КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (34), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.34.1349
УДК 004.8:004.932

Nazarii O. Dzhaliuk
PhD Student,
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0009-0006-9381-416X
nazarii.o.dzhaliuk@lpnu.ua

Volodymyr V. Khoma
DSc, Professor
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0000-0001-9391-6525
volodymyr.v.khoma@lpnu.ua

Yurii V. Khoma
DSc, Associate Professor
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0000-0002-4677-5392
yurii.v.khoma@lpnu.ua

COMPARATIVE EVALUATION OF DEEPFAKE DETECTION MODELS FOR SECURE REMOTE BIOMETRIC VERIFICATION

Abstract. This study examines how remote biometric verification systems hold up against next-generation video deepfakes. Financial and digital platforms increasingly use video-based onboarding, but these pipelines are vulnerable to image-to-video animation, which generates video from a single static image without leaving the spatial splicing or blending artifacts that older detectors look for. To evaluate this threat, we developed a balanced dataset of 51 videos, consisting of 26 authentic videos and 25 synthetic videos generated via face-merging and subsequent image-to-video animation. We benchmarked two commercial SaaS platforms (TruthScan and Hive Moderation) against six open-source deep learning models: EfficientNet-B7, a CLIP-based detector, GANomaly, two Vision Transformers, and GenConViT. Classical convolutional networks such as EfficientNet-B7 did not generalize to this synthesis method: recall fell to 0.240. The hybrid GenConViT model gave the most balanced performance of the open-source systems, with an accuracy of 0.843, an F1-score of 0.825 and a specificity of 0.923. That accuracy is close to the commercial platforms, which scored between 0.882 and 0.901.
These results indicate that hybrid convolutional-transformer architectures generalize to boundary-free synthesis in a way that purely convolutional detectors do not, and that an open-source model can approach commercial detection quality without transmitting biometric data to a third party.4

Keywords: deepfake, deepfake detection, image-to-video animation, GenConViT, Swin Transformer, model benchmarking, biometric verification, cybersecurity

INTRODUCTION

Remote video verification is now a standard identity check in digital banking and customer onboarding [11]. As financial institutions move away from in-branch verification, they depend on automated biometric systems to authenticate users. The spread of synthetic media undermines that dependence [1]. Tools that produce realistic artificial video of a specific person create a vulnerability in remote authentication, so existing defences have to be re-tested as those tools improve.

The volume of synthetic media has grown as generative engines became easier to access. DeepStrike reports that the number of online deepfakes rose from roughly 500,000 in 2023 to nearly 8 million in 2025, a 16-fold increase and an annual growth rate of about 900% [17].

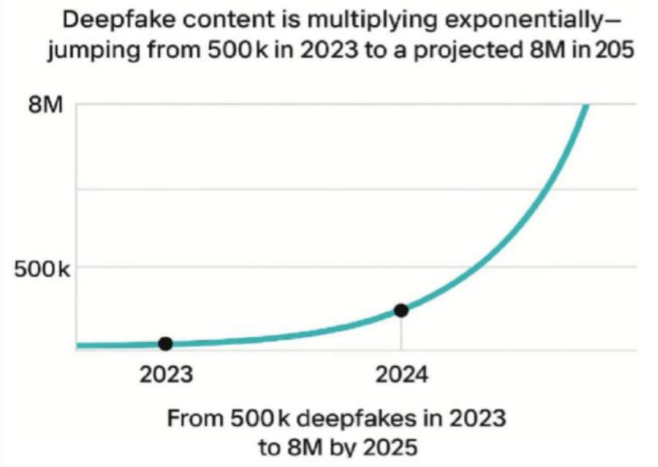

Figure 1. Deepfake volume growth curve

Beyond the direct threat of biometric spoofing, this proliferation accelerates the "liar's dividend" phenomenon, wherein the pervasive nature of deepfake technology allows malicious actors to plausibly deny the authenticity of genuine, incriminating digital evidence [2].

