Evaluating the Adversarial Robustness of Deepfake Detectors

Mihai-George STURZA, Daria-Maria PREDA
Bucharest University of Economic Studies
sturzamihai21@stud.ase.ro, daria.preda@csie.ase.ro

As eKYC pipelines see increasingly more usage in verifying identities across banking, insurance and retail applications, bad actors are developing new ways of bypassing validation and gaining trusted access in protected environments. Deepfake detectors have their unique place in such verification pipelines and act as a defense against synthetic forgeries delivered through injection attacks. This paper evaluates the adversarial robustness of four architecturally diverse detectors under white-box and black-box attacks in a cross-domain setting. Under white-box conditions all four detectors are fully compromised, even at perturbation magnitudes that remain metrically imperceptible. Transfer attacks show that adversarial examples crafted against a single, freely available, model can evade other detectors with high reliability and query-based attacks achieve comparable results without any knowledge of model internals. These results indicate that evaluated deepfake detectors do not withstand adversarial manipulation under realistic attack conditions, raising practical concerns for production eKYC deployments and for compliance with the EU AI Act's robustness requirement for high-risk biometric systems.

Keywords: *Deepfake detection, eKYC, Biometric security, Adversarial testing*
DOI: 10.24818/issn14531305/30.2.2026.04

1 Introduction

Generative models can now synthesize facial images realistic enough to pass human inspection, commonly referred to as deepfakes. These manipulations span several categories: identity swap, where the identity of one person is replaced with another in each video or image; face reenactment, where the expressions and movements of a source face are transferred onto a target; entire face synthesis, where entirely fictional yet photorealistic faces are generated [1]. Early deepfake generation methods produced visible artifacts that could be identified through manual inspection, but modern approaches based on Generative Adversarial Networks (GAN) and diffusion models have reached a level of fidelity where distinguishing synthetic from authentic content is no longer reliably possible by human observers [2]. Compounding this issue is the increasing accessibility of deepfake creation tools, ranging from open-source frameworks such as DeepFaceLab [3] to consumer-grade platforms that require minimal expertise.

Identity verification is among the domains most exposed to deepfake threats. Electronic Know Your Customer (eKYC) processes have become the standard method for remote identity onboarding across banking, insurance and fintech platforms, where submitted media is matched against identity documents on a server. While this model enables scalable onboarding, it introduces risk by trusting the client-submitted media to be genuine. Injection attacks exploit this assumption by substituting legitimate media streams with crafted content, as detailed in Section 3 [4].

The scale of this threat has escalated in recent years and has forced identity verification providers to adapt. According to the Entrust 2026 Identity Fraud Report, based on over one billion identity verifications across 195 countries, deepfakes account for one in five biometric fraud attempts, with injection attacks growing 40% annually [5]. Furthermore, the emergence of Deepfake-as-a-Service marketplaces has reduced the cost of producing synthetic identity to as little as 10 to 50 USD, making these attacks accessible even to low-skilled actors [6].

To counteract the growing threat of deepfake-based injection attacks, deepfake detection models have been integrated as a server-side

defense layer within eKYC pipelines. These detectors receive the submitted image or video frame and classify it as either authentic or manipulated by identifying learned artifacts that distinguish generated content from authentic captures. Over the past several years, the detection landscape has diversified considerably in terms of architectural approaches. Convolutional Neural Networks (CNN) based methods such as Xception [7] and EfficientNet-B7 [8] learn spatial features and compression artifacts through convolutional layers. More recently, detectors such as Uncovering Common Features (UCF) [9] generalize better across datasets by modeling global spatial relationships rather than local pixel-level patterns. Alternative paradigms, such as Reconstruction Classification Learning (RECCE) [10], frame detection as a reconstruction task, based on the observation that authentic faces can be reconstructed more faithfully than manipulated ones, while frequency-domain methods such as F3-Net [11] analyze spectral inconsistencies introduced during the synthesis process. DeepfakeBench [12], a unified evaluation platform encompassing 36 detection methods across 9 datasets, has established consistent baselines showing AUC scores above 0.95 for leading detectors on widely used FaceForensics++ dataset. From this perspective, deepfake detection appears to be tractable.

