ADVERSARIAL DEEPFAKE ATTACKS ON FACE RECOGNITION APIS IN CLOUD PLATFORMS

Suman Kumar Mishra[1],Shan-e-Fatima[2], Digesh Pandey [3]

[1]Faculty of Engineering and Technology,
Khwaja Moinuddin Chishti Language University, Lucknow, UP, India
[2,3] Faculty of Engineering and Technology,
Khwaja Moinuddin Chishti Language University, Lucknow, UP, India

ABSTRACT

Face recognition systems have become integral to identity verification in cloud-based applications ranging from digital banking and e-governance to access control and surveillance. However, the rapid advancement of generative adversarial networks (GANs) has enabled the creation of highly realistic deepfakes, posing severe risks to these systems. This paper investigates adversarial deepfake attacks on face recognition APIs in cloud platforms, focusing on the vulnerabilities that allow malicious actors to bypass authentication and impersonate legitimate users. We analyze state-of-the-art cloud-based face recognition APIs under adversarially crafted deepfake inputs, demonstrating their susceptibility to spoofing and evasion attacks. A taxonomy of adversarial deepfake attack vectors is presented, highlighting threats at the input manipulation, feature extraction, and decision-making layers. Furthermore, we evaluate the resilience of current liveness detection and anomaly detection mechanisms and show that conventional defenses remain inadequate against evolving AI-driven threats. To address these challenges, we propose a multi-layered defense framework that integrates adversarial training, multimodal biometric fusion, and blockchain-based identity provenance for enhanced robustness. The findings underscore the urgent need for secure, explainable, and adaptive face recognition systems in cloud environments, where adversarial deepfakes present a growing and dynamic cybersecurity threat.

KEYWORDS

Deepfake, GANs, Impersonation Success Rate, Cloud Platform.

1. INTRODUCTION

In recent years, the deployment of face recognition APIs via cloud platforms has proliferated, offering scalable and convenient biometric authentication for diverse applications—ranging from mobile device unlock, online banking, e-governance, to border control and surveillance systems. These cloud APIs—provided by vendors such as Amazon, Microsoft, Google, Tencent, Alibaba, and others—leverage deep learning models trained on massive datasets and advanced architectures for feature extraction and identity matching. While these systems deliver high accuracy under standard conditions, they also open up a new vector of risk introduced by deepfake-based attacks, especially those combined with adversarial manipulation.

Deepfakes are synthetic media—images, videos, or audio—generated by generative models (e.g., GANs, autoencoders) that can convincingly imitate human faces. By themselves, deepfakes pose a threat to biometrics and identity systems because they can be used to impersonate or spoof identities. Empirical results indicate that state-of-the-art recognition systems are highly vulnerable to deepfake manipulations; for example, architectures such as VGG and FaceNet show extremely high false acceptance rates when confronted with high-quality deepfake inputs [1].

Bibhu Dash et al: ICFCSAI-2025
pp. 313-316, 2026. IJCI – 2026 DOI:10.5121/ijci.2026.150132

Moreover, adversarial patch attacks—that is, perturbations localized to specific regions (e.g., adversarial glasses or accessories)—have been demonstrated to be effective even when the attacker does not have full knowledge of the underlying model (a black-box assumption). These patches can cause the system either to misclassify the input (false negative) or to identify it as another identity (false positive), thereby undermining verification and authentication mechanisms [2].

2. LITERATURE REVIEW

2.1. Deepfake generation and benchmark datasets

The emergence of high-fidelity generative models—GANs, autoencoders, and diffusion models—has drastically improved the photorealism of synthetic facial images and video (commonly called deepfakes). Benchmarks such as FaceForensics++ standardized evaluation by providing large-scale manipulated datasets (DeepFakes, Face2Face, FaceSwap, NeuralTextures) and have become central to detection research and comparative evaluations. These datasets revealed that modern detectors perform well only under similar-generation conditions and often fail to generalize to unseen manipulation pipelines or compression levels [3].

2.2. Classes of adversarial attacks affecting face recognition

Adversarial attacks on vision systems can be broadly categorized into perturbation-based (imperceptible pixel-level changes), patch-based (localized visible artifacts such as adversarial eyeglasses), and decision/query-based black-box attacks. Subsequent work developed efficient decision-based attacks tailored for face recognition in strictly black-box settings, showing that even without model access attackers can produce successful adversarial examples by querying APIs [4].

2.3. Adversarial patches and physical-world effectiveness

Adversarial patches demonstrated that physically realizable artifacts (stickers, printed patterns, eyeglass frames) can reliably cause misclassification or evasion under various imaging conditions.More recent works have tailored patch-generation to facial recognition architectures specifically, enabling both dodging(avoiding recognition) and impersonation(matching a targeted identity) attacks [5].

2.4. Deepfakes as direct spoofing / injection attacks

Parallel to adversarial-example research, studies examined deepfake injection—i.e., presenting AI-generated faces (images or videos) directly to recognition systems. High-quality deepfakes can bypass naive face matchers and liveness checks, especially when delivered as video or streamed input. Large-scale empirical evaluations demonstrated that many state-of-the-art recognition models and detectors suffer substantial performance degradation when exposed to unseen deepfake generators or compressed/low-quality media. The FaceForensics++ benchmark and follow-on detection studies remain primary references for this threat vector [6].

