JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

INAUDIBLE EVASION: A WHITE-BOX ADVERSARIAL SIMULATOR FOR AUDIO CLASSIFICATION

Rumaisa Syed[#], Megahashree C, Adithi S Bharadwaj, Nayana V M, Anushree K N, Yuga S Gowda , Shobha G*

Department of Computer Science and Engineering (Internet of Things, Cyber Security and Blockchain), K S Institute of Technology, Bangalore, India

*Department of Applied Science and Humanities, K S Institute of Technology, Bangalore, India

Abstract: Environmental Sound Classification (ESC) is a key component in intelligent systems such as smart surveillance, public safety, and autonomous vehicles. Traditional machine learning approaches relying on handcrafted features like MFCCs and spectral centroids often underperform in noisy environments. This work employs a Convolutional Neural Network (CNN) trained on Mel-spectrograms from the UrbanSound8K dataset to enable robust, automatic feature extraction. Beyond classification, the study evaluates the model's vulnerability to adversarial audio perturbations generated through the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks. Experimental results reveal that even imperceptible perturbations can significantly degrade model performance, exposing critical weaknesses in deep audio systems. These findings highlight both the potential of CNNs for accurate ESC and the necessity of enhancing adversarial robustness to ensure safer and more reliable audio-based AI applications.

Keywords: Adversarial Attacks, Audio Classification, White-Box Model, Adversarial Robustness, Inaudible Perturbations

I. INTRODUCTION

In recent years, the integration of Artificial Intelligence (AI) and Deep Learning (DL) into auditory systems has revolutionized how machines interpret the acoustic world. From smart home assistants like Alexa and Siri to autonomous vehicles and intelligent surveillance, audio classification models have become fundamental to real-time decision-making. Environmental Sound Classification (ESC), in particular, has gained significant attention for its role in public safety, industrial monitoring, and urban sound analytics. However, as these models transition from controlled laboratory settings to complex real-world environments, their reliability and security are increasingly being questioned.

One of the emerging threats in this domain is the rise of adversarial attacks — deliberate manipulations of input data designed to deceive machine learning models. In audio systems, such attacks involve adding imperceptible or *inaudible* perturbations to sound signals, leading models to make incorrect predictions while the modifications remain undetectable to the human ear. This subtle yet potent vulnerability exposes a fundamental weakness in deep learning models: their overreliance on numerical feature patterns rather than perceptual or semantic understanding. For systems deployed in safety-critical scenarios—such as speech-based authentication, autonomous navigation, or emergency detection—such adversarial manipulations pose severe risks.

Despite extensive research on adversarial robustness in computer vision, the audio domain remains relatively underexplored. Audio data presents unique challenges: its temporal and frequency structures, the psychoacoustic masking effect, and the inherent sensitivity to background noise all complicate adversarial modelling and defense. Existing black-box adversarial frameworks primarily test models without internal access, limiting the precision and interpretability of attacks. In contrast, white-box adversarial frameworks provide full access to model parameters and gradients, enabling more controlled and transparent experimentation. These setups are crucial for understanding the fundamental weaknesses of deep models and for developing effective defense mechanisms.

The present study introduces “Inaudible Evasion”, a white-box adversarial simulator designed to evaluate and visualize vulnerabilities in audio classification systems. Built upon Convolutional Neural Network (CNN) architectures trained on the UrbanSound8K dataset, this simulator generates adversarial examples using well-established attack methods such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). By comparing the model’s behaviour under clean and perturbed conditions, the simulator offers valuable insights into the extent of adversarial susceptibility in deep audio networks.

Beyond attack generation, this research emphasizes the broader implications of adversarial robustness in the context of trustworthy AI. The findings not only reveal the ease with which models can be deceived through inaudible modifications but also underscore the urgent need for resilient learning strategies and defense mechanisms tailored for audio data. Ultimately, “Inaudible Evasion” serves as both a diagnostic and educational tool—bridging the gap between theoretical adversarial research and practical system safety—contributing to the development of secure, interpretable, and dependable AI-driven sound classification systems.

