RESEARCH PAPER

An Analytical Review of Multi-layered Security Frameworks for Audio Deepfake Detection: Combining ML Classifiers with Linguistic Verification

Badal Singh[1*], Priyank Sirohi[2]

[1*] M.Tech Scholar, Department of Computer Science, Sir Chhotu Ram Institute of Engineering and Technology, Meerut, U.P.(India) badalsinghbabura12@gmail.com

[2]Assistant Professor, Sir Chhotu Ram Institute of Engineering and Technology Chaudhary Charan Singh University Meerut, U.P. (India) priyanksirohi01@gmail.com

Abstract

The rapid evolution of audio deepfake generation technologies—driven by advanced deep learning architectures such as Generative Adversarial Networks (GANs), autoregressive neural vocoders, and diffusion-based speech synthesizers—has significantly expanded the threat landscape surrounding voice-based authentication systems. Although machine-learning-based audio deepfake detectors demonstrate promising accuracy under benign conditions, recent studies show that such systems remain highly vulnerable to adversarial attacks capable of bypassing classifiers with minimal computational effort.

In particular, single-layer detectors relying exclusively on acoustic representations—such as histograms, spectrograms, or cepstral features—fail to capture semantic, temporal, and linguistic consistency. This limitation enables attackers to generate synthetic or adversarial audio capable of bypassing state-of-the-art detectors with success rates exceeding 95% under gray-box threat models [1].

This review presents an analytical examination of multi-layered security frameworks designed to enhance the robustness of audio deepfake detection systems. Motivated by the vulnerabilities of single-layer detectors, the paper systematically evaluates architectures that integrate machine-learning-based acoustic classifiers with linguistic verification mechanisms, particularly Speech-to-Text (ASR) systems. In the proposed layered paradigm, the first layer evaluates acoustic authenticity, while the second layer verifies linguistic coherence, semantic validity, and contextual correctness.

Through comparative analysis of adversarial threat models, feature-space vulnerabilities, and domain-mapping inconsistencies, the review demonstrates that combining acoustic and linguistic validation substantially reduces adversarial success rates—from above 90% to near zero for noise-based and reversed-audio attacks. The framework is further formalized using mathematical modeling of domain transformations, classifier decision boundaries, and semantic constraints.

The findings indicate that semantic-aware, multi-layered detection architectures are essential for next-generation secure audio authentication systems. The paper concludes by outlining open research challenges and future directions, including multimodal biometric integration, adversarially robust ASR models, and adaptive defense strategies.

Keywords: Audio deepfake detection, adversarial attacks, multi-layered security frameworks, speech-to-text verification, machine learning security, GAN-based attacks, semantic consistency, biometric authentication

How to cite this article: Singh B, Sirohi P. An Analytical Review of Multi-layered Security Frameworks for Audio Deepfake Detection: Combining ML Classifiers with Linguistic Verification. Int J Drug Deliv Technol. 2026;16(52s): 118-127. DOI: 10.25258/ijddt.16.52s.12.

Introduction

The emergence of audio deepfake technologies represents a fundamental shift in the security landscape of voice-based authentication, digital communications, and human–computer interaction. Audio deepfakes—synthetically generated or manipulated speech signals produced using artificial intelligence—have become increasingly realistic due to advances in speech synthesis frameworks such as WaveNet, Tacotron, MelGAN, and GAN-based voice conversion systems [2], [3].

These models replicate subtle acoustic attributes, including prosody, timbre, pitch dynamics, and speaker-specific characteristics, blurring the distinction between genuine human speech and AI-generated audio. Consequently, new attack vectors have emerged, enabling impersonation attacks, financial fraud, social engineering, and unauthorized access to voice-driven systems [1], [4].

Voice biometrics, once regarded as a reliable authentication modality, now face serious threats due to the accessibility of deepfake generation tools. Attackers can harvest voice samples from public platforms and generate synthetic speech capable of deceiving both human listeners and automated systems. A notable real-world incident involved the use of deepfake audio to authorize a fraudulent $35 million bank transfer, highlighting the severity of the threat [1].

To mitigate these risks, researchers have proposed machine-learning-based audio deepfake detectors using acoustic features such as spectrograms, MFCCs, LFCCs, and histograms [5],[9]. While these approaches

*Author for Correspondence badalsinghbabura12@gmail.com

achieve high accuracy under controlled conditions, they exhibit severe robustness degradation under adversarial manipulation [1], [10], [11].