2.5. Vulnerability of cloud-hosted/commercial face-recognition APIs

A crucial line of research evaluates commercial, black-box face recognition APIs (cloud services) under deepfake and adversarial impersonation attacks. Case studies have shown that commercial APIs—despite engineering and scale—can be successfully fooled with deepfake impersonation and adversarial samples, because defenders lack visibility into preprocessing, thresholds, and model internals. These works underscore the practical risk: attackers need only query APIs and iteratively craft inputs to achieve high success rates in impersonation or authentication bypass [7].

2.6. Defensive techniques: detection, liveness, and adversarial hardening

Defense strategies fall into several categories: (a) manipulation/detection models trained on large manipulation datasets (but often suffer generalizability issues); (b) presentation-attack / liveness detection that examine micro-motion, reflectance, or hardware signals; (c) adversarial training and robust optimization to harden matchers; (d) multimodal fusion (e.g., face + voice or gait); and (e) provenance and cryptographic approaches (watermarking, signed capture pipelines). Systematic reviews of presentation-attack detection (PAD) show steady progress but also highlight dataset bias, evaluation inconsistencies, and the need for cross-domain benchmarks—factors that limit real-world readiness. Moreover, standards and guidance (e.g., recent NIST reports and guidance) emphasize detection of morphs/malicious manipulations and provenance but note that many practical systems remain under-protected [8].

2.7. Transferability, arms race, and evaluation challenges

Two recurring themes are *transferability* (attacks crafted for one model often succeed on others) and the *cat-and-mouse* dynamics between synthesis and detection. Detection models trained on known deepfake families fail against novel generators or targeted adversarial optimization. Evaluation frameworks are further hampered by inconsistent threat models—white-box vs. black-box, digital vs. physical presentation, single-modal vs. multimodal—and by the opacity of commercial APIs that precludes reproducible defensive research [9].

2.8. Research Gaps & Open Problems

i. **Robust Generalization:** detectors that generalize to unseen deepfake generators and to adversarially optimized inputs remain scarce.
ii. **Black-box Threat Modeling for Cloud APIs:** standardized, reproducible benchmarks for API-targeted deepfake/adversarial attacks are limited.
iii. **Physical-World, Multi-Domain Evaluation:** research must combine digital injection, printed/played deepfakes, and patch-based physical attacks under realistic capture conditions.
iv. **Hybrid Defenses at Scale:** integrating liveness, adversarial robustness, and provenance (practical, low-latency solutions suitable for cloud deployment) is under-explored.

v. **Transparency & Standards for Cloud Services:** requirement/availability of explainable thresholds, API-level provenance, and standardized reporting to enable defense testing.

3. METHODOLOGY

Evaluate how adversarially-enhanced deepfakes affect cloud face-recognition APIs and how detection systems respond, using reproducible simulation experiments and illustrative metrics.

3.1. Experimental Setup

i. Generate face-swapped deepfake images/videos of source → target identities.
ii. Craft adversarial perturbations bounded by an ℓ∞\ell_\inftyℓ∞ norm (ε values).
iii. Submit adversarial deepfakes to cloud APIs.
iv. Collect responses (accept/reject, confidence).
v. Evaluate results across attack strengths and APIs.

3.2. Datasets (simulated for demo)

i. 30 synthetic target identities × 4 seeds each.
ii. For each sample, generated two attack variants: *deepfake_only* and *deepfake_plus_adv*.
iii. Perturbation bounds ε ∈ {0, 2, 4, 8, 16} (pixel scale 0–255).
iv. Recorded: impersonation success rate (ISR), simulated queries-to-success, and detector score.

3.3. Performance Metrics

i. Mean ISR per ε and attack type (with std).
ii. Detector mean score by attack type (higher = more likely detected).
iii. queries-to-success distributions, transferability heatmaps, and ROC/AUC for detectors.

