data. Models like Audio Spectrogram Transformer (AST) and Wav2Vec leverage attention mechanisms to capture contextual information from audio signals. These models show promise in improving generalization and detecting subtle manipulations but require large datasets and significant computational resources for training.

H. Self-Supervised Learning Approaches

Self-supervised learning (SSL) methods aim to learn useful representations from unlabeled audio data through pretext tasks like contrastive learning or masked prediction.

Techniques such as SimCLR, BYOL, or HuBERT enable the model to understand the structure of real speech without deepfake labels. SSL can improve performance in low-resource settings and help models adapt to unseen attacks, but effective fine-tuning strategies are still an open research area.

I. Contrastive Learning Methods

Contrastive learning focuses on learning representations by pulling similar samples closer and pushing dissimilar ones apart in the embedding space. In deepfake audio detection, it helps distinguish real from fake speech by learning discriminative features without needing large labeled datasets.

Models like SupCon or MoCo have shown improved performance in generalization and robustness, though they require careful sampling strategies and large batch sizes.

J. Graph-Based Detection Techniques

Recent works have explored graph neural networks (GNNs) to model relationships between different audio segments or features. By constructing graphs from audio features, GNNs can capture structural inconsistencies or unnatural transitions common in synthetic speech. Although promising, graph-based models are still in early stages for audio applications and face challenges in scalability and interpretability.

K. Real-Time and Lightweight Detection Models

For deployment in real-world systems, particularly on mobile or embedded devices, lightweight models like MobileNet, TinyML architectures, and pruned CNNs are being developed. These models prioritize low latency and low power consumption while maintaining acceptable detection accuracy. While suitable for real-time applications, they often trade off performance, especially when facing high-quality or adaptive deepfake attacks.

L. Dataset-Specific Fine-Tuning

Many detection models rely on dataset-specific fine-tuning to optimize performance on particular benchmark datasets such as ASVspoof, FakeAVCeleb, or WaveFake. While fine-tuning helps improve model accuracy within a given dataset, it often leads to overfitting, reducing the model’s ability to generalize to unseen data or deepfake generation techniques. This highlights the need for cross-dataset evaluation and robust training practices.

M. Adversarial Robustness

Deepfake detectors are increasingly being evaluated against **adversarial attacks**, where small, imperceptible perturbations are added to audio to fool the detection system. Some models integrate **adversarial training** or **defensive distillation** to improve robustness. However, balancing accuracy and resistance to adversarial inputs remains a challenge, especially as attack methods become more sophisticated and transferable across models.

N. Explainability and Interpretability

As deepfake detection systems are used in sensitive applications such as law enforcement or media verification, **explainability** has become crucial. Methods such as **Layer-wise Relevance Propagation (LRP)** or **SHAP** are being adapted to audio to help interpret model decisions. Despite this progress, most deep learning models remain black boxes, making it difficult to understand or justify predictions especially in high-stakes environments.

O. Cross-Lingual and Multilingual Detection

Deepfake audio detection systems often struggle with cross-lingual generalization, as most are trained on English or a single language. With the rise of multilingual deepfake tools, models need to handle speech across diverse languages, accents, and dialects. Some recent research incorporates multilingual datasets or language-agnostic features (e.g., prosody, rhythm) to improve cross-lingual performance, but building truly language-robust models remains a significant challenge.