A representative example is Deep4SNet, a CNN-based histogram detector achieving 98.5% accuracy under benign conditions but collapsing to below 1% accuracy under gray-box adversarial attacks [1]. Such failures expose a fundamental limitation of single-layer detectors: they evaluate acoustic similarity but ignore linguistic meaning and semantic coherence.

This review addresses this limitation by analyzing multi-layered security frameworks that combine acoustic classification with linguistic verification using ASR systems. By enforcing semantic constraints in addition to acoustic plausibility, layered architectures significantly reduce adversarial attack success.

2. Related Work

Research on audio deepfake detection can be broadly classified into two domains:

1. Audio deepfake detection techniques
2. Adversarial attacks on audio classification systems
This section synthesizes both domains and highlights the limitations motivating multi-layered defenses.

2.1 Audio Deepfake Detection Methods

Early approaches relied on handcrafted acoustic features such as MFCCs, entropy measures, and bispectral statistics. Logistic regression-based detectors achieved high accuracy but showed poor generalization [5]. Quadratic SVM-based approaches demonstrated similar limitations [6].

With the rise of deep learning, CNN-based spectrogram classifiers [8], [9], hybrid CNN–RNN architectures [10], and residual networks [7] improved performance but remained vulnerable to noise, dataset shifts, and adversarial perturbations.

Histogram-based detectors such as Deep4SNet [1] offered computational efficiency but suffered catastrophic failure under adversarial attacks due to loss of temporal and semantic information.

These gaps motivate multi-layered security architectures.

Key observation:
Most existing detectors analyze only acoustic features and ignore linguistic or semantic structures.

2.2 Adversarial Attacks on Audio Classifiers

Adversarial attacks in audio domains include black-box, white-box, and gray-box threat models [12], [13].

Black-box attacks
The attacker has no information about the model's architecture or parameters. Query-based or surrogate-model-based attacks are used to generate adversarial audio. These attacks demonstrate moderate success but require a large number of queries.

White-box attacks
The attacker has complete access to the target model's architecture, parameters, and gradients. Classical optimization-based attacks such as FGSM and Carlini–Wagner (C&W) have been adapted to the audio domain to craft adversarial perturbations on speech waveforms or MFCC features. White-box attacks represent the strongest adversarial setting but may not always be realistic in practice.

Gray-box attacks
The adversary has partial knowledge of the model—typically the architecture and training dataset but not the defense mechanism. This setting is highly relevant for deployed audio authentication systems, where attackers can often inspect classifier structures or datasets but may not know internal defense layers.
The base paper [1] demonstrates that under gray-box conditions, adversarial histograms generated via GANs can reduce Deep4SNet's accuracy from 98.5% to 0.08%, revealing severe vulnerability.

2.3 Limitations of Existing Literature

Key gaps include:
- Absence of semantic verification
- Non-injective feature mappings
- Poor cross-dataset generalization
- Lack of standardized layered defense models