However, the evaluations underlying these performance figures share a common assumption: that the attacker submits an unmodified deepfake. An attacker operating through injection attacks has full control over the submitted input and faces no physical constraints, unlike presentation attacks, where environmental factors degrade the spoofing attempt. Such bad actors can therefore iteratively refine the payload before submission, including the application of adversarial perturbations [13]. The result could be a deepfake image that remains visually identical to the original yet classified as authentic by the detector. Prior work has shown that imperceptible perturbations can reduce a forensic classifier's AUC from 0.95 to near zero in a white-box setting and to 0.22 under black-box conditions [14].

Existing benchmarks, including DeepfakeBench, evaluate accuracy, cross-dataset generalization and compression robustness, but not resilience to adversarial manipulation. A detector that achieves 0.98 AUC on FaceForensics++ may offer a false assurance if an attacker can reduce the figure to near chance level with an imperceptible perturbation. In the context of eKYC, where a single successful bypass can result in fraudulent account creation, unauthorized financial transactions or identity theft, this problem carries direct economic and regulatory consequences. The EU AI Act classifies biometric verification systems as high-risk and explicitly requires resilience against adversarial examples under Article 15 [15].

This paper conducts an adversarial robustness evaluation of deepfake detectors, framed explicitly within the security context of eKYC pipelines. Our research question is: how resilient are current deepfake detection models against standard adversarial attacks and what are the practical security implications for identity verification systems that rely on them? To answer this, we evaluate four detectors representing distinct architectural paradigms against white-box attacks as well as black-box transfer attacks.

2 Related Work

2.1. Deepfake Generation and Detection

Several benchmark datasets have been developed to evaluate deepfake detection methods across these manipulation categories. FaceForensics++ applies four manipulation techniques: Deepfakes, Face2Face, FaceSwap and NeuralTextures to 1000 source videos, providing evaluation protocols at multiple compression levels. However, early datasets exhibited visible artifacts that made detection relatively straightforward. Celeb-DF v2 [16] was introduced to address this limitation, featuring higher synthesis quality and fewer visible artifacts. This makes it a harder benchmark, relevant for eKYC scenarios where attackers are motivated to produce the highest quality forgeries possible.

Early detection approaches targeted specific

physiological inconsistencies in manipulated content, such as irregular eye blinking patterns or abnormal head pose distributions [17]. While effective against first-generation deepfakes, these methods did not generalize as synthesis quality improved. Deep learning detection models began with Rössler et al. [1] who demonstrated that the Xception architecture, adapted via transfer learning, could learn spatial and compression-level artifacts directly from data. EfficientNet-b4 subsequently emerged as a leading detector in the Deepfake Detection Challenge [18], achieving strong accuracy through compound scaling of network depth, width and resolution.

More recently, the detection approaches have diversified architecturally. UCF [9] is focusing on identify common forgery features that generalize across manipulation methods and datasets, rather than learning artifacts specific to a single forgery type. RECCE [10] takes a different approach, framing detection as a reconstruction task: the model learns to reconstruct input faces, with manipulated content producing a higher reconstruction error. This diversity matters for our analysis because different paradigms may respond differently to adversarial perturbations.

To address fragmented evaluation protocols across the literature, Yan et al [12] introduced DeepfakeBench, a unified platform integrating 36 detectors across 9 datasets with standardized preprocessing and training pipelines. We use DeepfakeBench pre-trained weights for all four detectors in our experiments.

2.2. Adversarial Attacks on Deep Neural Networks

Adversarial examples are inputs to a neural network that have been deliberately modified with imperceptible perturbations designed to cause misclassification [19]. For image classifiers, this involves adding a bounded noise pattern to the pixel values such that the resulting image appears visually unchanged to a human observer but produces a different model output. The magnitude of the perturbation is commonly measured under the $L\infty$ norm, which constrains the maximum change to any individual pixel, or the L2 norm, which constrains the overall Euclidean distance between the original and perturbed image. Adversarial vulnerability is a general property of deep neural networks, not specific to any architecture, and has been observed across virtually all deployment domains.

Several attack algorithms have been proposed, varying in computational cost, effectiveness and assumptions about attacker knowledge. The Fast Gradient Sign Method (FGSM) [13] computes a single gradient step in the direction that maximizes the classification loss, producing adversarial examples with limited optimization. Projected Gradient Descent (PGD) [20] extends FGSM through multiple iterative steps, projecting the perturbation back onto the allowed ε-ball after each update, and is widely regarded as the strongest first-order attack. The Carlini and Wagner (C&W) attack [14] formulates adversarial example generation as an optimization problem that directly minimizes perturbation magnitude while ensuring misclassification, typically under the L2 norm. These three methods operate in a white-box setting, where the attacker has full access to the model architecture and parameters.