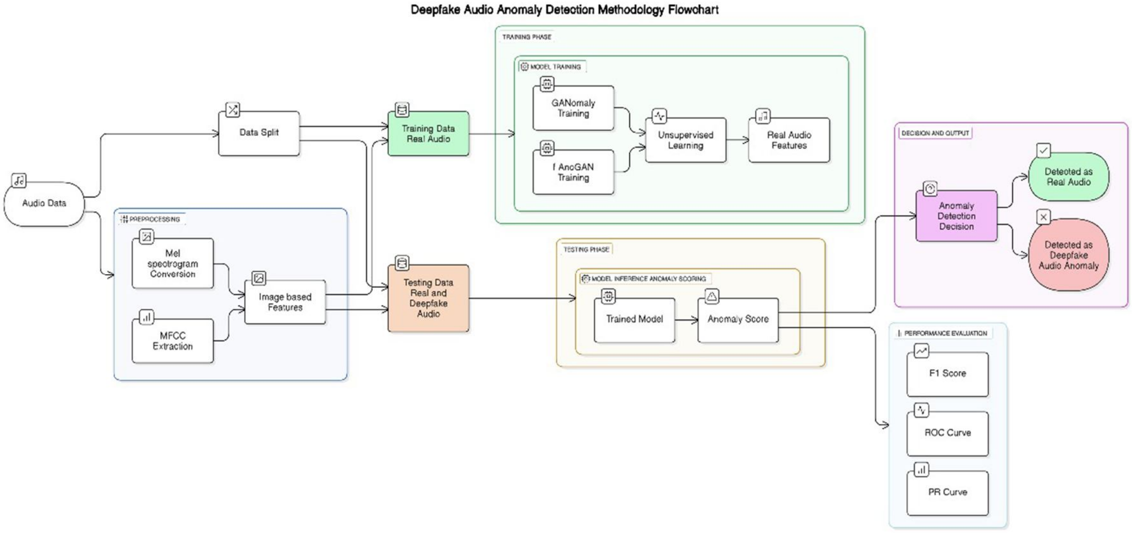
III. METHODOLOGY

The general architecture of GAN-based anomaly detection includes preprocessing, feature extraction, training of generator-discriminator networks, and computation of anomaly scores. Feature extraction methods include MFCCs, log-mel spectrograms, and raw waveform embeddings. The generator attempts to reconstruct input features, while the discriminator distinguishes real from generated outputs. Anomaly scores are derived from reconstruction error or feature matching losses.

The core of the architecture involves training a Generative Adversarial Network (GAN) composed of a generator and a discriminator. The generator, often designed as an encoder-decoder network, learns to reconstruct input audio features by modeling the distribution of authentic speech, while the discriminator’s role is to distinguish real from generated features, providing feedback that refines the generator’s outputs. Training employs adversarial learning, where the generator tries to fool the discriminator by producing realistic reconstructions, and the discriminator aims to correctly classify inputs, leading to a robust representation of genuine audio data. Once trained, anomaly detection is performed by calculating anomaly scores derived mainly from the reconstruction error—measuring the difference between input features and their reconstructions—and sometimes latent space discrepancies or feature matching losses based on discriminator activations. Inputs that deviate significantly from the learned authentic distribution yield higher anomaly scores, signaling potential deepfake audio.

Postprocessing aggregates these scores across segments and applies thresholding for classification. Additional considerations include training stability techniques, such as Wasserstein loss or spectral normalization, to ensure convergence, and the importance of diverse training data to enhance generalization. While GAN-based methods can be computationally intensive during training, optimizations like encoder pretraining help enable faster inference.

A. Flow Diagram



This flowchart illustrates a Deepfake Audio Anomaly Detection Methodology. It outlines the full pipeline for detecting fake audio using anomaly detection techniques.

B. Pre Processing

Preprocessing is a critical first step in any deepfake audio detection pipeline, as it prepares raw audio data for consistent and effective feature extraction and model training. The raw audio signals often vary in length, sampling rate, noise levels, and loudness, making normalization essential. Typically, audio is resampled to a fixed sampling rate (e.g., 16 kHz or 22.05 kHz) to ensure uniformity across the dataset. Noise reduction techniques may also be applied to filter out background noise or irrelevant artifacts, especially in real-world recordings, thereby enhancing the quality and clarity of the signal. Silence trimming is often used to remove non-informative segments at the beginning and end of recordings. Next, the audio is segmented into fixed-size overlapping or non-overlapping frames (e.g., 1–2 seconds) to create uniform input lengths suitable for batch processing and model input. Normalization techniques, such as z-score normalization or min-max scaling, are applied to maintain consistent amplitude and energy across samples, preventing model bias due to volume variations. Additionally, data augmentation techniques—such as pitch shifting, time-stretching, background noise addition, or reverberation—are sometimes used to increase the diversity of the training set, which helps prevent overfitting and improves model robustness to real-world distortions. Overall, effective preprocessing ensures that the input data is clean, standardized, and suitable for downstream feature extraction and anomaly detection tasks, forming the foundation for reliable deepfake audio detection using GAN-based architectures.