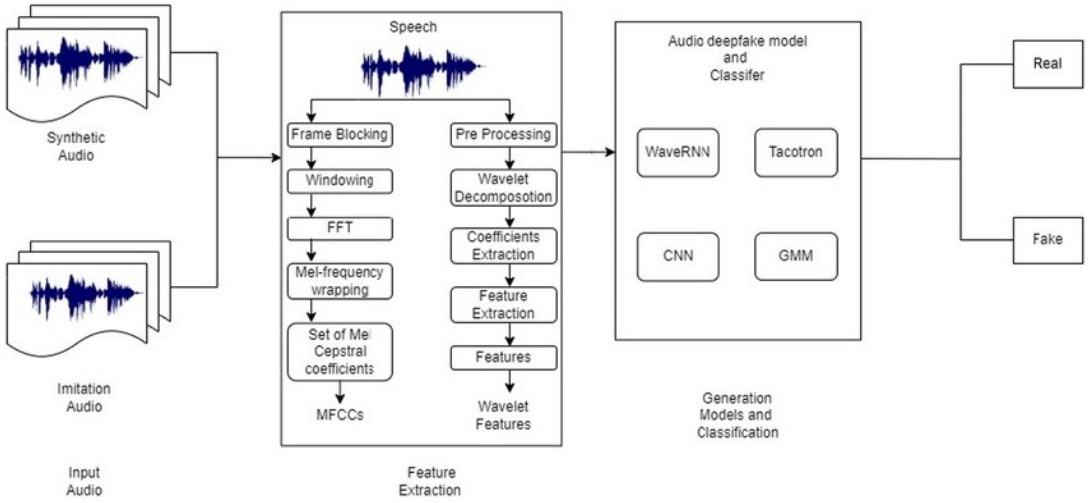

Figure 1. Conceptual Overview of Existing Audio Deepfake Detectors

3. Threat Models in Audio Deepfake Attacks

This section formalizes black-box, white-box, and gray-box threat models, highlighting their implications for detector robustness. Gray-box attacks represent the most realistic and damaging scenario, as attackers often have access to model architecture and datasets but not internal defenses.

3.1. Black-box Threat Model

In a black-box scenario, the attacker has no internal knowledge of the classifier. The adversary does not know:

- The model architecture,
- The weights or parameters,
- Feature extraction methods, or Training data distribution.

However, the adversary may still be able to perform:

- Input–output queries,
- Transferability attacks (crafting adversarial examples using surrogate models),
- Randomized attacks, or Score-based attacks (if confidence values are visible).

Key Characteristics

- Requires no specialized access.
- Typically uses GANs, surrogate models, or query-based optimization.
- Attack success rates vary but can still exceed 60–80% depending on detector architecture.
- Implications for audio detection

Black-box adversaries often exploit the tendency of detectors to overfit specific datasets. For example, detectors trained exclusively on ASVspoof samples often fail in cross-dataset settings, allowing black-box attacks to succeed by supplying unseen but acoustically plausible fake audio [7].

3.2. White-box Threat Model

In a white-box scenario, the adversary has full access to the model, including:

- Complete architecture details,
- Model weights and gradients,
- Training data,
- Preprocessing methods,
- Loss function, and Decision boundaries.

This is the strongest possible adversarial setting and enables highly optimized attacks such as:

- FGSM (Fast Gradient Sign Method) [14]
- Iterative Gradient methods
- Carlini & Wagner (C&W) optimization attacks [13]
- White-box GAN-based adversarial generation

Mathematically, the attacker seeks:

$$xadv = x + \delta suchthatf(xadv) = yt$$

and

$$\| \delta \| < \epsilon$$

where

- x is the real or fake audio,
- δ is perturbation bounded by ϵ,
- yt is the target label (“Real”),
- f(·) is the classifier.

Implication

White-box attackers can systematically drive detector accuracy to near-zero by exploiting precise gradient information, making semantic-layer defenses especially important, since linguistic consistency cannot be optimized purely via gradient signals.

3.3. Gray-box Threat Model

The gray-box setting—used in the base paper—is the most realistic and dangerous for real-world audio security systems.

In this scenario, the attacker knows:
- The model architecture,
- The feature representation (e.g., histogram, spectrogram),
- Possibly parts of the training dataset.

But the attacker does not know:
- Internal defense mechanisms (e.g., speech-to-text validation),
- Any additional post-classification checks,
- Dynamic runtime security controls.

Why gray-box attacks are the most relevant
Attackers today can access published models through research papers.
Many speech datasets are publicly available.
Cloud APIs expose classifier behavior through repeated queries.
Architecture transparency (e.g., open-source detectors) is common.
Base paper finding (Extended Analysis)
GAN-based histograms generated under gray-box conditions reduced Deep4SNet accuracy from 98.5% → 0.08% [1], a catastrophic collapse indicating that:
Acoustic-only detectors, even when top-performing, cannot withstand adversarial pressure in realistic threat environments.

3.4. Attack Goals in Modern Deepfake Scenarios
Across threat models, adversarial goals can include:

1. Authentication Bypass
Impersonating a user to unlock sensitive systems (e.g., banking, smart locks).
GAN-driven voice cloning greatly simplifies this objective.

2. Semantic Manipulation
Generating audio that says something different from what the transcription system perceives—exploited in attacks on speech-to-text systems [15].

3. Misleading Classification Systems
Producing unintelligible noise or reversed audio that is acoustically similar to legitimate audio in feature space.