In practice, full model access is almost never available. Two strategies address this constraint. Transfer attacks exploit the empirical observation that adversarial examples crafted against one model frequently cause misclassification in other models, even when architectures differ. This property is concerning in security contexts, as an attacker can train or obtain a surrogate detector and generate perturbations that transfer to the unknown deployed model. Query-based black-box attacks take a different approach by repeatedly querying the target model and observing its output and the Square Attack [21] is a representative method in this category, using randomized square-shaped perturbations and a scored search strategy that does not require gradient computation. The studies reviewed above collectively establish that adversarial vulnerability is a real concern for deepfake detection systems. We aim to build on the existing body of work by extending the analysis to a broader set of modern detector architecture, evaluated under a

unified experimental protocol across white-box and black-box attack scenarios.

3 Threat Model

In a typical remote identity verification flow, a user captures a selfie or short video on a mobile device, which is transmitted over a network to a server-side verification pipeline. The server performs a sequence of checks: face detection and alignment, liveness verification, deepfake detection and face matching against a reference identity document. The final access decision depends on all checks passing. This architecture places the deepfake detector as a primary defense against synthetic media that has already bypassed the device-level capture. Importantly, the server receives only the final image or video payload. It has no direct guarantee that the media originated from a physical camera rather than from a virtual camera software, a device emulator or a crafted API request (if not using some form of signatures for requests). This is the assumption that injection attacks exploit.

We consider an attacker whose objective is to pass the identity verification process using a deepfake of a target victim. The attacker generates a face swap using a set of tools, producing a synthetic set of images or videos that resembles the victim. Before submitting this content to the eKYC pipeline, the attacker applies adversarial perturbations specifically optimized to cause mislabeling of generated content. The submission is performed through an injection vector described above which gives the attacker total control over the input and removes any physical-world constraints on the perturbation.

Attack effectiveness depends on how much the attacker knows about the target detector. We consider two scenarios that reflect a spectrum of lab-created and realistic capabilities. In the white-box setting, the attacker has full access to the detector's architecture and trained weights. In a black box setting, the attacker has access to a different detector, potentially open-source or based on SotA and generates adversarial examples against it, relying on cross-model transferability to fool the unknown deployed detector.

Regardless of the knowledge level, the attacker operates under the constraint that the adversarial perturbations must remain imperceptible. A visibly corrupted image would either raise suspicion during manual review or be rejected by upstream quality checks in the pipeline. An attack is considered successful when a deepfake image that would have been correctly rejected by the detector is instead classified as authentic after perturbation.