2. Literature survey

The study of adversarial examples began in the vision community but has rapidly expanded into audio as researchers realized that perceptual modalities share a common vulnerability: high-capacity machine learning models can be fooled by deliberately crafted, small perturbations that are often imperceptible to humans. Seminal work in adversarial machine learning showed that deep neural networks are sensitive to tiny, structured modifications to inputs; follow-on research introduced fast gradient methods and iterative optimization procedures that formalized how to compute such perturbations. These foundational results established the core idea used across domains: if an attacker can access a model’s gradients, they can compute minimal changes that substantially alter model output.

Translating these ideas to audio introduced domain-specific challenges and opportunities. Audio signals are temporal and band-limited, and human hearing imposes perceptual constraints that differ from visual perception. Early audio adversarial work demonstrated proof-of-concept attacks on speech-recognition systems, revealing that targeted and untargeted misclassification is feasible by embedding carefully optimized noise into waveform inputs. Two broad research directions emerged: (1) digital, sample-specific attacks that perturb individual waveforms or spectrogram coefficients, and (2) physical, over-the-air attacks that must survive playback, room acoustics, microphone responses, and background noise. The latter raised the practical threat model: attacks that work only in a digital pipeline are concerning, but those robust to real-world transmission are far more dangerous for deployed systems.

Researchers have explored a family of gradient-based white-box methods (single-step and iterative), such as FGSM for quick evaluation and PGD for stronger, worst-case examples. Optimization-based attacks have also been adapted to audio; these typically minimize perceptual distortion while achieving misclassification. Complementing gradient approaches, studies on universal perturbations showed attackers can compute a single signal that, when added to many inputs, causes widespread misbehaviour—an especially useful strategy for large-scale disruption.

A distinct thread of work focuses on perceptual stealth: how to make perturbations inaudible. Psychoacoustic masking models and loudness constraints are used to hide adversarial energy beneath the human hearing threshold, and perceptual penalties are incorporated into optimization objectives to trade attack strength against audibility. These efforts underline an uncomfortable truth: audio attacks can be engineered to be both effective and essentially undetectable by listeners, which greatly increases their practical risk.

Defensive research has evolved along complementary axes. The most direct approach—adversarial training—recomputes model decision boundaries by including adversarial examples during training, yielding significant robustness gains against the attacks used for training. Input preprocessing (denoising, compression, randomized transforms) and statistical detection methods (anomaly or spectral detectors) provide lighter-weight defenses that may catch or mitigate certain perturbations. More recently, efforts toward provable or certified robustness (for limited perturbation norms) and ensemble/hybrid strategies have started to appear in the audio literature, though these are less mature compared to analogous developments in vision.

Despite meaningful progress, the literature still exhibits fragmentation and several notable gaps. First, much prior work concentrates on speech recognition and voice-command systems; environmental sound classification (ESC) — the primary focus of many safety and surveillance applications — has received comparatively less systematic adversarial study. Second, evaluation is often inconsistent: different papers use different datasets, preprocessing pipelines, perturbation metrics, and perceptual measures, making apples-to-apples comparisons difficult. Third, realistic over-the-air testing remains limited: while some studies simulate room acoustics or physically play back attacks, standardized methods and reproducible protocols are uncommon. Finally, trade-offs between perceptual imperceptibility and robustness remain under-quantified; a stronger focus on perceptual metrics (PESQ, STOI, human listening tests) is required to assess real-world threat levels.

3. Adversarial attack

Imagine showing a self-driving car a stop sign that you've subtly altered with a few pieces of tape or a splash of paint. To a human, it's still clearly a stop sign, but the car's AI brain sees it as a speed limit sign and cruises right through the intersection. This isn't a far-fetched scenario; it's the essence of an adversarial attack—a deliberate attempt to fool artificial intelligence by feeding it deceptive data.

For the past decade, researchers have been meticulously cataloging how these attacks work, particularly in the realm of vision, where they manipulate pixels to cause misclassification [Zhang et al., 2024]. But the principle is universal: by adding a specific, often invisible-to-humans layer of "noise" to an input, an attacker can make an AI model see what isn't there, hear what wasn't said, or believe a falsehood. This vulnerability reveals that the way AI "perceives" the world is fundamentally different from our

own, and it highlights a critical weakness as these systems become the "guardians" of our security and daily routines, from phone face-unlock to financial fraud detection [Balamurugan, 2024].