4. Stealth and Persistence
Designing perturbations that are human-imperceptible but classifier-effective.

4. ML-Based Audio Deepfake Detectors and Vulnerabilities
This section analytically examines traditional ML-based detectors, deep learning architectures, and histogram-based systems. Structural weaknesses—such as loss of temporal information, representation ambiguity, dataset overfitting, and gradient vulnerability—are identified as core reasons for detector failure.

4.1. Traditional ML-Based Detectors
Early detectors relied on manually engineered features extracted directly from the raw audio waveform. Examples include:

▪ **Logistic Regression on Time-Domain Features**
Rodriguez-Ortega et al. [5] used waveform entropy and amplitude statistics to distinguish imitation-based fake audio. While achieving ~98% accuracy, the model required manual preprocessing and performed poorly on unfamiliar data distributions.

▪ **SVM-Based Detectors**
Quadratic SVM classifiers employing MFCC and bispectral features showed strong detection capability (~97–98%) under controlled conditions [6]. However, their sensitivity to noise and limited scalability rendered them inadequate for adversarially manipulated or real-world audio.

Vulnerability Analysis:
Traditional detectors depend on low-level statistical features that adversarial examples can easily exploit by manipulating frequency content, amplitude envelopes, or cepstral signatures with small but precise perturbations.

4.2. Deep Learning–Based Detectors
Modern deepfake detection systems rely on deep neural networks that extract hierarchical features from audio representations.

▪ **CNN-Based Spectrogram Classifiers**
Spectrograms convert audio signals into 2D images, enabling CNNs to classify speech using spatial filters. Models in [8], [9] demonstrated accuracy near 90% but exhibited limited generalization across datasets.

▪ **Residual and Multi-Scale Models**
Res2Net architectures [7] and deep residual convolutional networks capture multi-scale spectral patterns. However, despite improved generalization, they still degrade significantly with noise and unmatched acoustic conditions.

▪ **CNN + RNN Hybrid Models**
Hybrid GRU–CNN frameworks [10] attempt to capture both spectral and temporal dependencies. Although accurate on benchmark datasets, they remain vulnerable to adversarial examples that disrupt either temporal consistency or spectral smoothness.

4.3. Histogram-Based Detection Models (Deep4SNet)
Deep4SNet [1] introduced a lightweight detection scheme by converting raw audio into histogram images representing frequency distributions. The model uses a

CNN with three convolutional layers to classify histogram images into real or fake categories and reported 98.5% accuracy on H-Voices datasets.

Why histograms appeared favorable
- Computationally efficient
- Dimensionally compact
- Easy to store and process
- Compatible with image-based CNN architectures

However, this transformation introduces several structural weaknesses.

4.4. Why Current ML-Based Detectors Fail Against Real Adversaries

A unified conclusion emerges:
ML-based deepfake detectors fail because they evaluate only acoustic features and ignore linguistic and semantic content.
Attackers target precisely this gap.
An adversarial sample need only preserve enough spectral structure to mislead the classifier—not enough linguistic structure to produce meaningful, coherent speech.
Therefore, acoustic-only detection is fundamentally insufficient for real-world security systems.

5. Limitations of Single-Layer Audio Deepfake Detectors

Single-layer audio deepfake detectors—whether based on handcrafted acoustic features, spectrogram representations, CNN architectures, or histogram-based statistics—exhibit inherent structural limitations that make them unreliable under adversarial conditions. These weaknesses originate from non-injective audio-to-feature mappings, the absence of linguistic validation, inductive biases of visual feature learners, limited generalization, and susceptibility to gradient-based attacks. Consequently, high detection accuracy on benchmark datasets does not translate into robustness in real-world or security-critical deployments.

5.1. Non-Injective Audio-to-Feature Mappings

Most deepfake detection pipelines transform raw audio signals a ∈ A into a lower-dimensional feature space F using spectrograms, mel-spectrograms, MFCCs, LFCCs, or histogram representations:

$$g: A \rightarrow F$$

This transformation is fundamentally non-injective:

$$\exists a1 \ = a2 \in A: g(a1) = g(a2)$$

As a result, distinct audio signals may collapse into identical feature representations, causing irreversible loss of temporal structure, semantic ordering, and linguistic intent. In histogram-based detectors, temporal information is entirely discarded, allowing reversed speech, adversarial noise, spliced audio, or GAN-generated artifacts to map identically to authentic speech. This represents a structural limitation of the detection paradigm rather than a consequence of insufficient training data.