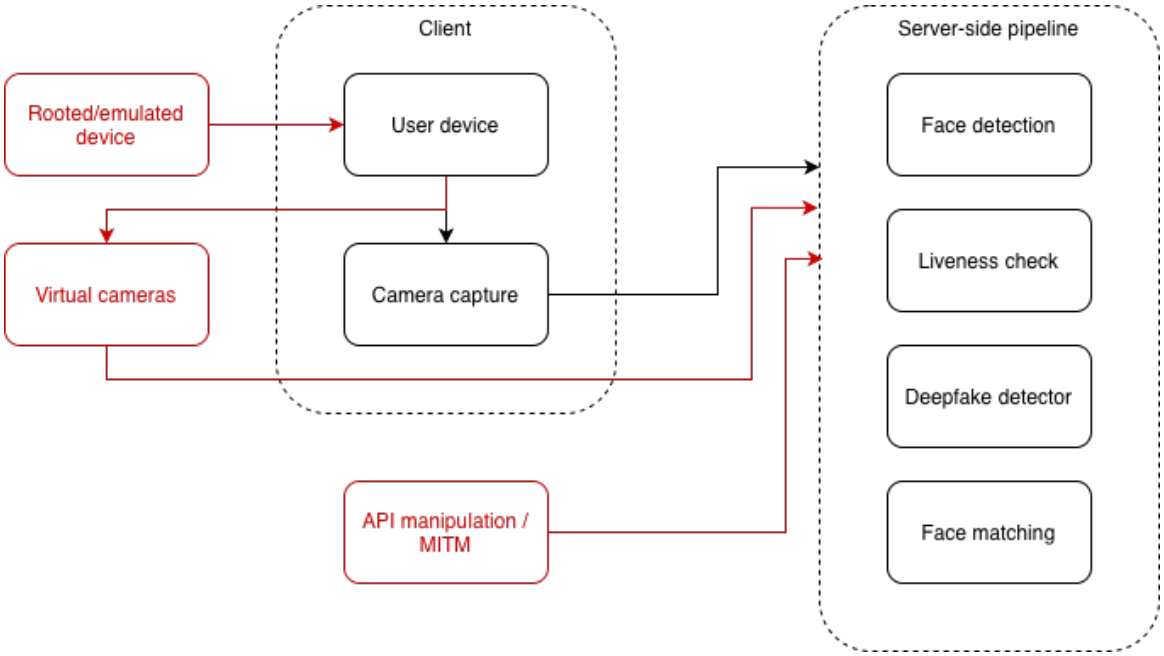

Fig. 1. eKYC pipeline with attack vectors

4 Methodology

4.1. Dataset and Preprocessing

The evaluation is conducted on the test split of Celeb-DF v2, as defined by the provided List_of_testing_videos.txt partition file provided with the dataset. The test split comprises both real and synthesized celebrity face videos. Frames are extracted from each test video and faces cropped and aligned following the standard DeepfakeBench processing pipeline. The label convention follows the DeepfakeBench standard, where 0 denotes a real face and 1 denotes a manipulated face. In total, the test set contains 16420 frames. All frames are resized to 256x256 pixels and normalized using a mean and standard deviation of 0.5 per channel. Due to the computational cost of C&W and Square Attack, these are evaluated on a stratified subset maintaining equal class balance when sample limits are specified, while all other attacks are evaluated on the full test set.

4.2. Detectors Under Evaluation

Four deepfake detection models are selected for evaluation, chosen to represent the principal architectural paradigms in current use.
All models are loaded with their respective pre-trained weights from DeepfakeBench, trained on FaceForensics++ with c23 compression. No architectural modifications or additional fine-tuning are applied, and the detectors are evaluated as distributed.

Table 1. Evaluated deepfake detectors

Detector	Backbone	Parameters	Approach
Xception	Xception	22.8M	CNN with depthwise separable convolutions
EfficientNet-b4	EfficientNet-b4	19.3M	CNN with compound scaling
UCF	Xception	46.8M	Multi-task CNN
RECCE	Custom encoder/decoder	47.7M	CNN with reconstruction model

4.3. Attack Configuration

The adversarial evaluation covers both white-box and black-box scenarios, as defined in the threat model.
FGSM is evaluated at four perturbation magnitudes $\varepsilon \in \{0.05 / 255, 0.1 / 255, 2 / 255, 8 / 255\}$. The inclusion of very small budgets (0.05 and 0.1) serves to identify the threshold at which adversarial perturbations begin to meaningfully affect detection while 8/255 represents a standard upper bound commonly used in the literature. PGD is configured with $\varepsilon = 8/255$, step size $\alpha = 2/255$ and 50 iterations at the highest perturbation budget, serving as the strongest first-order white-box attack in our evaluation. The C&W attack operates under the L2 norm and uses 1000 optimization steps. As a sanity check, we also evaluate random uniform noise at $\varepsilon = 0.1/255$ to confirm that arbitrary noise of comparable magnitude does not degrade detector performance, isolating the effect of adversarial optimization from generic noise sensitivity.
For the black-box evaluation, the Square Attack is configured with 5000 queries per image, simulating an attacker who can repeatedly probe the verification endpoint, in what would probably represent a testing environment in a practical example. The transfer attack evaluation generates PGD adversarial examples from each of the four detectors and evaluates them against all others, producing a transfer matrix. The design captures both within-family transfers and cross-family transfers.

4.4. Evaluation Metrics

Detection performance is measured using the Area Under the Receiver Operating Characteristic Curve (AUC), which evaluates classification quality across all decision thresholds. A high AUC indicates better separation between real and fake inputs. Since eKYC systems process video sequences rather than isolated frames, we also compute video-level AUC by averaging the per-frame fake-class

probabilities within each video and computing AUC over these aggregated scores.

The effectiveness of adversarial attacks is measured using the Attack Success Rate (ASR), defined as the proportion of fake samples that were correctly classified under clean conditions but are misclassified as real after adversarial perturbation. An ASR of 1.0 indicates that the attack flipped the model's decision on every sample, while 0.0 indicates no effect.

