Benchmarking Forensic Reliability: A Comparative Analysis of Automated AI Models versus Human Perception in Detecting Low-Bitrate Deepfakes

Sai Bhavani Venkatesh Pasupuleti
Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
pasupuleti1222002@gmail.com

Abstract

The rapid progress of Generative Adversarial Networks (GANs) has enabled highly realistic facial manipulations, creating a serious threat to the integrity of digital evidence. Modern deepfake detectors, particularly convolutional neural network (CNN) based architectures, report strong performance on curated benchmarks; however, their practical reliability is rarely tested under the heavy lossy compression applied by social and messaging platforms. This work presents a comparative forensic study of three automated detection pipelines XceptionNet, MesoNet, and a temporal CNN–RNN model against documented human performance when videos are degraded using H.264 compression at C23 and C40 levels. Using 150 samples from the FaceForensics++ dataset [1], a new metric, the GAN Fingerprint Survivability Index (GFSI), is introduced to quantify the fraction of high-frequency forensic cues that survive compression. Empirically, GFSI ≈ 0.0004 at C40, indicating near-total removal of spectral GAN fingerprints. Experimental results show that XceptionNet collapses into an extreme "fake" bias, MesoNet loses 12% accuracy, and the temporal model performs below random chance on low-quality videos, whereas humans maintain approximately 70% accuracy by relying on semantic cues. These findings demonstrate that automated detectors alone are unsafe for low-bitrate forensic evidence and motivate hybrid human-AI workflows.

Keywords: Deepfake detection, GAN fingerprints, H.264 compression, digital forensics, human–AI comparison.

1. Introduction

Deepfake videos generated by GAN-based pipelines have become increasingly convincing and accessible, with direct implications for privacy, cybersecurity, and legal proceedings. High-quality benchmark datasets such as FaceForensics++ [1] have enabled rapid progress in training and evaluating automated detectors. Many reported systems achieve high accuracy under controlled conditions, giving the impression that the deep-fake detection problem is largely solved.

In real forensic workflows, however, video evidence is rarely pristine. Content is frequently captured on mobile devices, shared across applications, and repeatedly transcoded. Social platforms typically apply strong H.264/H.265 compression at low bitrates, acting as a low-pass filter that suppresses the high-frequency artifacts often exploited by CNN-based detectors. As a result, the spectral "fingerprints" of GAN artifacts may be largely removed before analysis. Human observers operate differently. Cognitive and perceptual studies show that people rely on holistic and semantic cues in faces, including biological plausibility, natural blinking, smooth head motion, and synchrony between speech and lip movement [6]. These cues tend to survive even aggressive compression because they are expressed at coarser spatial and temporal scales.

This study focuses on quantifying the reliability gap between automated detectors and human observers under realistic compression. Three open-source deepfake detection pipelines

D. R. Reddy et al. (eds.), *Proceedings of the First International Conference on Advances in Forensics and Cyber Technologies (ICFACT 2025)*, Advances in Computer Science Research 127,
https://doi.org/10.2991/978-94-6239-610-4_5

XceptionNet, MesoNet, and a temporal CNN–RNN model [2–5] are evaluated on FaceForensics++ videos compressed to C23 and C40. A mathematical index, termed the GAN Fingerprint Survivability Index (GFSI), is proposed to measure the proportion of high-frequency GAN-related information preserved after compression. The goal is to understand not only how much performance is lost, but also why AI systems fail while humans continue to perform reasonably well.

2. Materials and Methods

2.1 Dataset selection

The experiments use a subset of 150 videos drawn from the FaceForensics++ benchmark [1]. The subset consists of 75 pristine (real) samples and 75 manipulated samples, including representative examples of common face manipulation methods such as Face2Face and DeepFake. All clips were standardized to a duration of approximately 10 seconds at 30 frames per second, and audio tracks were removed to focus solely on visual evidence. Dataset acquisition followed the official FaceForensics++ access procedure, and videos were downloaded using the authors provided utilities [8].

2.2 Compression pipeline

To emulate real-world messaging workflows, each video was transcoded using an H.264 encoder at two distinct compression levels:

- CRF 23: moderate compression, representing "platform-friendly" quality;
- CRF 40: heavy compression, approximating messaging application constraints with bitrates below roughly 500 kbps.

Transcoding was performed with a standard libx264 pipeline. These two conditions are referred to as "HQ" (original FaceForensics++ quality) and "LQ" (C40 compressed) in the analysis.