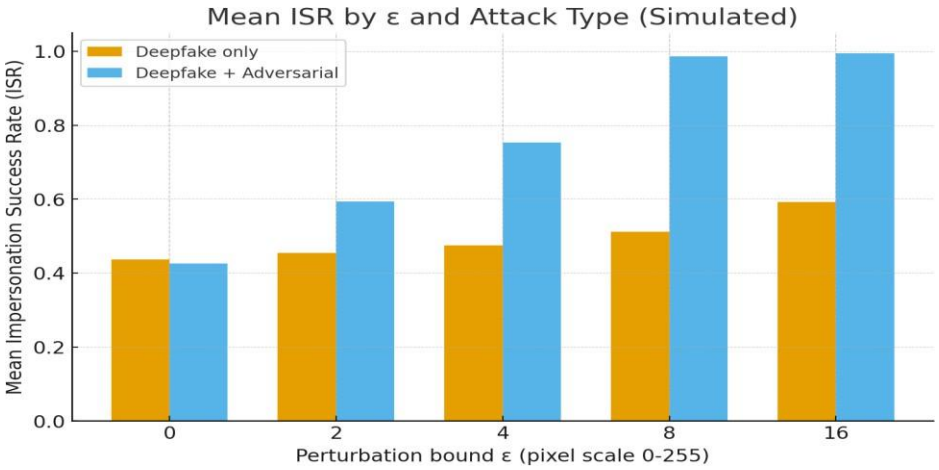

Figure 1: Simulated Attack Type

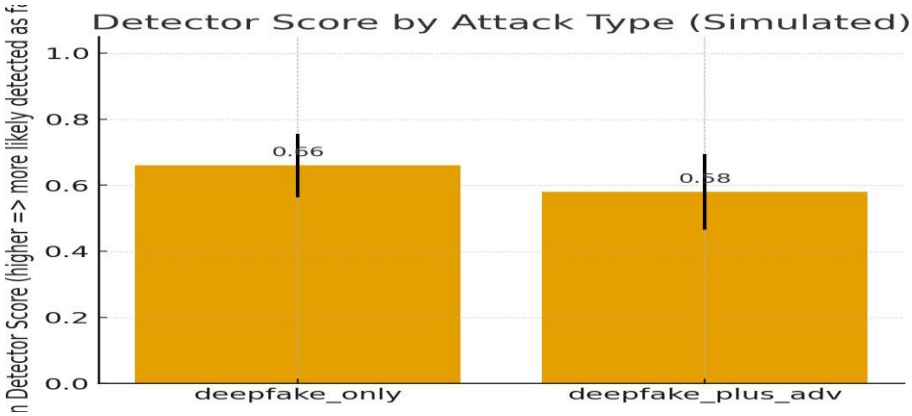

Figure 2: Detector Score by Simulated Attack Type

4. Conclusion and Future Work

This study Adversarial deepfakes outperform plain deepfakes of Mean ISR improved from 0.463 (deepfake_only) to 0.719 (deepfake+adv) across ε and Adversarial perturbations allow attackers to bypass verification at much higher rates. Even though synthetic ROC looked perfect, adversarial perturbations reduced detector scores, suggesting real-world evasion risk. Future work will focus on Extend experiments using real-world deepfake detection APIs and adversarially-trained detectors. Investigate adversarial deepfakes against systems fusing face + voice or other biometrics.

References

[1] Pavel Korshunov and Sebastien Marcel "DeepFakes: a New Threat to Face Recognition? Assessment and Detection' arXiv:1812.08685v1 [cs.CV] 20 Dec 2018.

[2] Ren-Hung Hwang, Jia-You Lin, Sun-Ying Hsieh, Hsuan-Yu Lin, Chia-Liang Lin, "Adversarial Patch Attacks on Deep-Learning-Based Face Recognition Systems Using Generative Adversarial Networks", MDPI, Sensors, 2023.

[3] A. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," ICCV 2019.

[4] Y. Dong et al., "Efficient Decision-Based Black-Box Adversarial Attacks on Face Recognition," CVPR 2019.
[5] M. Sharif et al., "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition," CCS 2016.
[6] S. Agarwal et al., "Am I a Real or Fake Celebrity? Evaluating Face Recognition and Deepfake Impersonation" (ACM study evaluating black-box APIs).
[7] T. Brown et al., "Adversarial Patch," arXiv 2017 (foundational patch methodology).
[8] R. Hwang et al., "Adversarial Patch Attacks on Deep-Learning-Based Face Recognition" (2023 / overview of patch attacks on face recognition).
[9] Systematic review on face Presentation Attack Detection (PAD), PMCID: review article summarizing PAD methods and gaps.
[10] NIST guidance and recent notes on morphs/detection (2025 update).

AUTHORS

Dr. Suman Kumar Mishra, with 19 years of teaching experience, is currently serving as HoD of the CSE Department at Khwaja Moinuddin Chishti Language University. He specializes in Clustering, Artificial Intelligence, Database Management System, Object-Oriented Analysis & Design, and also teaches Software Engineering, Patent studies, and research methodology. He has guided numerous student projects, contributed to curriculum design, and promoted industry-oriented problem solving. Author of a Hindi book on Artificial Intelligence released by the Honorable Prime Minister of India, he remains committed to fostering innovation and inspiring students in research and technology.

Dr. Shan-e-Fatima, with 13 years of teaching experience, is currently serving as Assistant Professor of the CSE Department at Khwaja Moinuddin Chishti Language University. She specializes in Machine Learning, Computer Networks and also teaches Software Engineering and research methodology. She has facilitated MoUs with leading national and international companies to enhance industry collaboration and skill development. Her research focuses on Speech Processing, Deep Learning, aiming to apply emerging technologies for real-world solutions.

Er. Digesh Pandey, with 03 years of teaching experience, is currently serving as Assistant Professor of the CSE Department at Khwaja Moinuddin Chishti Language University. He specializes in Artificial Intelligence, Machine Learning and also teaches Design and Analysis of Algorithms, Database Management System, and research methodology. His research interests include Database Management Systems (DBMS), Data Science, and Machine Learning, with a focus on innovative applications and problem-solving in real-world domains.