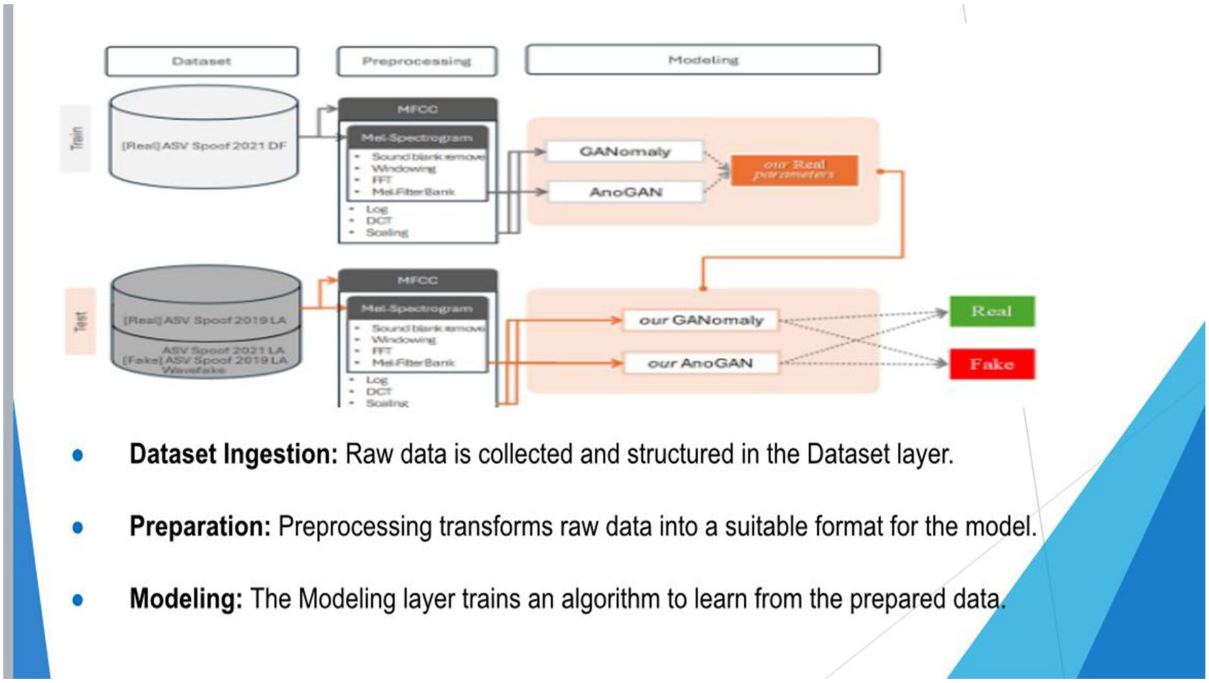
C. DATA SPLIT

To effectively train, validate, and evaluate GAN-based deepfake audio detection models, the dataset is typically divided into three subsets: training, validation, and testing. The training set consists solely of genuine (real) audio samples when using anomaly detection frameworks, as the goal is to model the normal distribution of authentic speech. This enables the generator to learn accurate reconstructions of real audio and helps the discriminator distinguish real from generated features. The validation set is used during training to fine-tune hyperparameters, monitor model performance, and prevent overfitting. It may include a mix of genuine and synthetic audio, especially when early evaluation of detection ability is required. Finally, the test set contains both genuine and deepfake audio samples, often generated from multiple synthesis methods (e.g., voice conversion, TTS models like Tacotron or WaveNet). This allows for robust evaluation of the model’s generalization capability to unseen or diverse attack types. Common splits follow an 80:10:10 or 70:15:15 ratio for training, validation, and testing, respectively.