2.3 GAN Fingerprint Survivability Index (GFSI)

Deepfake detectors often rely on subtle high-frequency differences between real and manipulated frames. To formalize the effect of compression on these cues, the GAN Fingerprint Survivability Index is defined as

$$GFSI = \frac{\sum_{u,v>4} |F_q(u,v)|}{\sum_{u,v>4} |F(u,v)|}$$

Where F(u,v) denotes the two-dimensional discrete cosine transform (DCT) coefficients of the original frame and F_q(u,v) represents the coefficients after compression. The summation is taken over higher-frequency components (indices u,v > 4), which primarily capture fine-grained textures and noise patterns. A GFSI value close to zero indicates that the compression process has destroyed most of the high-frequency energy that might carry GAN-related fingerprints. In the reported experiments, GFSI values around 0.0004 were observed at C40, implying that only about 0.04% of the original high-frequency content survives.

2.4 Automated detection models

Three deepfake detection pipelines were evaluated:

1. XceptionNet-based detector, initialized with pretrained weights trained on FaceForensics++ for different compression levels [5].
2. MesoNet (Meso4-DF), a mesoscopic architecture designed for detecting facial forgeries [2].
3. A temporal CNN–RNN detector that combines convolutional feature extraction with recurrent layers for sequential analysis [3].

All models were used as released in their respective open-source repositories [2–5], with appropriate input preprocessing and no additional fine-tuning.

2.5 Inference pipeline

To keep the evaluation consistent across models, the following pipeline was applied:

1. Sampled every 5th frame from each video.
2. Resized frames to match each model's expected input resolution.
3. Performed forward passes through the model to obtain per-frame predictions.
4. Aggregated frame-level predictions using majority voting with a decision threshold of 0.5.

This produced a single "real" or "fake" prediction per video for each model and compression condition.

2.6 Human benchmark

As a reference for human-level performance, accuracy estimates were taken from the user study reported in the FaceForensics++ work [1], which found that human raters typically achieve around 70% accuracy on low-quality compressed deepfakes. In addition, an informal review of a subset of degraded clips was performed to confirm that semantic anomalies such as unnatural eye motion or lip-sync errors remained visible to a human observer even when high-frequency texture information was heavily suppressed.

2.7 Results

This section presents the performance evaluation of three automated deepfake detection models XceptionNet, MesoNet (Meso4-DF), and a Temporal CNN–RNN pipeline under two compression conditions: high-quality (HQ) FaceForensics++ input and low-quality (LQ) C40-recompressed videos. For each model, accuracy and false negative rates (FNR) were computed across 150 samples.

3. Performance of Automated Detectors

Table 1. Performance of Automated Detectors under High-Quality (HQ) and Low-Quality (LQ, C40) Compression

Model	HQ Accuracy (%)	LQ Accuracy (%)	Accuracy Drop (%)	LQ False Negative Rate (%)
XceptionNet	66.7	66.7	0.0	0.0

MesoNet (Meso4-DF)	62.7	50.7	12.0	36.0
Temporal CNN–RNN	44.0	36.0	8.0	86.0

Key Observations

i. **XceptionNet** shows identical accuracy in HQ and LQ conditions, but closer inspection reveals that the model predicts nearly all samples as *fake*. This leads to:

   1. Artificially stable accuracy,
   2. Extremely high false positive rate, making the model operationally unreliable.

ii. **MesoNet** demonstrates the clearest degradation pattern. Its **12% accuracy drop** and **36% FNR** indicate that compression directly impairs its mesoscopic-texture features.

iii. The **Temporal CNN–RNN** model collapses under compression, with accuracy falling below random chance and an FNR of **86%**, meaning it fails to identify the majority of manipulated videos.

a. Comparison to Human Performance

Human evaluators, based on the FaceForensics++ perceptual study, achieve ∼**70% accuracy** on similarly compressed deepfake videos. This highlights the **resilience of semantic-level reasoning**, which survives compression, unlike the spectral cues used by CNN models.

b. The Compression Gap

The results reveal a clear discrepancy between the theoretical performance of deepfake detectors and their real-world reliability under heavy video compression, a phenomenon referred to as the Compression Gap. H.264 encoding at high CRF levels (e.g., C40) functions as a strong low-pass filter, removing nearly all high-frequency forensic cues that automated models rely upon. The proposed GAN Fingerprint Survivability Index (GFSI ≈ 0.0004) confirms that more than 99.96% of high-frequency DCT energy is eliminated after compression, leaving automated detectors with insufficient discriminative information.