Deepfakes have already been used as weapons in war, corporate fraud and elections. In information warfare, the March 2022 surrender-order deepfake of Ukrainian President Volodymyr Zelenskyy marked the first documented deployment of a weaponized deepfake during an active military conflict [5]. The video had visible flaws, including disproportionate head-to-body scaling, unnatural movement and boundary artifacts, but it showed that deepfakes had moved from a theoretical risk to a wartime tool [5]. In the corporate sector, threat actors have successfully executed highly coordinated financial fraud. A notable 2024 incident in Hong Kong resulted in a multinational firm transferring 25 million USD to attackers following a multi-person video conference where all participants, including the Chief Financial Officer, were real-time deepfakes [6]. This was a significant escalation from earlier social engineering exploits, such as the 2019 voice-cloning theft of 220,000 EUR [6]. Furthermore, electoral manipulation was observed during the 2023 Slovakian parliamentary election, where a fake audio leak of a candidate discussing vote buying was released less than 48 hours before voting, which left no time for forensic debunking [3].

These attacks expose a gap in traditional banking facial recognition systems. Legacy detection algorithms were predominantly engineered to identify structural anomalies associated with standard "face-swapping" techniques, which typically leave spatial seams, skin-tone mismatches, and blending artifacts near the facial boundaries [4, 10]. However, next-generation image-to-video animation methodologies bypass these heuristics entirely. By generating temporally coherent movements from a single static merged face, image-to-video animation produces continuous video sequences with no spatial splicing boundaries [7, 8, 9]. Classical convolutional detectors built only for spatial anomaly detection therefore do not generalize across domains, and remote banking identity pipelines are left exposed to synthetic injection attacks.

RELATED WORKS

The technological foundations of video deepfake creation rest primarily upon Generative Adversarial Networks (GANs) and Denoising Diffusion Probabilistic Models (DDPMs) [7, 9]. Proposed by Goodfellow et al., GANs operate via a minimax game between two competing neural networks: a generator G that captures the data distribution and a discriminator D that estimates the probability that a sample came from the training data rather than G [7]. Advances such as StyleGAN and StyleGAN2 refined this paradigm by introducing a style-based generator architecture where latent vectors are mapped to intermediate style spaces W and W+, enabling fine-grained control over facial attributes, lighting, and expressions [8]. In contrast, diffusion models, formalized by Ho et al., generate media via a two-step stochastic process: a forward process that gradually degrades image structure by adding Gaussian noise over time steps t in

{1, ..., T}, and a learned reverse process that iteratively eliminates noise to synthesize high-fidelity images from pure noise distributions [9]. Contemporary engines such as OpenAI's Sora, Runway Gen-3, SadTalker, and AnimateDiff use these diffusion mechanisms to achieve temporal consistency and realistic motion synthesis from static imagery [4, 9].

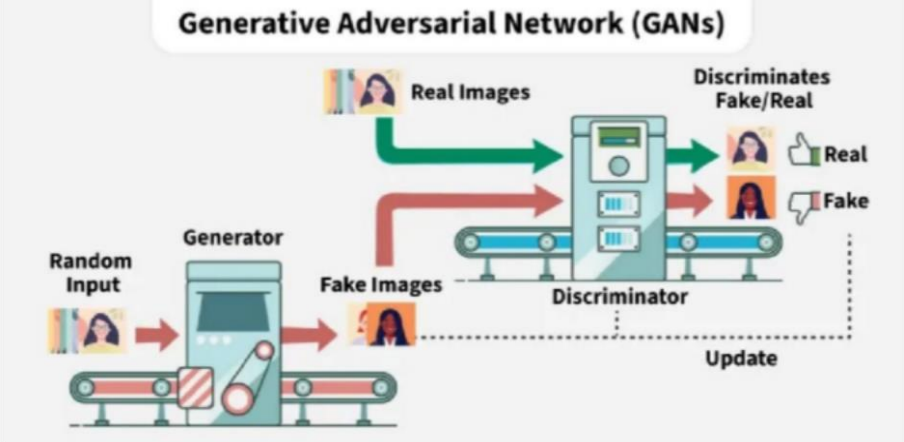

Figure 2. GAN architecture flowchart

Synthesized media falls into four facial forgery classes: face-swap, face-reenactment, lip-sync and image-to-video animation [4]. Face-swapping involves replacing the facial identity of a target individual in a video with that of a source subject while preserving the target's background and posture [4, 10]. Face-reenactment alters the expressions and head pose of a target subject to match the movements of a driving actor [4]. Lip-syncing modifies the lower facial region to synchronize mouth movements with an arbitrary acoustic driving track [4]. The hardest of the four, image-to-video animation, builds a temporally coherent video from a single photograph with no source-to-target splicing boundary [4, 9].