In cross-dataset evaluation, the training and testing sets may be drawn from entirely different datasets to simulate real-world deployment scenarios. Stratified sampling or speaker-independent splitting is often used to ensure balanced representation and prevent data leakage.

D. Training Phase

The training phase of GAN-based anomaly detection models is critical for learning the underlying distribution of genuine speech and enabling the detection of deviations indicative of deepfake audio. During this phase, only real (authentic) audio samples are typically used, especially in unsupervised or semi-supervised settings. The training process involves two main neural networks: the generator (G) and the discriminator (D), which are trained simultaneously in an adversarial fashion. The generator is usually implemented as an encoder-decoder architecture that attempts to reconstruct input features (e.g., MFCCs or log-mel spectrograms) from a latent representation. The discriminator, on the other hand, learns to differentiate between real features and those reconstructed by the generator. As training progresses, the generator improves its ability to produce realistic reconstructions of genuine speech, while the discriminator becomes increasingly capable of identifying imperfections or discrepancies.



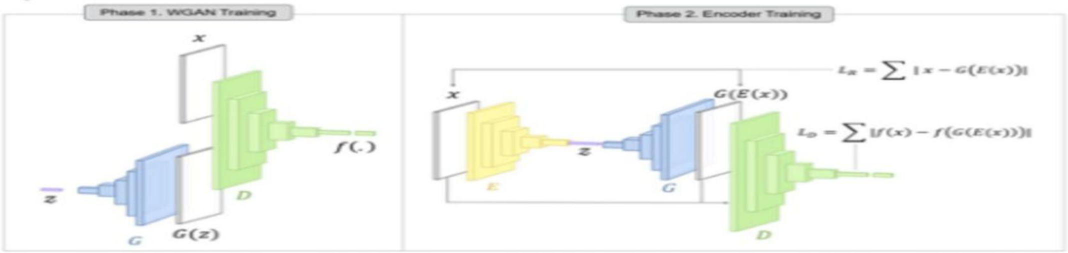
- **Dataset Ingestion:** Raw data is collected and structured in the Dataset layer.
- **Preparation:** Preprocessing transforms raw data into a suitable format for the model.
- **Modeling:** The Modeling layer trains an algorithm to learn from the prepared data.

E. Framework

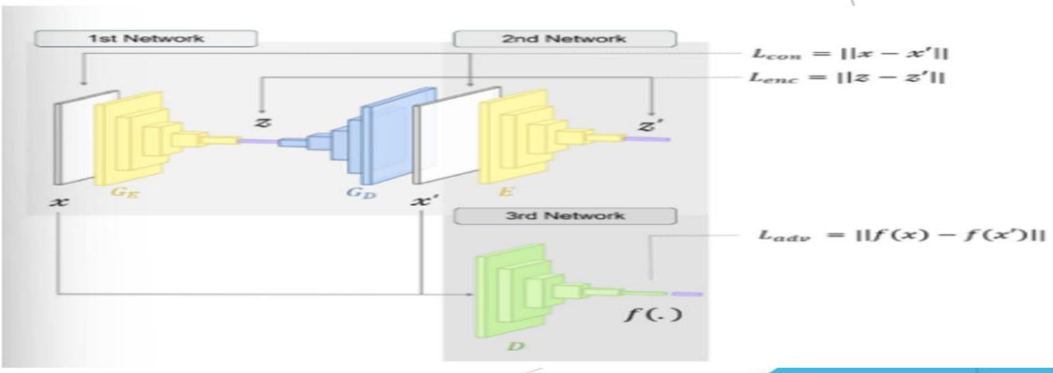
The proposed framework for GAN-based anomaly detection in deepfake audio consists of a systematic pipeline that includes preprocessing, feature extraction, GAN-based training, anomaly scoring, and final classification. It is designed to learn the distribution of genuine speech data and flag any input that deviates from this distribution as potentially fake or manipulated. The process begins with preprocessing, where raw audio samples undergo normalization, noise reduction, and segmentation into fixed-length frames. These steps ensure consistency across samples and prepare them for feature extraction. In the feature extraction stage, audio signals are converted into informative representations such as Mel-Frequency Cepstral Coefficients (MFCCs), log-mel spectrograms, or raw waveform embeddings. These features capture the key spectral and temporal properties of speech, making them suitable inputs for the GAN.