To quantify whether adversarial perturbations remain imperceptible, we report four complementary measures computed between the original and perturbed fake images. Peak Signal-to-Noise Ratio (PSNR) expresses the perturbation magnitude relative to the image's dynamic range, with higher values indicating less visible modifications (values above 40dB are generally considered imperceptible). Structural Similarity Index (SSIM) [22] captures perceptual degradation in terms of luminance, contrast and structural information with values close to 1.0 indicating near-identical images. Finally, Learned Perceptual Image Patch Similarity (LPIPS) [23] uses deep features from a pre-trained AlexNet to measure perceptual distance as perceived by a neural network, with lower values indicating greater similarity.

5 Results

5.1. Baseline Performance

The results closely match the official cross-dataset evaluation scores reported by DeepfakeBench. All four detectors achieve AUC scores between 0.74 and 0.77 on Celeb-DF v2. These values are notably lower than the in-domain performance reported on FaceForensics++ c23, where the same models achieve AUC scores above 0.96. Cross-dataset setting more accurately represents the conditions under which these models would operate in a production environment. UCF achieves the highest discriminative performance on both metrics.

Table 2. Baseline detection performance on Celeb-DF v2

Detector	AUC	AUC (DeepfakeBench)	Video AUC
Xception	0.7403	0.7365	0.8164
EfficientNet-b4	0.7505	0.7487	0.8081
UCF	0.7717	0.7527	0.8374
RECCE	0.7411	0.7319	0.8232

5.2. White-box Attack Results

Table 3 presents the entire white-box evaluation results across all detector combinations. The most immediate observation is the contrast between random noise and adversarial perturbation. At ε=0.1/255, random uniform noise produces ASR below 0.4% across all detectors yet FGSM at half that budget achieves ASR between 71% and 91%. Even a single-step attack at minimal perturbation budgets flips most correct decisions. The vulnerability is therefore not a matter of large, potentially visible distortions. Degradation accelerates rapidly with increasing perturbation budget. At FGSM ε=0.1/255, all detectors fall below 0.21 AUC, with RECCE exhibiting the steepest decline to 0.0894. By ε=2/255, frame-level AUC approaches zero for Xception and remains below 0.11 for all models, with ASR exceeding 99.3%.

Table 3. White-box attack results on Celeb-DF v2

Attack types	Metrics	Xception	EfficientNet-b4	UCF	RECCE
Random-0.1	AUC	0.7403	0.7505	0.7718	0.7416
	ASR	0.0024	0.0015	0.0034	0.0029
FGSM-0.05	AUC	0.3124	0.432	0.2993	0.2676
	ASR	0.7887	0.7104	0.8604	0.9156
FGSM-0.1	AUC	0.1078	0.2042	0.1145	0.0894

	ASR	0.9694	0.9505	0.9444	0.996
FGSM-2	AUC	0.022	0.0729	0.0549	0.1077
	ASR	0.9996	0.9936	0.974	0.9979
FGSM-8	AUC	0.5895	0.4558	0.7296	0.52
	ASR	0.9967	0.9276	0.9718	0.9996
PGD-50	AUC	0	0	0	0
	ASR	1	1	1	1
CW-L2	AUC	0.365851	0.4944	0.6328	0.5514
	ASR	1	1	0.9774	1

At ε=8/255, however, FGSM exhibits an apparent recovery as AUC rises to 0.59 for Xception and 0.73 for UCF, despite ASR remaining above 92% across all detectors. This divergence between metric behaviors reveals a known limitation of single-step gradient methods: overshooting. At ε = 8/255, the large single FGSM step violates the local linearity of the loss landscape, pushing prediction scores of the manipulated images past the optimal evasion minimum. Because these overshot scores still comfortably cross the binary classification threshold, the ASR remains exceptionally high. However, since AUC measures relative ranking rather than a fixed threshold crossing, the model is still able to rank these overshot fakes slightly higher than unperturbed real images and it artificially inflates the AUC, giving a false impression of model resilience when the detector's operational threshold has been bypassed.

PGD-50, configured at the same perturbation budget (ε=8/255), eliminates this limitation entirely as it iteratively updates the projection onto the ε-ball, preventing overshooting. Over 50 iterative gradient steps, the attack reduces the model ability to classify fake samples completely.

The C&W attack, operating under L2 minimization with 1000 optimization steps, achieves ASR between 97.7% and 100% while maintaining lower perturbation magnitudes as seen at Section 5.4. This makes C&W the most efficient attack in terms of the trade-off between evasion success and perturbation imperceptibility.