To counter these forgery classes, forensic detection methodologies have evolved across three primary analytical paradigms: spatial analysis, temporal analysis, and biological signal extraction [2, 4, 13]. Spatial detection focuses on identifying intra-frame artifacts, including abnormal skin textures, boundary blending seams, color space inconsistencies, and structural eye defects [4, 13]. Temporal detection evaluates frame sequences to capture inter-frame flickering, geometric facial deformations, and unnatural motion transitions over time [2, 4]. Biological signal extraction measures physiological indicators, such as remote photoplethysmography (rPPG) to detect cardiovascular blood volume pulses through micro-changes in facial skin coloration, or irregular eyeblink frequencies [2]. Classical passive forensics fails on modern deepfakes. This includes Error Level Analysis, PRNU sensor noise identification, JPEG blocking artifact quantification and EXIF metadata validation [12].

Generative models synthesize coherent pixel distributions without traditional editing footprints, while lossy video compression and social media re-encoding pipelines strip away high-frequency sensor noise signatures [13].

To evaluate detection performance against image-to-video animation, six representative open-source neural network architectures were selected for benchmarking. The first model, dfdc (selimsef), employs an EfficientNet-B7 backbone optimized with compound scaling across network depth, width, and input resolution (380 x 380 px) [14]. As the top-performing convolutional architecture in the 2020 Facebook DeepFake Detection Challenge (DFDC), it relies heavily on detecting localized spatial splicing anomalies learned from benchmarks such as FaceForensics++ [10, 14]. The second architecture, CLIP-based (yermandy), uses OpenAI's visual foundation encoder (CLIP ViT-L/14) fine-tuned with a lightweight classification head on FaceForensics++ [10, 15]. Grounded in the methodology of Yermakov et al., this model uses rich visual embeddings pre-trained on hundreds of millions of image-text pairs to identify distribution shifts characteristic of synthetic manipulation [15]. The third model, GANomaly, represents semi-supervised anomaly detection; developed by Akcay et al., it combines an encoder-decoder-encoder pipeline with adversarial learning, training exclusively on authentic facial samples to identify synthetics as statistical distribution deviations [18].

The remaining evaluated architectures incorporate Vision Transformers and hybrid multi¬modal frameworks. Models four and six, prithivMLmods v1 and Deep-Fake-Detector-v2, use Vision Transformer (ViT-base) backbones [19, 20] that partition facial crops into sequences of non-overlapping patches, employing multi-head self-attention mechanisms to model global long-range spatial dependencies across the facial plane rather than relying strictly on localized receptive fields. The fifth model, GenConViT, developed by Wodajo et al., employs a hybrid architecture that combines ConvNeXt blocks for multi-scale local feature extraction with Swin Transformer modules for shifted-window hierarchical self-attention [16]. Additionally, GenConViT incorporates parallel autoencoder (AE) and variational autoencoder (VAE) branches to perform latent-space distribution analysis, allowing it to detect both visible spatial defects and statistical generative anomalies [16].