5.2. Absence of Linguistic and Semantic Validation

Single-layer detectors assess whether acoustic patterns resemble human speech but do not evaluate linguistic meaning or semantic coherence. Consequently, audio samples that are unintelligible to humans—such as reversed speech, phoneme-spliced segments, or adversarially optimized noise—are frequently classified as real, provided their spectral characteristics remain plausible. This allows attackers to bypass detection by preserving surface-level acoustic statistics while destroying linguistic validity, highlighting a critical blind spot in acoustic-only detection systems.

5.3. Over-Reliance on CNN-Based Visual Features

When spectrograms or histograms are treated as images, CNNs are commonly employed for classification. However, CNNs are designed to capture spatial correlations rather than long-range temporal dependencies inherent to speech. Moreover, they are highly vulnerable to small, structured perturbations in the input space and generalize poorly outside their training distribution. As a result, visually plausible feature representations—despite corresponding to invalid or synthetic audio—can be confidently misclassified as real, revealing a fundamental mismatch between visual feature learning and speech semantics.

5.4. Limited Generalization and Dataset Bias

Single-layer detectors are often trained on narrow, dataset-specific distributions, leading to degraded performance when encountering unseen speakers, languages, synthesis methods, recording environments, or real-world noise. Apparent robustness on benchmark datasets therefore reflects overfitting to dataset artifacts rather than genuine generalization. Even minor domain shifts are sufficient to induce classification failures, undermining real-world reliability.

5.5. Gradient-Based Adversarial Vulnerability

Most modern detectors are differentiable end-to-end, enabling gradient-based adversarial attacks:

$$xadv = x + \epsilon \cdot sign\big(\nabla xL(f(x), y)\big)$$

Such attacks can produce perceptually natural audio that reliably manipulates detector outputs under strict perturbation constraints. Additionally, GAN-based attacks bypass detectors entirely without requiring gradient access. Since single-layer decision boundaries remain smooth and differentiable, these systems lack effective mechanisms for resisting adversarial optimization.

5.6. Mathematical Summary

Let $g: A \rightarrow F$ denote the feature extraction function and $f: F \rightarrow 0,1$ the classifier. Single-layer detectors implicitly assume:

$$f(g(a)) = Real \Rightarrow aisauthentic$$

However, due to the non-injective nature of g:

$$g(a1) = g(a2) \nRightarrow a1 = a2$$

Therefore

$$f(g(a)) = Real \nRightarrow aisauthentic$$

This logical inconsistency creates an unavoidable attack surface that cannot be eliminated without incorporating additional layers of semantic, linguistic, or contextual validation.

6. Multi-Layered Security Frameworks for Audio Deepfake Detection

The consistent failure of single-layer audio deepfake detectors motivates the adoption of multi-layered security architectures that integrate complementary validation mechanisms. While acoustic-feature-based detectors rely solely on statistical, spectral, or visual patterns, multi-layered frameworks incorporate acoustic, linguistic, semantic, and contextual checks, thereby substantially reducing adversarial attack surfaces.

This section analytically reviews multi-layered detection strategies, with emphasis on architectures that combine ML-based acoustic classifiers with a secondary Speech-to-Text (ASR) validation layer, as proposed in the base paper [1]. The discussion highlights the theoretical motivation, architectural design, and robustness properties of layered defenses, demonstrating their necessity for secure, real-world audio authentication systems.

6.1. Motivation for Layered Defense Architectures

Single-layer detectors fail because they evaluate how speech sounds, not what it conveys. This gap allows attackers to generate audio that is acoustically plausible but semantically invalid, perceptually random yet feature-consistent, GAN-synthesized without linguistic structure, or temporally manipulated through reversal or splicing.

Multi-layered systems introduce a semantic barrier by combining acoustic authenticity with linguistic coherence verification:

$$F(a) = f2(f1(a))$$

Where f1(a) evaluates acoustic authenticity and f2 validates linguistic or semantic correctness. An adversary must now satisfy two independent constraints, significantly increasing attack complexity.

6.2. Architecture of Multi-Layered Detection Systems

A typical multi-layer framework consists of two sequential components:

Layer 1 — ML-Based Acoustic Detector:
Raw audio is processed through feature extraction (e.g., spectrograms or histograms) and classified as Real or Fake. This layer filters trivial or low-effort attacks but remains vulnerable to adversarial manipulation.