Across architectures, no detector demonstrates meaningful adversarial resilience. UCF is the only model where C&W does not achieve complete evasion, with marginal differences that disappear entirely under PGD-50. RECCE's reconstruction-based approach shows no inherent advantage; it is in fact the most sensitive to low-epsilon FGSM, suggesting that alternative learning objectives do not confer resilience against gradient-based attacks.

5.3. Black-box Attack Results

Table 4. Transfer video AUC (rows = source, columns = target, PGD-50 at ε=8/255)

	Xception	**EfficientNet-b4**	**UCF**	**RECCE**
Xception	0.0000	0.2926	0.0174	0.0000
EfficientNet-b4	0.4933	0.0000	0.2225	0.5044
UCF	0.0244	0.7123	0.0000	0.0007
RECCE	0.0000	0.5636	0.0000	0.0000

Table 5. Transfer ASR (rows = source, columns = target, PGD-50 at ε=8/255)

	Xception	**EfficientNet-b4**	**UCF**	**RECCE**
Xception	1.0000	0.9058	1.0000	1.0000
EfficientNet-b4	0.9226	1.0000	0.9958	0.8747
UCF	0.9969	0.4098	1.0000	1.0000
RECCE	1.0000	0.6297	1.0000	1.0000

Transfer attacks are made with adversarial examples generated using PGD-50 (ε=8/255) against each source detector and evaluated on all target detectors. Results reveal that Xception serves as the most effective surrogate as adversarial examples generated against it achieve 90% to 100% ASR on other detectors, with video AUC dropping to near zero on three of the four targets. This is significant given that Xception is freely available, widely documented and among the most used deepfake detectors, making it an obvious starting point for an attacker.

EfficientNet-b4 stands out as the most transfer resistant target. When attacking using surrogates other than itself, it receives ASR of 41% to 90% and its video AUC remains at 0.29 to 0.71 (degraded but not fully collapsed). This partial resilience likely reflects architectural differences in how EfficientNet-b4 processes features compared to the other three detectors, resulting in lower adversarial transferability.

UCF and RECCE, despite their distinct designs, are vulnerable to transfer. The reconstruction inductive bias of RECCE provides no meaningful protection against transferred adversarial perturbations, reinforcing the findings from the white-box attacks that architecture diversity does not offer adversarial robustness.

Table 6. Square attack results

Detector	AUC	Video AUC	ASR
Xception	0.4791	0.4554	0.9991
EfficientNet-b4	0.5564	0.5531	1.0000
UCF	0.4983	0.3861	1.0000
RECCE	0.5426	0.4881	1.0000

Without any knowledge of model internals, the Square attack achieves ASR between 99% and 100% on all four detectors. These results are achieved using only output scores and the attack does not require model gradients, architecture information or training data. In an eKYC context, this corresponds to an attacker who can repeatedly submit verification attempts and observe accept/reject decisions, a capability that is preventable in production systems.

5.4. Perturbation Imperceptibility

We have established that adversarial perturbations should remain imperceptible as a visibly corrupted image would be rejected by upstream quality checks or flagged during manual reviews. The perturbation magnitudes at which detection collapses are very small. Lower epsilon FGSM produces perturbations metrically indistinguishable from the unperturbed image.

Table 7. Perturbation quality metrics (Xception)

Attack	L2	PSNR	SSIM	LPIPS
FGSM-0.05	0.087	74.20	0.9999	0.0000
FGSM-0.1	0.173	68.18	0.9999	0.0000
FGSM-2	3.458	42.16	0.9703	0.0340
FGSM-8	13.806	30.13	0.7029	0.4072
PGD-50	10.09	33.04	0.8147	0.2424
CW-L2	0.382	68.35	0.9994	0.0003
Square	11.21	43.45	0.7876	0.2057

At these perturbation levels the ASR obtained compared to the effect it has on the perturbed image illustrated that this vulnerability is not a function of magnitude but of perturbation direction. Notably, The C&W attack is operating under L2 minimization, and it achieves

97% to 100% ASR while maintaining PSNR between 68.35 for the Xception model.