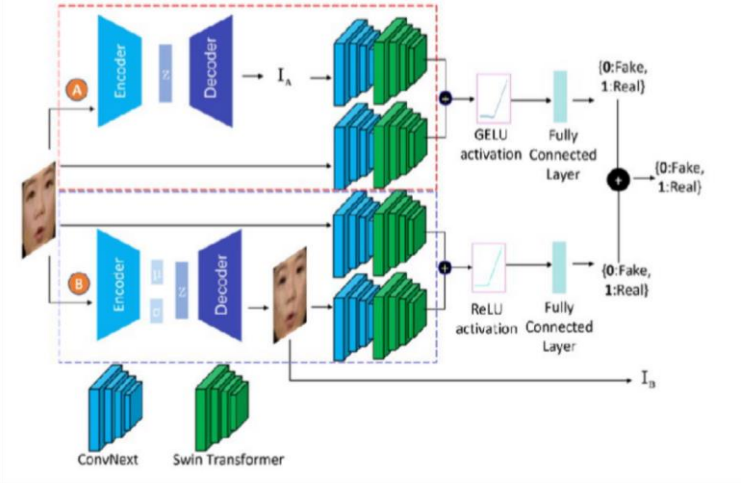

Figure 3. GenConViT schema

AIM OF RESEARCH

The objective of this research is to evaluate and benchmark the detection accuracy and cross-domain generalization capabilities of commercial SaaS solutions (TruthScan and Hive Moderation) and six open-source deep learning architectures (dfdc EfficientNet-B7, CLIP-based detector, GANomaly, prithivMLmods v1 ViT, GenConViT, and Deep-Fake-Detector-v2) against next-generation video deepfakes generated via face-merging and image-to-video animation. To achieve this, the investigation uses a custom-built, balanced dataset of 51 videos (26 authentic and 25 synthetic) designed to evaluate detection resilience against seamless temporal animation lacking traditional spatial splicing boundaries.

DETECTION PIPELINE AND EVALUATION METHODOLOGY

Conceptual Architecture of Video Deepfake Detector

Automated detection of synthetic media in cybersecurity applications requires an end-to-end processing framework capable of handling unconstrained video input. Modern biometric verification systems use a structured pipeline that turns raw video into a security decision. The pipeline standardizes preprocessing, isolates the facial region and aggregates frame-level model outputs into a single authenticity verdict.

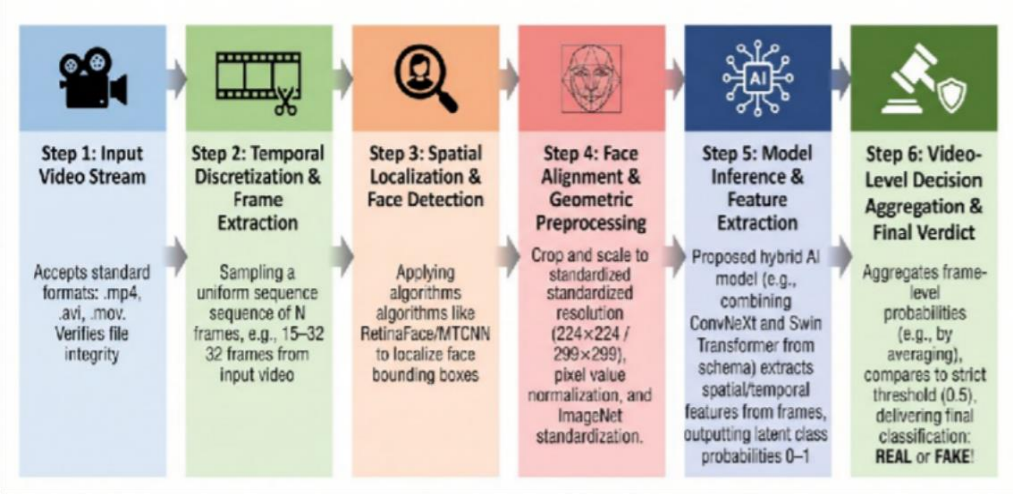

Figure 4. Conceptual Architecture Flowchart

Video Upload Stage: The system ingests digital video files in standard container formats (.mp4, .avi, .mov) with durations ranging between 4 and 6 seconds. Initial file validation checks ensure container integrity, codec compatibility, and file readability before passing data to downstream modules.

Frame Extraction Stage: To preserve computational efficiency while maintaining temporal representation, continuous video streams undergo uniform temporal sampling. A discretized subset of N frames (where 13 <= N <= 32) is extracted per video clip based on file duration.

Face Detection and Alignment Stage: Facial regions within sampled frames are localized using face detection frameworks such as RetinaFace or Multi-Task Cascaded Convolutional Networks (MTCNN). The detector identifies key facial landmarks, including eye centers, nose tip, and mouth corners. Using these coordinates, affine transformations geometrically align the face to correct for spatial tilt. Bounding box crops with padded borders isolate the facial region, which is subsequently resized to the fixed input dimensions required by the classifier architecture (224 x 224 pixels or 380 x 380 pixels).

Preprocessing Stage: Cropped face tensors undergo color space standardization (converting BGR to RGB channels). Pixel intensity values are normalized from integer ranges [0, 255] to the floating-point continuous range [0, 1]. Tensors are then standardized using standard ImageNet mean vectors [0.485, 0.456, 0.406] and standard deviation vectors [0.229, 0.224, 0.225] to match training distribution characteristics.