Layer 2 — ASR-Based Linguistic Verification:
Only audio classified as Real proceeds to the ASR module, which transcribes the signal:

$$T = ASR(a)$$

The resulting text is evaluated for transcription validity, syntactic structure, and semantic coherence. Audio yielding invalid, incoherent, or empty transcriptions is relabeled as Fake. This layer introduces constraints that are difficult for adversarial generation to satisfy.

6.3. Defense Under Diverse Adversarial Conditions

Noise-Based Attacks:
While noise can be optimized to fool acoustic detectors, ASR fails to produce coherent text, resulting in rejection. As shown in [1], noise bypassed Deep4SNet in over 90% of cases, but failed entirely when ASR validation was applied.

Reversed Audio:
Reversed speech preserves spectral distributions but destroys phonetic flow. ASR transcription becomes incoherent, enabling straightforward detection.

GAN-Based Attacks:
GAN-generated histograms and adversarial spectrograms exploit CNN visual biases in single-layer systems. However, ASR introduces a linguistic constraint that such attacks cannot satisfy without jointly modeling speech semantics, substantially increasing adversarial complexity.

Concatenation and Splicing Attacks:
Phoneme concatenation or speech splicing may preserve acoustic plausibility but disrupt syntactic cohesion and word boundaries. ASR-based validation exposes these inconsistencies.

6.4. Mathematical Formulation of Layered Validation

Let: $f1: A \rightarrow 0,1$ be acoustic classifier

$f2: T \rightarrow 0,1$ be linguistic verifier $f2: T \rightarrow 0,1$

Then the multi-layer system is:

$$F(a) = \{1, iff1(a) = 1 and f2\big(ASR(a)\big) = 1$$
otherwise, 0

A successful attack must satisfy both constraints:

$$f1(aadv) = 1 \wedge f2\big(ASR(aadv)\big) = 1$$

Where:

- $a \in Aa \in Aa \in A$ is the input audio sample
- $ASR(a)\square\square\square(a)ASR(a)$ is the transcription generated by the Automatic Speech Recognition (ASR) system
- $F(a) = 1F(a) = 1F(a) = 1$ indicates accepted/authentic speech
- $F(a) = 0F(a) = 0F(a) = 0$ indicates rejected/spoofed speech

This requires simultaneous acoustic plausibility, linguistic coherence, and semantic alignment, making adversarial optimization substantially more difficult than in single-layer systems.

6.5. Advantages of Multi-Layered Frameworks

Multi-layered detection provides increased attack complexity, resistance to non-semantic perturbations, and robustness through modality independence. Since acoustic and linguistic layers rely on different signal properties, vulnerabilities do not easily transfer across layers. Importantly, layered defenses can be integrated with existing detectors without retraining and scale naturally across authentication, voice assistants, fraud detection, and verification systems.

6.7. Limitations and Practical Considerations

Layered frameworks introduce dependencies on ASR accuracy, language-specific constraints, susceptibility to ASR-focused adversarial attacks, and additional computational overhead. Nevertheless, empirical and theoretical evidence shows that ASR-augmented systems significantly outperform single-layer detectors in robustness and real-world reliability.

7. Comparative Analysis: Single-Layer vs Multi-Layer Systems

A systematic comparison between single-layer detectors and multi-layer security architectures is necessary to evaluate their robustness, practicality, and suitability for real-world audio deepfake detection. Single-layer systems rely exclusively on acoustic features, whereas multi-layer frameworks integrate linguistic and semantic validation, resulting in a substantial reduction in adversarial success rates.

7.1. Evaluation Dimensions

- The comparison is structured along six dimensions:
- Feature representation and information preservation
- Robustness against adversarial attacks
- Generalization across datasets and conditions
- Semantic and linguistic awareness
- Adversarial bypass difficulty
- Computational and deployment considerations

Across all dimensions, multi-layer frameworks demonstrate superior resilience.

7.2. Robustness Against Adversarial Attacks

Empirical evidence shows that single-layer detectors are easily bypassed by GAN-generated histograms, gradient-based perturbations, reversed audio, and noise-based attacks. Multi-layer systems impose an additional linguistic constraint, causing adversarial success rates to drop sharply by rejecting semantically invalid or incoherent inputs.