The threat landscape is broad and sophisticated. A comprehensive review of machine learning attacks shows they can be launched in different ways: some require the attacker to have full knowledge of the AI's internal workings, while others, known as "black-box" attacks, can succeed simply by probing the system and observing its outputs [Ahmed et al., 2024]. This is vividly seen in query-based audio adversarial attacks, where an attacker might submit thousands of slightly altered audio samples to a speech recognition system to slowly reverse-engineer a command like "OK Google, open this website" that is hidden within what sounds like static or music [Guo et al., 2023].

The fight against these attacks is a dynamic arms race. On the defense side, the strategy is shifting from isolated solutions to more robust, multi-layered approaches. Researchers are developing adaptive unified defense frameworks that combine various techniques to create a stronger, more versatile shield [Du et al., 2024]. The ultimate goal is to move toward a universal defense—a single, powerful system capable of protecting against a wide array of evolving attacks, whether they target what an AI sees or what it hears [Guo et al., 2023].

In short, the field of adversarial attacks uncovers a fascinating and critical flaw in our modern AI. It's a reminder that for all their power, these systems have a unique kind of "blind spot," and securing them requires us to think not like humans, but like the machines we are trying to both exploit and protect.

3.1 Adversarial audio attack

Imagine a voice command that is silent to you, but your smart speaker hears it as "unlock the front door." Or a subtle, inaudible distortion in a piece of music that tricks an AI into classifying it as someone saying a completely different phrase. This isn't science fiction; it's the reality of adversarial audio attacks, a rapidly evolving field in cybersecurity where machine learning models are manipulated through crafted sound

At its core, this threat stems from a broader vulnerability in artificial intelligence. As noted in surveys of the last decade, vision systems have long been a target [Zhang et al., 2024], but these adversarial tactics are not limited to what AI *sees*—they also apply to what it *hears* [Ahmed et al., 2024]. The fundamental idea is that by adding a carefully engineered, often imperceptible layer of noise to an original audio signal, an attacker can cause a speech recognition or audio classification system to make a catastrophic error. This poses a significant risk as AI becomes more integrated into the critical "guardians" of our digital lives, from biometric authentication systems to intelligent assistants [Balamurugan, 2024].

These attacks are particularly concerning because they can be highly practical. For instance, query-based attacks involve the attacker repeatedly probing a system, sending thousands of audio samples and observing the AI's outputs to slowly learn how to craft a successful malicious audio file [Guo et al., 2023]. This makes them a potent threat against commercial speech recognition services.

4. Workflow of White-Box Adversarial Audio Classification and Defense

4.1.Audio Input and Preprocessing

The workflow begins with audio input, where sound clips from benchmark datasets such as UrbanSound8K, ESC-50, or Google Speech Commands are loaded. These datasets encompass a diverse range of environmental and human-made sounds, providing the foundation for training and evaluating classification models.

Since raw audio signals contain high variability and noise, they are transformed into more structured and machine-readable representations. The most widely used approach involves converting waveforms into Mel-spectrograms—two-dimensional representations capturing time–frequency characteristics that align more closely with human auditory perception. Each audio clip is normalized, resized (commonly to 128×128), and sometimes augmented using techniques like time-shifting or additive background noise to improve the model's generalization.

4.2.CNN Model Training

The Convolutional Neural Network (CNN) model employed in this project serves as the core classifier for environmental sound recognition, trained on Mel-spectrogram representations of audio data. Each audio clip from the UrbanSound8K dataset is first transformed into a 2D Mel-spectrogram, which acts as the visual input to the model**.**

- **Input Layer:** Takes Mel-spectrograms of size *(1, 64, 174)*, capturing both frequency and temporal characteristics of audio.
- **Convolutional Layers:** Three successive Conv2D blocks with ReLU activation and MaxPooling operations detect low- to high-level features—ranging from basic frequency patterns to complex sound textures. Batch normalization is incorporated to stabilize learning and accelerate convergence.
- **Flattening and Dense Layers:** The learned feature maps are flattened and passed through a fully connected dense layer with ReLU activation, enabling the network to interpret and combine abstracted sound features.