Model Inference Stage: Standardized face tensors are passed through the deep learning classification model. The network processes each frame tensor independently to generate frame-level probability distributions, outputting a fake probability score P_fake^(i) for frame i.

Video-Level Decision Aggregation Stage: Individual frame scores are collected across the sampled sequence. The overall video authenticity score P_fake_bar is computed via the arithmetic mean of the individual frame probabilities:

$$P\_fake\_bar = (1 / N) * sum\_\{i=1\}^\{N\} P\_fake^(i)$$

The final classification verdict is determined by comparing P_fake_bar against a fixed decision threshold of 0.5. If P_fake_bar >= 0.5, the video is classified as FAKE (synthetic/deepfake); otherwise, it is assigned the classification REAL (authentic).

Mathematical Metrics for Evaluating Detection Quality

To rigorously evaluate and compare the performance of deepfake detection systems, a formal evaluation framework based on binary classification metrics is established [12, 13]. In this domain, the positive class is defined as FAKE (synthetic or AI-generated video content), representing the security threat, while the negative class is defined as REAL (authentic video content recorded from genuine users) [13].

System performance is fundamentally anchored in the four operational outcomes of the confusion matrix, laid out in Table 1:

- True Positive (TP): A synthetic deepfake video correctly identified and flagged as FAKE by the model.
- True Negative (TN): An authentic video correctly recognized and validated as REAL by the model.
- False Positive (FP): A Type I error, wherein a genuine user video is falsely classified as FAKE (false alarm).
- False Negative (FN): A Type II error, wherein a synthetic deepfake attack bypasses detection and is wrongly classified as REAL (missed attack).

Table 1.

Confusion Matrix for Deepfake Detection

True Class \ Model Prediction	**True Class \ Model Prediction**	**True Class \ Model Prediction**
AI Video (Actual FAKE)	AI Video (Actual FAKE)	AI Video (Actual FAKE)
TP (True Positive): Correctly detected fake	TP (True Positive): Correctly detected fake	TP (True Positive): Correctly detected fake

Based on these fundamental outcomes, five standardized performance metrics are calculated to quantify detection capabilities across different security dimensions:

Accuracy: measures the overall proportion of correct decisions made by the classifier across the entire dataset:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

While intuitive, Accuracy assumes balanced class distributions and fails to distinguish between the business costs of false positives and false negatives [13].

Precision: quantifies the exactness of positive classifications, representing the proportion of actual deepfakes among all samples flagged as fake by the system:

Precision = TP / (TP + FP)

High Precision ensures that when the detector triggers a security alert, the flag is highly reliable, minimizing false accusations.

Recall (Sensitivity): measures the detection rate or completeness of the system, defined as the proportion of actual synthetic deepfakes correctly intercepted:

Recall = TP / (TP + FN)

F1-Score: represents the harmonic mean of Precision and Recall, serving as a unified, balanced metric that heavily penalizes extreme imbalances between exactness and completeness:

F1 = 2 * (Precision * Recall) / (Precision + Recall)

Specificity: measures the true negative rate, evaluating the model's ability to correctly identify genuine, unmanipulated videos:

Specificity = TN / (TN + FP)

From a cybersecurity and risk-management perspective, these metrics convey distinct operational priorities. In biometric authentication and identity assurance pipelines, maintaining a high Recall is critical to security, as missing a deepfake attack (False Negative) allows unauthorized threat actors to breach financial systems [12]. Conversely, maintaining high Specificity is essential for commercial viability and user retention [11]. A high false positive rate rejects legitimate clients, which creates operational friction and damages the bank's reputation. An optimal defense system must therefore achieve a balanced trade-off between Recall and Specificity [11, 13].

BENCHMARK OF AI MODELS

Performance Evaluation of Commercial SaaS Solutions

To establish an empirical baseline for deepfake detection quality, commercial Cloud-based Software-as-a-Service (SaaS) platforms were evaluated first [11]. Objective benchmarking requires avoiding data leakage, meaning any overlap between the evaluation media and the training data of the systems under test. Standard public benchmarks such as FaceForensics++, Celeb-DF, and DFDC have been used widely in model training over recent years [10, 14]. Consequently, evaluating platforms on legacy datasets produces artificially inflated performance metrics that do not reflect real-world defensive efficacy against emerging threat vectors.