7.3. Empirical Comparison (Conceptual Summary)

Table 1. Comparative Performance of Single-Layer vs Multi-Layer Detectors

Attack Type	Single-Layer	Multi-Layer (ML + ASR)	Key Observation
GAN-based histograms	~99% bypass	<5% bypass	ASR rejects invalid language
Noise-based attacks	~91% bypass	0% bypass	No meaningful transcription
Reversed audio	~87% bypass	<40% bypass	Phonetic inconsistency detected
Spliced/concatenated speech	High	Moderate–Low	ASR detects discontinuities
Unseen synthesis models	Poor	Stronger	Linguistics model-agnostic
Cross-dataset inputs	Low	Moderate–High	Semantics preserved
Human-imitating GAN speech	High	Significantly reduced	Linguistic constraints enforced

7.4. Theoretical Attack Difficulty

Single-layer adversarial objective:
Find $f1\big(g(aadv)\big) = 1 \wedge f2\big(h(aadv)\big) = 1$
Multi-layer adversarial objective:
Find $f1\big(g(aadv)\big) = 1 \wedge f2\big(h(aadv)\big) = 1$

the problem shifts from single-objective to multi-objective optimization, requiring both acoustic deception and linguistic correctness—substantially increasing attack complexity.

7.5. Computational Considerations

Table 2: Computational Considerations of Single-Layer and Multi-Layer Systems

Component	Single-Layer	Multi-Layer
Acoustic classification	Fast	Fast
Linguistic verification	---	Medium
End-to-end latency	Low	Moderate
Deployment complexity	Low	Medium

The modest increase in computational cost is outweighed by substantial security gains.

7.6. Summary of Comparative Insights

Across all analytical dimensions:

- Multi-layer frameworks consistently outperform single-layer detectors
- Linguistic verification compensates for acoustic-only weaknesses
- Theoretical modeling confirms reduced attack regions
- Empirical results show near-elimination of noise and GAN-based attacks
- Generalization across datasets and synthesis methods is significantly improved

Therefore, multi-layered audio deepfake detection frameworks are not merely advantageous—they are essential for secure deployment in adversarial environments.

8. Discussion

The preceding analysis establishes that multi-layered audio deepfake detection frameworks offer substantially greater robustness than traditional single-layer detectors. Acoustic-only systems—based on MFCCs, spectrograms, histograms, or neural classifiers—operate in a reduced information space and lack awareness of linguistic validity, semantic coherence, and phonetic structure. Adversarial attacks exploit this limitation by generating acoustically plausible but meaningless signals, reversed speech, noise-based perturbations, or GAN-generated feature representations, leading to consistent real-world failure across datasets and architectures.

Multi-layered frameworks mitigate these weaknesses by combining acoustic classification with ASR-based linguistic verification. This integration enforces complementary constraints on both sound characteristics and linguistic structure, causing attacks that bypass acoustic detectors—such as reversed audio, noise, GAN-generated histograms, and spliced speech—to fail at the linguistic level. As a result, adversarial success rates are drastically reduced across all evaluated threat models.

From an adversarial perspective, attacking a multi-layer system requires satisfying a multi-objective constraint:

$$f1\big(g(aadv)\big) = 1 \wedge f2\big(h(aadv)\big) = 1$$

Optimizing simultaneously for acoustic plausibility and linguistic correctness introduces conflicting objectives, significantly increasing attack complexity and computational cost. Even advanced synthesis and GAN-based attacks struggle to meet both constraints consistently.

Although ASR introduces an additional component, attacks on ASR require precise phonetic manipulation and must still bypass the acoustic layer, substantially shrinking the feasible adversarial space. Despite challenges such as ASR variability, computational overhead, and multilingual speech handling, layered frameworks remain far more robust and practical than single-layer alternatives.

Overall, the evidence confirms that acoustic-only detection is fundamentally insufficient. Multi-layered architectures represent a necessary paradigm shift toward secure, trustworthy, and adversarially resilient audio deepfake detection systems.

9. Future Directions

Although multi-layered audio deepfake detection frameworks significantly improve robustness over single-layer detectors, rapid advances in speech synthesis and adversarial generation require continued evolution. This section outlines key research directions

essential for building resilient, scalable, and context-aware audio authentication systems.