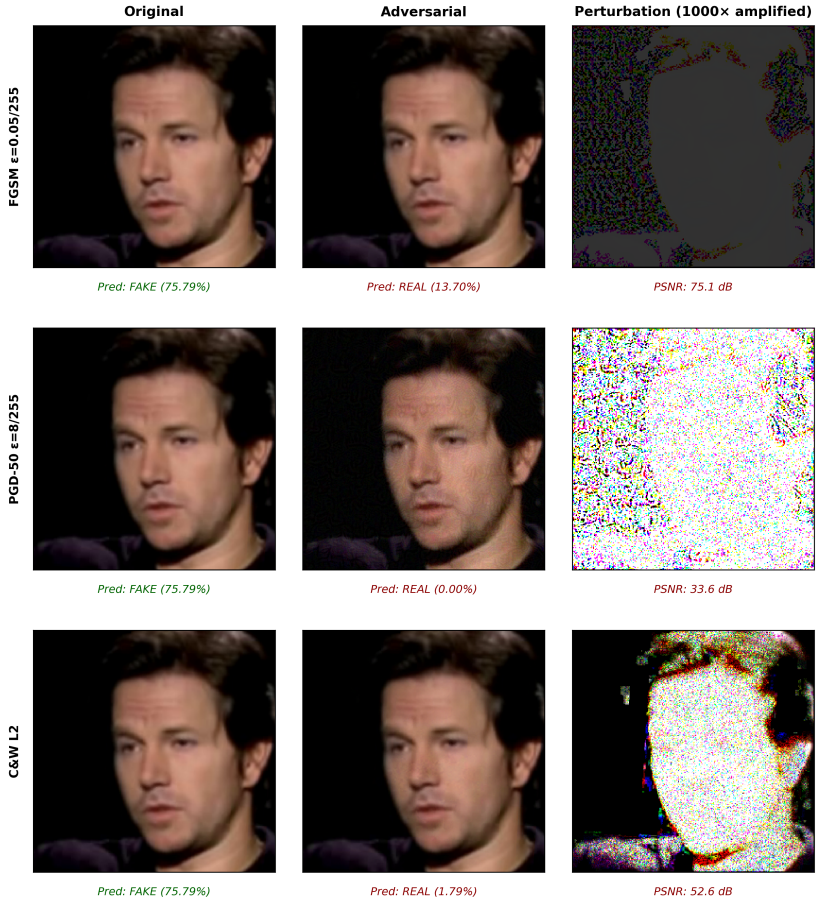

Fig. 2. Adversarial perturbation examples on Xception

6 Conclusions

6.1. Security and Architectural Implications

The results show that current deepfake detectors, regardless of architectural paradigm, offer no meaningful resistance to adversarial perturbation. Under white-box PGD, all four detectors were completely compromised, and this vulnerability was seen to be transferable across different model architectures. These findings gain weight when considered alongside the operational context established in Section 1 that injection attacks are the fastest growing threat vector in eKYC. The deepfake detector, positioned as the primary server-side defense against such attacks, can be bypassed.

6.2. Regulatory Compliance

The EU AI Act classifies biometric verification systems as high-risk and requires, under Article 15, that such systems achieve appropriate levels of robustness including explicit resilience against adversarial examples. However, the primary technical standard governing biometric attack detection, the ISO/IEC 30107 focuses on presentation attacks such as printed photos, masks and screen replays, with no provisions for adversarial perturbation attacks. In practice, a biometric system that satisfies ISO 30107 testing may nonetheless be completely ineffective against an adversary who applies adversarial optimization to their payload.

6.3. Limitations

Several limitations should be considered when interpreting these results. All experiments use Celeb-DF v2 in a cross-dataset setting. Results may differ on other datasets or under in-domain training conditions. Our evaluation operates at the frame level and video level temporal defenses that exploit inter-frame consistency are not considered.

A practical limitation must be noted regarding

the high ASR observed at sub-pixel perturbation budgets (e.g., $\varepsilon = 0.05/255$ and $0.1/255$). The experimental evaluations are computed in a continuous floating-point space, where an ε of $0.05/255$ represents a fractional change equivalent to just $1/20^{th}$ of a single 8-bit pixel intensity value. In the context of our established eKYC injection threat model, a real-world attacker must submit their payload via methods that ingest discrete 8-bit image formats (e.g. PNG, JPEG). If such small perturbations are serialized and quantized back to standard 8-bit integer space, the adversarial noise is rounded to zero, which would effectively destroy the attack payload. A practical injection attack would require a perturbation budget of at least $\varepsilon \geq 1/255$ to survive image serialization and network transmission.
Finally, we do not evaluate defensive countermeasures, and the focus of this work is quantifying vulnerability.