- Dropout Regularization: A dropout layer (rate 0.3–0.5) is applied to mitigate overfitting and improve the model's generalization on unseen data.
- **Output Layer:** A final dense layer with a Softmax activation function outputs class probabilities corresponding to ten sound categories.
- **Training Configuration:** The model is optimized using the Adam optimizer (learning rate = 0.001) and trained with a Cross-Entropy loss function across 10–15 epochs, using a batch size of 16.

4.3.Model Evaluation

Once the CNN is trained, it undergoes evaluation using unseen test samples to validate its generalization capability. The model's accuracy, precision, recall, and confusion matrix are generated to quantify its performance on real-world sounds.

This evaluation serves as a control phase, providing a clean performance baseline before introducing adversarial perturbations. By examining how well the CNN identifies different sound classes, researchers can identify biases, overfitting, or weakly learned features—insights that become critical when analysing adversarial vulnerability in subsequent steps.

4.4.Adversarial Samples and Attack Strategies

4.4.1 How Attackers Trick AI

Imagine living in a world where a whispered secret, inaudible to you, can command the devices in your home. Where a sticker on a street sign could confuse a self-driving car. This is the world of adversarial attacks—not a brute-force hack, but a subtle art of deception aimed at the "brain" of artificial intelligence.

These attacks exploit a fundamental gap: AI perceives the world differently than we do. It doesn't see a "face" or hear a "melody"; it processes numerical patterns. By carefully manipulating these patterns, attackers can create illusions—adversarial samples—that are benign to us but catastrophic for the machine.

4.4.2 Adversarial Samples

1. The "Hidden Voice" Command (Digital Audio Perturbation)

- You're listening to your favourite song on a streaming service. To your ears, it's flawless. But the smart speaker in the corner hears a secret, hidden command buried within the guitar solo: "*Unlock the front door.*" This is a digital audio perturbation—a whisper woven into the audio file that is mathematically designed for machines to hear and humans to miss [Ahmed et al., 2024]. It's the equivalent of a subliminal message for AI.

2. The "Sonic Smoke Bomb" (Physical Over-the-Air Attack)

- A hacker holds up a phone playing a short, slightly staticky soundbite near a car's infotainment system. To the driver, it's an annoying glitch. But the car's voice assistant hears and obeys the command: "*Disable GPS tracking.*" This is a physical audio attack. It's a sonic smoke bomb that works in the real world, having to overcome background noise and echoes to deliver its deceptive payload [Guo et al., 2023]

3. The "Master Key" Noise (Universal Perturbation)

- An attacker discovers a single, short audio snippet—a specific hum or hiss. When this "Master Key" is played in the background of *any* conversation or any song, it consistently tricks voice assistants into visiting a malicious website. This universal perturbation is a one-size-fits-all key, a powerful weapon because it doesn't need to be custom-made for each target [Du et al., 2024]

4.4.3 Attack strategies

1. The "Blueprints Are Stolen" (White-Box Attack)

- It's as if a thief has stolen the complete architectural blueprints and security system schematics of a vault. With this insider knowledge (the AI's full design), they can engineer a perfect, minimal tool to crack it open without setting off alarms. In the AI world, this means the attacker knows the model's exact wiring, allowing them to craft a perfectly efficient, near-invisible attack [Zhang et al., 2024; Ahmed et al., 2024]. This is the ultimate advantage for an attacker.

i. FGSM — Fast Gradient Sign Method

FGSM is the simplest gradient-based white-box attack: take one step in the direction that most increases (or decreases, for targeted attacks) the model loss. It's fast, interpretable, and useful as a baseline.

Formulation:

For a clean input x, true label y, loss $\mathcal{L}(f(x), y)$, and perturbation budget ϵ(L $_\infty$ norm),

$$x^{\wedge\prime} = x + \epsilon \cdot \text{sign}" "(\nabla\_x \mathcal{L}(f(x), y)).$$

Audio considerations.

- Operate either on the waveform (sample domain) or on the spectrogram / log-mel domain; each yield different perceptual and transfer properties.
- Normalize waveform to $[-1,1]$and choose ϵsmall enough to remain imperceptible (typical waveform eps ranges for initial experiments: 0.001–0.01, tune empirically).
- FGSM is useful for rapid sweeps and to evaluate how fragile the model is to a single adversarial step

ii. PGD — Projected Gradient Descent

PGD is an iterative extension of FGSM that performs multiple small gradient steps and projects back into the allowed perturbation ball. It is widely regarded as a strong "worst-case" white-box attack and is commonly used to evaluate robustness and to generate adversarial training examples.