To address this challenge, a custom, balanced dataset of 51 video clips with durations between 4 and 6 seconds was constructed. The dataset has two balanced classes:

Authentic (REAL) Class: 26 videos recorded from live subjects under natural lighting conditions without facial retouching, filters, or post-processing modifications.

Synthetic (FAKE) Class: 25 synthetic videos generated via a two-stage process modeling advanced image-to-video animation. First, static photographs of two real individuals were combined using face-merging algorithms to synthesize a unified novel identity. Second, the merged static image was animated using modern generative engines (such as SadTalker and AnimateDiff) to produce realistic head movements, natural eye blinking, and micro-expressions [4, 9].

Traditional face-swapping inserts a source face into a target frame and leaves blending seams, skin-tone boundaries or resolution discontinuities. Image-to-video animation generates the whole sequence from one static image, so the face has no boundary at all, which makes it harder to detect [4, 7, 9].

Two leading commercial platforms offering public web-based verification were evaluated against this custom dataset: TruthScan and Hive Moderation. Each of the 51 videos was uploaded independently without submitting metadata or class labels. Table 2 gives the resulting confusion matrices and Table 3 the derived metrics.

Table 2.

Commercial Solutions Confusion Matrix

Service	TP	TN	FN	FP	Total Errors
TruthScan	20	25	5	1	6 out of 51
Hive Moderation	21	25	4	1	5 out of 51

Table 3.

Commercial Solutions Comparative Metrics.

Service	Accuracy, %	Precision, %	Recall, %	F1-Score, %	Specificity, %
TruthScan	88.2	95.2	80.0	8.69	96.1
Hive Moderation	90.1	95,4	84.0	89.3	96.1

Both platforms performed reliably on authentic content, achieving an identical Specificity of 0.961 by generating only a single false positive error across 26 real videos. This minimizes unnecessary friction for genuine banking customers [11]. However, both platforms exhibited a notable drop in Recall (0.800 for TruthScan and 0.840 for Hive Moderation), failing to detect 16% to 20% of the image-to-video deepfake attacks. This indicates that commercial cloud detectors remain partially susceptible to next-generation synthesis techniques that lack explicit spatial boundary artifacts.

Commercial SaaS solutions also have structural limitations in financial cybersecurity. Relying on external APIs requires transmitting highly sensitive customer biometric video files across public networks to third-party cloud infrastructure. This architecture introduces data leakage risks, creates compliance conflicts with stringent privacy regulations such as the European Union General Data Protection Regulation (GDPR), and incurs substantial, recurring API transaction costs at commercial banking scale.

Performance Benchmarking of Open-Source Neural Networks

To identify an effective, privacy-compliant alternative that can be deployed locally on-premise without exposing customer biometric media to external clouds, six open-source deep learning models representing distinct architectural paradigms were benchmarked [14-16, 18-20]. All offline experiments were executed within a unified Google Colaboratory GPU environment using a single Nvidia Tesla T4 GPU (16 GB VRAM) running PyTorch.

For models processing individual image frames rather than direct video containers, uniform frame extraction isolated 13 to 32 frames per video clip based on duration. Video-level classification decisions were established by computing the arithmetic mean of frame-level fake probabilities and applying a standard decision threshold of 0.5. Table 4 gives the confusion matrices for all six models and Table 5 the derived metrics.

Table 4.

Open-Source Models Confusion Matrix.

No.	Model Name	TP	TN	FN	FP
1	dfdc EfficientNet-B7	6	24	19	2
2	CLIP-based (yermandy)	2	21	23	5
3	GANomaly	15	14	10	12
4	prithivMLmods/v1 (ViT)	19	12	6	14
5	GenConViT	19	24	6	2
6	Deep-Fake-Detector-v2 (ViT)	23	5	2	21

Table 5.

Open-Source Models Comparative Metrics.