6.4. Future Directions and Conclusions

A natural extension of this work is to investigate whether adversarial training can restore robustness without significantly degrading clean detection performance. Preliminary results in the broader adversarial robustness literature suggest that this kind of training introduces inherent accuracy tradeoffs [24].
This paper has presented an adversarial robustness evaluation of four architecturally diverse deepfake detectors in the context of eKYC security. The complete implementation is publicly available at https://github.com/sturzamihai/eardd.

References

[1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Nießner, FaceForensics++: Learning to Detect Manipulated Facial Images, 2019.
[2] S. Baliah, Q. Lin, S. Liao, X. Liang and M. H. Khan, Realistic and Efficient Face Swapping: A Unified Approach with Diffusion Models, arXiv, 2024.
[3] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, Dpfks, C. S. Facenheim, L. RP, J. Jiang, S. Zhang, P. Wu, B. Zhou and W. Zhang, DeepFaceLab: Integrated, flexible and extensible face-swapping framework, arXiv, 2021.
[4] H. Felouat, H. H. Nguyen, T.-N. Le, J. Yamagishi and I. Echizen, "eKYC-DF: A Large-Scale Deepfake Dataset for Developing and Evaluating eKYC Systems," IEEE Access, vol. 12, pp. 30876-30892, 2024.
[5] Entrust, "IDENTITY FRAUD REPORT," 2026.
[6] Group-IB, "Weaponized AI: Inside The Criminal Ecosystem Fueling The Fifth Wave of Cybercrime," 2026.
[7] F. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, 2017.
[8] M. Tan and Q. V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, arXiv, 2020.
[9] Z. Yan, Y. Zhang, Y. Fan and B. Wu, UCF: Uncovering Common Features for Generalizable Deepfake Detection, arXiv, 2023.
[10] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding and X. Yang, "End-to-End Reconstruction-Classification Learning for Face Forgery Detection," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
[11] Y. Qian, G. Yin, L. Sheng, Z. Chen and J. Shao, Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues, arXiv, 2020.
[12] Z. Yan, Y. Zhang, X. Yuan, S. Lyu and B. Wu, DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection, arXiv, 2023.
[13] I. J. Goodfellow, J. Shlens and C. Szegedy, Explaining and Harnessing Adversarial Examples, arXiv, 2015.
[14] N. Carlini and D. Wagner, Towards Evaluating the Robustness of Neural Networks, arXiv, 2017.
[15] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, pp. L 2024/1689, Art. 15, 2024.
[16] L. Yuezun, Y. Xin, S. Pu, Q. Honggang

and L. Siwei, Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics, 2020.
[17] Y. Li and S. Lyu, Exposing DeepFake Videos By Detecting Face Warping Artifacts, arXiv, 2019.
[18] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang and C. Canton Ferrer, The DeepFake Detection Challenge (DFDC) Dataset, arXiv, 2020.
[19] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, Intriguing Properties of Neural Networks, arXiv, 2014.
[20] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, Towards Deep Learning Models Resistant to Adversarial Attacks, arXiv, 2019.
[21] M. Andriushchenko, F. Croce, N. Flammarion and M. Hein, Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search, arXiv, 2020.
[22] Z. Wang, A. Bovik, H. Sheikh and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Transactions on Image Processing, vol. 13, pp. 600-612, 2004.
[23] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, arXiv, 2018.
[24] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner and A. Madry, Robustness May Be at Odds with Accuracy, arXiv, 2019.
[25] H. Kim, Torchattacks: A PyTorch Repository for Adversarial Attacks, arXiv, 2020.

Mihai-George STURZA has graduated from the Faculty of Cybernetics, Statistics and Economic Informatics at the Bucharest University of Economic Studies and is currently pursuing a master's degree in IT&C Security at the same university. His research interests include artificial intelligence, with a focus on computer vision, natural language processing and large language models, as well as adversarial robustness and biometric security. He also runs a software engineering company, offering AI-driven solutions, including computer vision, speech recognition and natural language processing for business applications.

Daria-Maria PREDA has graduated from the Faculty of Economic Cybernetics, Statistics and Informatics at the Bucharest University of Economic Studies, earning a Bachelor's degree in 2023 and a Master’s degree in 2025 in the field of Economic Informatics. She began her professional career in 2022 as a Software Developer, focusing on building scalable and reliable software systems.