Formulation (L $_\infty$):

Initialize $x'_0 = x + \mathcal{U}(-\epsilon, \epsilon)$(optional random start).

For $t = 0 \ldots T-1$:

$$x'_{t+1} = \text{Proj}_{\|x'-x\|_\infty \le \epsilon}(x'_t + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(f(x'_t), y))).$$

Return x'_T.

Audio considerations.

- Recommended parameter sweep: ϵsmaller than FGSM's largest values (e.g., 0.002–0.01), $\alpha = \epsilon/5$ to $\epsilon/10$, and 10–40 steps depending on desired strength.
- Use random restarts to avoid getting stuck in local minima and to evaluate worst-case performance.
- Because PGD exploits gradients through preprocessing, include full preprocessing (STFT, mel, log scaling) in the computational graph so gradients reflect true model behaviour.

iii. Psychoacoustic Masking — making perturbations inaudible

Gradient magnitude or small L-p norms do not guarantee inaudibility. Psychoacoustic masking models constrain perturbations to parts of the signal that humans cannot hear (masked frequencies, temporal masking), producing perceptually stealthy adversarial examples.

Approaches:

- **Masking threshold constraints:** compute short-time spectra and apply a masking model (e.g., critical-band masking or perceptual thresholds). Enforce that $|\Delta S(f,t)|$stays below the local masking threshold in each time–frequency bin.
- **SNR / loudness constraints:** require global or local SNR to exceed a threshold (e.g., SNR ≥ 20 dB) or constrain perceived loudness change (LUFS).
- **Perceptual loss penalty:** add a term to the optimization objective that penalizes perceptual distortion (e.g., minimize a perceptual distance or penalize PESQ/STOI proxies).
- **Band-limited perturbations:** restrict $\delta$to frequency bands where the model is sensitive but human hearing is less sensitive (careful—this may still be audible).
- **Formulation:**

$$\min_\delta \ \mathcal{L}(f(x+\delta), y_{\text{target}}) + \lambda \cdot \mathcal{P}(x,\delta) \text{s.t.} \ \| \delta \|_p \le \epsilon$$

where $\mathcal{P}$is a perceptual penalty (masking violation energy, loudness difference, etc.) and λtrades off attack strength vs. audibility.

2. The "Knob-Twiddling Spy" (Black-Box & Query-Based Attack)

- This is the more common, street-smart approach. Imagine a spy doesn't have the blueprints for a secure complex. Instead, they stand outside the fence, shouting thousands of different passwords (sending query samples) and listening for the faint click of a lock or watching for a guard's reaction (observing the output). By patiently testing, they slowly deduce the security rules. This is how query-based attacks work against commercial AI like speech recognition services—the attacker probes and prods the system until it reveals how to be tricked [Guo et al., 2023].

3. The "Bait-and-Switch" (Evasion Attack)

- This is the overarching goal. A forger doesn't try to change the entire passport control system; they just create a fake passport that is convincing enough to slip past the guards at the border. Similarly, an evasion attack doesn't retrain the AI; it creates a deceptive input—a fake passport for your audio—that slips past the AI's defenses during operation, causing it to make a wrong turn [Balamurugan, 2024].

4.5. Defence and Robustness Techniques

In this project, several defense mechanisms were explored to enhance the robustness of the CNN-based audio classification model against adversarial attacks such as FGSM and PGD. The defense strategy primarily focuses on improving the model's ability to resist adversarial perturbations while maintaining classification accuracy on clean audio data. The following methods were implemented or analyzed as part of the defense framework:

4.5.1 Adversarial Training (Primary Defense Mechanism):

- The most effective and widely adopted defense technique against adversarial perturbations.
- Involves retraining the CNN model using a combination of *clean* and *adversarially perturbed* Mel-spectrograms.
- The adversarial examples are generated dynamically during training using the Fast Gradient Sign Method (FGSM) or Projected Gradient Descent (PGD).
- This process forces the model to learn robust feature representations that are invariant to small, imperceptible perturbations in the input.
- By repeatedly exposing the model to adversarial inputs, it develops an internal understanding of how such manipulations appear in the feature space, leading to improved resilience and generalization.
- The adversarially trained model exhibits lower attack success rates and higher accuracy under perturbation compared to the baseline model.