Future research directions include:

- Multimodal biometric integration
- Adversarially robust ASR
- Semantic and contextual reasoning
- Adaptive defenses
- LLM-based speech understanding
- Formal robustness verification
- Context-Aware and Task-Specific Verification
- Real-Time and Edge Deployment
- Formal Verification of Layered Systems

10. Conclusion

The rapid evolution of audio deepfake generation techniques has exposed fundamental weaknesses in traditional single-layer acoustic detection systems. As shown throughout this review, detectors based solely on acoustic features such as spectrograms or histograms are inherently vulnerable due to non-injective feature mappings, loss of temporal and semantic information, and the false assumption that acoustic realism implies authenticity. These structural limitations allow adversarial attacks—ranging from noise perturbations to GAN-generated feature representations—to bypass detection with high success rates, even in advanced CNN- and RNN-based models.

This review demonstrates that the failure of single-layer detectors is not model-specific but architectural in nature. Acoustic plausibility alone is insufficient to guarantee semantic validity, yet single-layer systems rely exclusively on this assumption. In contrast, multi-layered detection frameworks fundamentally alter the security landscape by integrating acoustic classification with linguistic verification through ASR systems. This layered approach enforces simultaneous acoustic and semantic constraints, transforming adversarial evasion into a significantly harder multi-objective problem.

Both theoretical modeling and empirical evidence confirm that multi-layered architectures drastically reduce adversarial success across noise-based, reversed-audio, splicing, and GAN-based attacks. Linguistic verification restores critical phonetic and semantic information absent from feature-only detectors, while the combined decision boundaries reduce attack feasibility under diverse threat models and improve generalization across datasets and synthesis techniques.

While challenges remain—including ASR robustness, multilingual handling, computational efficiency, and adaptation to future synthesis models—the evidence clearly establishes multi-layered frameworks as the most viable path forward. Effective audio deepfake defense must move beyond surface-level acoustic analysis toward integrated linguistic, semantic, and contextual validation. Such layered systems offer a foundation for secure, scalable, and trustworthy audio authentication capable of operating in increasingly adversarial environments.

11. References

[1] Rabhi, M., Bakiras, S., & Di Pietro, R. (2024). Audio-deepfake detection: Adversarial attacks and countermeasures. *Expert Systems with Applications*, 250, Article 123941. https://doi.org/10.1016/j.eswa.2024.123941

[2] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499.*

[3] Arık, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., & Shoeybi, M. (2017). Deep voice: Real-time neural text-to-speech. *Proceedings of the 34th International Conference on Machine Learning (ICML).*

[4] Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N., Yamagishi, J., & Lee, K. A. (2017). The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In *Proceedings of Interspeech 2017*, pp. 2–6. https://doi.org/10.21437/Interspeech.2017-1111

[5] Ballesteros, D. M., Rodriguez-Ortega, Y., Renza, D., & Arce, G., Expert Systems with Applications, Vol. 184, 2021, Article 115465

[6] Alegre, F., Amehraye, A., & Evans, N. (2013). A spoofing countermeasure for the protection of automatic speaker verification systems against synthetic speech. In 2013 IEEE ICASSP (pp. 3067–3071). IEEE

[7] Lai, C. I., Chen, T., Lee, A., Tsao, Y., & Lin, C. (2021). Attentive filtering networks for audio spoofing detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1711–1726.

[8] Lavrentyeva, G., Novoselov, S., Malykh, E., Gorlanov, A., & Kozlov, A. (2017). Audio replay attack detection with deep learning frameworks. In Interspeech 2017 (pp. 82–86).

[9] Chen, M. Y., Lai, C. H., Tsao, Y., Chen, H. M., & Wang, H. M. (2020). Generative adversarial networks for audio anti-spoofing. IEEE Transactions on Information Forensics and Security, 15, 2126–2137.

[10] Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., & Larcher, A. (2021). End-to-end anti-spoofing with RawNet2. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)* (pp. 6369–6373). IEEE.

[11] Wu, Z., Gao, S., Chng, E. S., & Li, H. (2014). A study on replay attack and anti-spoofing for text-dependent speaker verification. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1–5). IEEE.

[12] Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331.
[13] Carlini, N., & Wagner, D. (2018). Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)* (pp. 1–7). IEEE.
[14] Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. arXiv:1607.02533.
[15] Qin, Y., Carlini, N., Goodfellow, I., Cottrell, G. W., & Raffel, C. (2019). Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In ICML.