No.	Model	Accuracy, %	Precision, %	Recall, %	F1-Score, %
1	dfdc EfficientNet-B7	58.8	75.0	24.0	36.3

2	CLIP-based	45.0	28.5	0.80	12.4
3	GANomaly	56.8	55.5	60.0	57.6
4	prithivMLmods/v1	60.7	57.5	76.0	65.4
5	GenConViT	84.3	90.4	76.0	82.5
6	Deep-Fake-Detector-v2	54.9	52.2	92.0	66.6

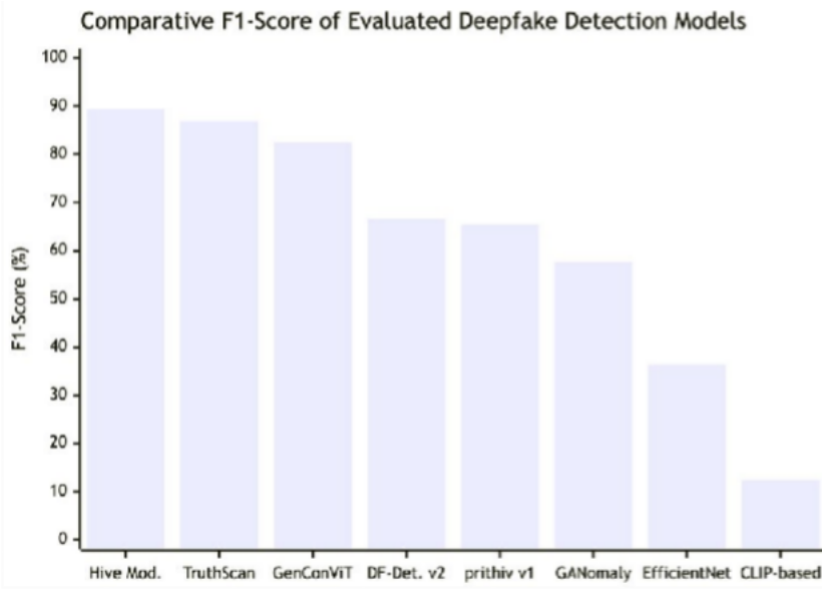

Figure 5. Comparative F1-score across all eight models

Performance varied sharply across architectures when confronted with image-to-video animationPerformance varied sharply across architectures when confronted with image-to-video animation:

1. dfdc (selimsef, EfficientNet-B7). Despite winning the 2020 Facebook DFDC competition, this classical convolutional architecture failed to generalize across domains [14]. It preserved a high specificity (0.923) but missed 19 of the 25 synthetic videos, giving a recall of 0.240. Because EfficientNet-B7 was optimized to detect spatial splicing boundaries and blending artifacts from legacy face-swap datasets, it did not transfer to boundary-free image-to-video animation [14].
2. CLIP-based (yermandy). The CLIP ViT-L/14 visual encoder failed to isolate facial manipulation artifacts, missing 23 of the 25 synthetic samples. This indicates that foundation models fine-tuned without dedicated face-cropping pre-processing pipelines cannot reliably isolate micro-level generative anomalies in unconstrained natural scenes [15].
3. GANomaly. As a semi-supervised anomaly detector trained exclusively on authentic samples, GANomaly struggled to differentiate subtle generative face variations from natural human facial expressions, producing 12 false alarms and a specificity of 0.538.
4. prithivMLmods v1 (ViT-base). Recall was moderate at 0.760, but the model generated 14 false positives out of 26 authentic videos. A specificity of 0.461 is unacceptable in commercial banking, as it would falsely flag over half of legitimate onboarding clients as fraudsters.
5. GenConViT. GenConViT was the best open-source model tested. Its hybrid framework generalized across domains: ConvNeXt handles local spatial features, Swin Transformers handle global self-attention, and the AE/VAE branches handle latent space anomaly detection [16]. It generated only 2 false positives (specificity 0.923) while reaching an F1-score of 0.825, close to commercial SaaS performance (0.869 to 0.893).
6. Deep-Fake-Detector-v2 (ViT). This Vision Transformer exhibited extreme out-of-distribution bias. It intercepted 23 of the 25 deepfakes (recall 0.920) but classified 21 authentic user videos as AI-generated, producing a specificity of 0.192.

GenConViT was the strongest open-source candidate overall. It has the highest accuracy (0.843), the highest precision (0.904) and a specificity of 0.923, with a balanced F1-score of 0.825. That is close to commercial detection quality, and because it can be run locally, it avoids both the cloud privacy risk and the subscription cost.