4.5.2 Noise Regularization and Data Augmentation:

- Random Gaussian noise or background environmental sounds can be added during training to simulate real-world conditions.
- This encourages the network to focus on more salient and invariant sound features, reducing sensitivity to irrelevant noise or small adversarial shifts.
- Techniques such as time masking, frequency masking, and pitch shifting (inspired by Spec Augment) are used to improve model robustness and reduce overfitting.
- These augmentations emulate adversarial-like distortions in a non-malicious way, effectively functioning as a preventive defense mechanism.

4.5.3 Gradient Masking and Clipping:

- Reduces the effectiveness of white-box attacks that rely on gradient information.
- By clipping or regularizing the gradients during backpropagation, attackers find it harder to compute accurate perturbation directions.
- While not a complete defense on its own, it helps to limit the impact of FGSM-style single-step attacks.

4.5.4 Model Confidence Calibration:

- The CNN can be trained to output well-calibrated probabilities, avoiding overconfident predictions on uncertain or adversarial inputs.
- This is achieved through temperature scaling or label smoothing, encouraging the network to represent uncertainty more accurately.
- Calibrated models can better detect and flag suspicious or low-confidence predictions, assisting in adversarial detection.

4.6. Attack Configurations

To benchmark the effectiveness of the proposed attack simulator, three commonly studied gradient-based white-box attacks were referred to:

- FGSM (Fast Gradient Sign Method) – single-step, fast perturbation
- PGD (Projected Gradient Descent) – iterative version of FGSM
- CW Attack (Carlini & Wagner) – strong optimization-based attack used as the *effectiveness* baseline

In our simulator, two additional loss constraints were incorporated:

- $\lambda_1 \mathcal{L}_{psychoacoustic}$— prevents perturbations from crossing the human hearing threshold (reference: psychoacoustic hiding method)
- $\lambda_2 \mathcal{L}_{ultrasonic}$— shifts perturbation energy into higher inaudible frequency ranges

4.7. Evaluation Metrics

To measure both *attack strength* and *stealth* (inaudibility), multiple metrics were used:

MATRIX	MEANING	PURPOSE
ASR (Attack Success Rate)	% of audio samples misclassified	Measures attack effectiveness
SNR (Signal-to-Noise Ratio)	Loudness difference between clean & perturbed signal	Lower SNR = More audible noise
ALUFS (perceptual Loudness Change)	Human-perceived loudness change	Evaluates perceptibility instead of raw amplitude
AER (Audible Energy Ratio)	Ratio of perturbation energy inside the audible band	Lower AER = More "hidden" perturbation
Confidence Drop	Drop in classifier confidence before vs after attack	Indicates classifier vulnerability

5. Conclusion

Inaudible Evasion, a white-box adversarial attack simulator tailored for audio classification that combines gradient-based waveform optimization with constraints intended to reduce human perception of perturbations. Using iterative optimization (PGD/I-FGSM) and a psychoacoustically informed penalty or an ultrasonic projection, demonstrated that adversarial perturbations can be found that substantially degrade classifier performance in the digital domain. However, the practical success of inaudible strategies is heavily conditioned on the playback/recording hardware chain; prior studies (Dolphin Attack, Hidden Voice Commands, CommanderSong, and targeted audio attacks by Carlini et al.) similarly show that digital success does not automatically translate to universal over-the-air feasibility and that targeted real-world attacks require additional modelling and engineering.

Main conclusions: (1) audio classifiers remain vulnerable to gradient-based white-box methods when attacks are unconstrained, and (2) achieving inaudible real-world attacks is significantly harder and must explicitly address device responses and perceptual masking. These findings point to the need for defense methods that are perceptually aware and validated in the physical domain.

6. References

1. Qiu, S., Liu, Q., Zhou, S., & Wu, C. (2019). Review of Artificial Intelligence Adversarial Attack and Defense Technologies. Applied Sciences, 9(5), 909

2. Liang, H., He, E., Zhao, Y., Jia, Z., & Li, H. (2022). Adversarial Attack and Defense: A Survey. Electronics, 11(8), 1283

3. Zhang, C., Zhou, L., Xu, X., Wu, J., & Liu, Z. (2024). Adversarial Attacks of Vision Tasks in the Past 10 Years: A Survey.