F. Models Used

In this framework for deepfake audio detection, the primary model employed is a Generative Adversarial Network (GAN) specifically tailored for anomaly detection. Among the variants of GANs, models such as GANomaly, f-AnoGAN, and Adversarial Autoencoders (AAE) are widely used due to their ability to learn the distribution of genuine audio features without requiring labeled fake data during training. The core structure of these models involves two components: a generator and a discriminator.



f-AnoGAN model framework.



GANomaly model framework.

G. Datasets

One of the most widely used datasets is ASVspoof (Automatic Speaker Verification Spoofing and Countermeasures Challenge), with multiple versions such as ASVspoof 2015, 2019, and 2021. These datasets provide a standardized benchmark for evaluating anti-spoofing systems. They contain both bona fide (genuine) speech and spoofed audio generated using text-to-speech (TTS), voice conversion (VC), and replay attacks. ASVspoof 2019, for instance, includes logical access (LA) and physical access (PA) subsets, focusing on synthetic and replay attacks respectively. Another important dataset is **WaveFake**, which includes audio samples synthesized using multiple state-of-the-art speech synthesis models such as Tacotron, WaveNet, and ParallelWaveGAN. This dataset is particularly useful for evaluating generalization, as it contains audio from diverse speakers and synthesis systems.For GAN-based anomaly detection systems, training is often conducted only on real (genuine) samples, so the dataset must offer a sufficient quantity and diversity of bona fide speech. Evaluation, however, requires deepfake samples generated from various attack types to test generalization and robustness. Cross-dataset testing is also commonly used to simulate real-world scenarios, where the model must detect fake audio generated by unseen synthesis techniques.

IV. DISCUSSION

GAN-based anomaly detection models like GANomaly and f-AnoGAN are effective in detecting deepfake audio without needing labeled fake data. They learn the distribution of genuine speech and identify outliers, making them more adaptable to unseen attack methods. These models often struggle with training instability, poor reconstruction quality on variable audio, and overfitting to specific datasets. Inference speed can also be an issue, especially with complex architectures like f-AnoGAN. While performance is strong on known data, cross-dataset generalization remains limited. Models may fail to detect fakes from unfamiliar sources or synthesis methods. Improvements may come from lighter models, adversarial robustness, cross-lingual support, and interpretable outputs. Combining GANs with transformers or self-supervised learning could enhance robustness and accuracy. Future research should explore lightweight architectures for faster inference, adversarial training for robustness against sophisticated attacks, and multimodal approaches combining audio with other data types. Incorporating self-supervised learning and transformer-based models may further boost performance and adaptability. Emphasis on explainability and interpretability of anomaly scores will also be critical for real-world deployment.

V. CONCLUSION

Deepfake audio detection remains a critical challenge as synthetic speech technologies advance rapidly. GAN-based anomaly detection frameworks offer a promising solution by learning the distribution of genuine speech and identifying manipulated audio as anomalies, without the need for extensive labelled fake data. Models like Generator and Discriminator demonstrate strong potential in adapting to unseen attacks and improving detection robustness. However, challenges such as training instability, reconstruction quality degradation, and limited cross-dataset generalization persist. Addressing these issues through improved architectures, diverse datasets, and hybrid learning strategies will be essential for creating scalable and reliable detection systems. Future work should focus on enhancing model interpretability, reducing computational overhead, and ensuring robustness against sophisticated adversarial attacks. By advancing GAN-based anomaly detection, researchers can contribute significantly to securing digital audio media against the growing threat of deepfake manipulation.