Consequently, XceptionNet collapses into a high-bias "fake" prediction pattern, MesoNet experiences a significant 12% accuracy decline, and the temporal CNN–RNN pipeline fails almost entirely, reaching an 86% false-negative rate. Humans, however, maintain approximately 70% accuracy because semantic cues such as blink timing, motion coherence, and lip-sync consistency survive even aggressive compression. Thus, the Compression Gap highlights a fundamental difference in the feature spaces exploited by AI systems and human observers. Compression disproportionately destroys the spectral signals essential for AI-based detection while preserving the semantic cues used in human perception, underscoring the need for hybrid forensic analysis workflows.

4. Discussion

The experimental results highlight a key mismatch between the assumptions made by many deep-fake detection models and the conditions encountered in operational forensic environments. CNN-based detectors typically exploit local texture irregularities, subtle color shifts, or periodic noise patterns that are preserved in relatively high-quality video. However, once videos are aggressively encoded for bandwidth efficiency, much of that information is removed or heavily distorted.

The near-zero GFSI values measured for C40 compression support the conclusion that most GAN-related spectral traces do not survive real-world encoding. Under these conditions, detectors such as XceptionNet and MesoNet are effectively deprived of the signals they were trained to exploit. This explains why one model collapses to a trivial high-bias decision rule and another suffers substantial accuracy loss.

Human observers, in contrast, rely on semantic and temporal cues that operate at coarser scales. Examples include whether the mouth movement matches the speech pattern, whether eyes blink at plausible intervals, and whether head pose transitions are physically natural. These cues are largely invariant to the smoothing and quantization performed by video encoders, which explains why human performance remains relatively stable even at low bitrates.

The findings suggest that simply training larger or deeper CNNs on more data will not fully solve the robustness problem if the features they rely on are inherently destroyed by compression. Instead, models may need to emulate human-like reasoning about motion patterns, facial dynamics, and semantic coherence, or incorporate explicit robustness mechanisms that are less sensitive to low-frequency distortions.

5. Conclusion

This study systematically examined the reliability of three automated deepfake detectors and human observers when faced with low-bitrate, heavily compressed video evidence. Using FaceForensics++ samples subjected to H.264 compression at C23 and C40, the GAN Fingerprint Survivability Index (GFSI) was introduced to quantify the fraction of high-frequency forensic information that survives encoding. A measured value of approximately 0.0004 for GFSI at C40 indicates that nearly all spectral fingerprints are eliminated.

Under these conditions, XceptionNet exhibited extreme prediction bias, MesoNet experienced a notable 12% accuracy drop, and the temporal CNN–RNN detector failed to generalize, performing worse than chance. Human accuracy, by contrast, remained around 70% due to reliance on semantic and temporal cues. These results strongly suggest that fully automated pipelines are unsafe for low-bitrate forensic decision-making and that human-in-the-loop verification should be treated as a requirement rather than an optional safeguard.

All experiment scripts, preprocessing utilities, model configurations, and result files used in this study are provided in a reproducible form in the dataset 'Deepfake Compression Forensics: C23 vs C40' hosted on Kaggle (https://www.kaggle.com/datasets/psbvenkatesh/deepfake-compression-forensics-c23-vs-c40).

Acknowledgement:

The author gratefully acknowledges the creators of the FaceForensics++ dataset and the maintainers of the XceptionNet, MesoNet, and temporal deepfake detection repositories for releasing pretrained models and code that enabled this study.

Conflict Of Interest:

The author declares that there are no financial or personal conflicts of interest related to this work.

References:

[1] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019. FaceForensics++ utilities and scripts: https://github.com/ondyari/FaceForensics

[2] Afchar, D., Nozick, V., Yamagishi, J., Echizen, I. MesoNet: A Compact Facial Forgery Detection Network. In: IEEE International Workshop on Information Forensics and Security (WIFS), 2018. MesoNet repository and Meso4_DF weights: https://github.com/DariusAf/MesoNet

[3] Chinmay Rane. Deep-Fake Video Detection (ResNeXt + LSTM Temporal Model). GitHub Repository: https://github.com/ChinmayRane16/Deep-Fake-Video-Detection

[4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. Generative Adversarial Networks. In: Advances in Neural Information Processing Systems (NeurIPS), 2014.

[5] Tapia, R. XceptionNet FF++ Pretrained Models (c23, c40). GitHub Repository: https://github.com/ReneDTapia/DeepFake

[6] Tanaka, J. W., Farah, M. J. Parts and Wholes in Face Perception. Quarterly Journal of Experimental Psychology, 46(2), 225–245 (1993).

[7] MesoNet Repository (Additional Resources). GitHub: https://github.com/DariusAf/MesoNet

[8] FaceForensics++ Dataset Tools and Download Scripts. GitHub: https://github.com/ondyari/FaceForensics