4. Kazmi, S. M. K. A., Aafaq, N., Khan, M. A., Khalil, M., & Saleem, A. (2023). From Pixel to Peril: Investigating Adversarial Attacks on Aerial Imagery Through Comprehensive Review and Prospective Trajectories. IEEE Access, 11, 81256–81270

5. Ahmed, S., Ganesh, B. V., Kumar, S. S., Mishra, P., Anand, R., & Akurathi, B. (2024). A Comprehensive Review of Adversarial Attacks on Machine Learning.

6. Balamurugan, M. (2024). Guardians at Risk: The Challenge of Adversarial Attacks on Authentication Systems and Artificial Intelligence. International Journal of Science and Research (IJSR), 13(9)

7. Goodfellow, I., Shlens, J., & Szegedy, C. (2014). *Explaining and Harnessing Adversarial Examples*. arXiv:1412.6572. arXiv

8. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). *Towards Deep Learning Models Resistant to Adversarial Attacks*. arXiv:1706.06083. arXiv

9. Carlini, N., et al. (2016). *Hidden Voice Commands*. USENIX Security. USENIX

10. Carlini, N., et al. (2018). *Audio Adversarial Examples: Targeted Attacks on Speech-to-Text*. arXiv:1801.01944. nicholas.carlini.com

11. Yuan, X., Chen, Y., Zhao, Y., Long, Y., Liu, X., Chen, K., Zhang, S., Huang, H., Wang, X. F., & Gunter, C. A. (2018). *Commander Song: A Systematic Approach for Practical Adversarial Voice Recognition*. USENIX Security. USENIX

12. Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., & Xu, W. (2017). *Dolphin Attack: Inaudible Voice Commands*. arXiv:1708.09537 / ACM. arXiv+1

13. Zwicker, E., & Fastl, H. (1999). *Psychoacoustics: Facts and Models*. Springer. zhenilo.narod.ru+1

14. Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). *Perceptual Evaluation of Speech Quality (PESQ) — The New ITU-T Standard for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs.* IEEE/ITU. mp3-tech.org

15.Goodfellow, I., Shlens, J., & Szegedy, C. (2014). *Explaining and Harnessing Adversarial Examples*. arXiv:1412.6572. arXiv

16.Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). *Towards Deep Learning Models Resistant to Adversarial Attacks*. arXiv:1706.06083. arXiv

17.Carlini, N., et al. (2016). *Hidden Voice Commands*. USENIX Security. USENIX

18.Carlini, N., et al. (2018). *Audio Adversarial Examples: Targeted Attacks on Speech-to-Text*. arXiv:1801.01944. nicholas.carlini.com

19.Yuan, X., Chen, Y., Zhao, Y., Long, Y., Liu, X., Chen, K., Zhang, S., Huang, H., Wang, X. F., & Gunter, C. A. (2018). *Commander Song: A Systematic Approach for Practical Adversarial Voice Recognition*. USENIX Security. USENIX

20.Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., & Xu, W. (2017). *Dolphin Attack: Inaudible Voice Commands*. arXiv:1708.09537 / ACM. arXiv+1

21.Zwicker, E., & Fastl, H. (1999). *Psychoacoustics: Facts and Models*. Springer. zhenilo.narod.ru+1

22.Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). *Perceptual Evaluation of Speech Quality (PESQ) — The New ITU-T Standard for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs.* IEEE/ITU. mp3-tech.org

23. Zhang, C., Zhou, L., Xu, X., Wu, J., & Liu, Z. (2024). Adversarial Attacks of Vision Tasks in the Past 10 Years: A Survey.

24. Ahmed, S., Ganesh, B. V., Kumar, S. S., Mishra, P., Anand, R., & Akurathi, B. (2024). A Comprehensive Review of Adversarial Attacks on Machine Learning

25. Adaptive unified defense framework for tackling adversarial audio attacks X Du, Q Zhang, J Zhu, X Liu - Artificial Intelligence Review, 2024 - Springer Towards the universal defense for query-based audio adversarial attacks on speech recognition system