CONCLUSIONS

Image-to-video animation built on merged facial geometries gets past detection pipelines that look for boundaries. Because the frames are generated from one static image, there are no spatial seams, skin-tone mismatches or blending artifacts for those pipelines to find. EfficientNet-B7 shows the problem clearly. It won the 2020 DeepFake Detection Challenge, but on boundary-free animation its recall dropped to 0.240, missing 76% of the synthetic videos. Open-source hybrid architectures are a workable, privacy-preserving alternative to commercial SaaS platforms. GenConViT gave the most balanced profile of the open-source models tested: accuracy 0.843, F1-score 0.825, precision 0.904, and a specificity of 0.923, meaning a low false alarm rate. GenConViT combines ConvNeXt local feature extraction, Swin Transformer global self-attention and parallel autoencoders for latent space anomaly detection, and this combination picked up small generative variations. Its accuracy is close to Hive Moderation (0.901), and because it runs locally, the bank avoids the compliance exposure and per-transaction costs of sending biometric data to an external API.

Further work will extend the dataset to other multimodal threats, including voice cloning, acoustic lip-syncing and face-reenactment, and fine-tune GenConViT on those vectors to raise its recall. Enlarging the evaluation set beyond 51 clips and testing against additional image-to-video engines would also strengthen the generalization claims made here.

REFERENCES

1. Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 39–52. https://doi.org/10.22215/timreview/1282
2. Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys, 54*(1), Article 7, 1–41. https://doi.org/10.1145/3425780
3. European Union Agency for Cybersecurity. (2023). *ENISA threat landscape 2023*. ENISA. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023
4. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion, 64*, 131–148. https://doi.org/10.1016/j.inffus.2020.06.014
5. Pearson, J. (2022). Deepfake video of Zelenskyy could be “tip of the iceberg” in info war, experts warn. *NPR*. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia
6. Chen, H., & Magramo, K. (2024). Finance worker pays out $25 million after video call with deepfake “chief financial officer.” *CNN*. https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems, 27*. https://papers.nips.cc/paper/5423-generative-adversarial-nets
8. Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/1812.04948
9. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems, 33*. https://arxiv.org/abs/2006.11239
10. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. https://arxiv.org/abs/1901.08971
11. Sarker, I. H. (2021). Deep cybersecurity: A comprehensive overview from neural network and deep learning perspective. *SN Computer Science, 2*, Article 154. https://doi.org/10.1007/s42979-021-00545-3
12. Verdoliva, L. (2020). Media forensics and DeepFakes: An overview. *IEEE Journal of Selected Topics in Signal Processing, 14*(5), 910–932. https://doi.org/10.1109/JSTSP.2020.3002101
13. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) dataset. *arXiv*. https://arxiv.org/abs/2006.07397

14. Seferbekov, S. (2020). *Deepfake Detection Challenge 1st place solution*. Kaggle. https://www.kaggle.com/c/deepfake-detection-challenge
15. Yermakov, A., Cech, J., & Matas, J. (2025). Unlocking the hidden potential of CLIP in generalizable deepfake detection. *arXiv*. https://arxiv.org/abs/2503.19683
16. Wodajo, D., Atnafu, S., & Akhtar, Z. (2023). Deepfake video detection using generative convolutional vision transformer. *arXiv*. https://arxiv.org/abs/2307.07036
17. DeepStrike. (2025). *Deepfake statistics 2025: AI fraud data & trends*. https://deepstrike.io/blog/deepfake-statistics-2025
18. Akcay, S., Atapour-Abarghouei, A., & Breckon, T. P. (2018). GANomaly: Semi-supervised anomaly detection via adversarial training. *arXiv*. https://arxiv.org/abs/1805.06725
19. prithivMLmods. (2024). *Deep-Fake-Detector-Model* [Machine learning model]. Hugging Face. https://huggingface.co/prithivMLmods/Deep-Fake-Detector-Model
20. prithivMLmods. (2025). *Deep-Fake-Detector-v2-Model* [Machine learning model]. Hugging Face. https://huggingface.co/prithivMLmods/Deep-Fake-Detector-v2-Model

Отримано редакцією журналу / Received: 10.02.26
Прорецензовано / Revised: 20.02.26
Схвалено до друку / Accepted: 24.09.26