RCMOS – Multidisciplinary Scientific Journal O Saber.
ISSN: 2675-9128. São Paulo, Brazil.

Year III, vol. 2, 2023. | Submitted: October 21, 2023 | Accepted: October 23, 2023 | Published: October 25, 2023

Deepfake Forensics: Integration of Computer Vision and Unsupervised Learning Techniques for Video Authenticity Analysis

Matheus de Oliveira Pereira Paula holds a Bachelor's degree in Information Systems from the Federal Institute of Education, Science and Technology Fluminense and is enrolled in the MSc in Data Science and Artificial Intelligence at Université Côte d'Azur.

ABSTRACT

The *deepfake* landscape demands a **paradigm shift** in detection, moving from Supervised Learning (SL) models that search for known artifacts to systems based on **Unsupervised Learning (UL)**, capable of identifying **anomalies and statistical deviations** from authentic media. This paper provides a detailed analysis of integrating **Advanced Computer Vision** and **UL Techniques** to create robust *Deepfake Forensics* pipelines. The focus is on developing systems that do not depend on a predefined *deepfake* dataset, making them ideal for detecting **zero-generation** manipulations or never-before-seen forgery techniques. We explore the use of *Autoencoders* and Generative Adversarial Networks (*GANs*) in their ability to model the distribution of normality (real videos) and the subsequent application of reconstruction metrics and latent mapping deviation to isolate the anomalous patterns that characterize the forgery. The practical application of these methodologies is vital for **cybersecurity** and **content moderation** on social media platforms, offering an **authenticity verification** mechanism that is resilient to the constant evolution of synthetic media technologies.

Keywords: Deepfake Forensics; Unsupervised Learning; Computer Vision; Autoencoders; Anomaly Detection; Cybersecurity; Zero-Generation.

1. INTRODUCTION: THE FRAGILITY OF SUPERVISED LEARNING IN THE FACE OF ZERO-GENERATION ATTACKS

The first generation of *deepfake* detectors, based on **Supervised Learning (SL)** and on architectures such as Convolutional Neural Networks (CNNs), achieved high performance but revealed a critical vulnerability: its dependence on **data labels (fake/real)** for *known* manipulation techniques. These models are trained to identify the *fabrication artifacts* specific to a given generator (*e.g.*, FaceSwap, StyleGAN2). However, with every new *deepfake* technique or post-processing variation, the accuracy of SL models collapses, a problem known as **failure to generalize** (*cross-dataset*). This fragility is unsustainable in a scenario of **Adversarial Warfare**, in which new *forgery* techniques (the zero generation of *deepfakes*) emerge continuously.

The challenge lies in the fact that the domain of authentic videos (the "real" data) is fixed and relatively well defined by the laws of physics and biology (lighting coherence, movement, heartbeat, etc.), whereas the domain of *deepfakes* (the "fake" data) is **unlimited and mutable**. Supervised Learning fails because it tries to learn the boundaries between a finite set of known forgeries and the real, leaving the system blind to any forgery that was not seen during

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

training. For **Deepfake Forensics** systems that aim at **cybersecurity** and **information integrity**, this dependence on prior knowledge is a fatal flaw.

This article proposes the exploration of **Unsupervised Learning (UL)** as the fundamental solution to mitigate the zero-generation vulnerability. Instead of learning what is "fake," UL focuses on **modeling the distribution of "normality"**, that is, what is inherently **authentic** and statistically coherent. Any input sample that **deviates significantly** from this distribution of normality (in terms of reconstruction, latency, or statistical patterns) is classified as an **anomaly** and, consequently, as a *deepfake*. This approach inverts the problem: detection is no longer a search for a known artifact but the identification of a **break in the statistical coherence of reality**.

The analysis will concentrate on integrating advanced **Computer Vision** techniques to extract relevant forensic *features*, which are then processed by UL algorithms such as *Autoencoders* (AEs) and variations of Generative Adversarial Networks (GANs) trained for anomaly detection. The goal is to demonstrate the feasibility of building detection *pipelines* that are **inherently resilient** to the evolution of the *deepfake*, with immediate application in **content verification on social media** and in the **forensic analysis** of cybersecurity incidents, where speed and robustness against the unknown are critical.

2. AUTOENCODERS AND THE MODELING OF THE DISTRIBUTION OF NORMALITY

Autoencoders (AEs), a prominent class of neural networks in **Unsupervised Learning (UL)**, are the cornerstone of anomaly detection in Computer Vision data and represent a powerful tool for *deepfake* forensics. The working principle of the AE is simple and elegantly suited to the problem: it is trained exclusively on a vast dataset of **authentic videos ("normal" samples)** to learn a **compact representation (the latent code)** of these data and then reconstruct them with the highest possible fidelity. The AE effectively learns the **statistical distribution of normality**.

The architecture of the AE consists of an **Encoder** (which maps the input image to a lower-dimensional latent space) and a **Decoder** (which reconstructs the image from this latent code). Once trained on real data, the AE becomes highly efficient at reconstructing **authentic faces and scenes**. However, when presented with a **manipulated video (*deepfake*)**, which contains visual and statistical patterns the AE has never encountered (the *forgery* artifacts), the network fails to map the *deepfake* accurately into its latent space and, consequently, fails to reconstruct it. The model, in trying to reconstruct what it does not understand, produces an image with a **high reconstruction error**.

The **Reconstruction Error** (generally measured by metrics such as the *Mean Squared Error*, MSE) becomes the primary metric for anomaly detection. In normal samples, the MSE is low; in *deepfakes*, the MSE is significantly higher, because the model cannot encode the forgery artifacts efficiently. This difference in reconstruction error serves as an **unsupervised discriminator**. The anomaly threshold is defined statistically from the distribution of reconstruction errors of the authentic training data (the "normal" data), and any error that

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

exceeds this threshold is classified as a *deepfake*.

The great strength of the AE in *Deepfake Forensics* is its **inherent robustness against zero-generation attacks**. Because it does not learn what is false but what is real, it is able to detect **any statistical deviation** from reality, regardless of the manipulation technique used (whether *FaceSwap*, *Face Reenactment*, or *Diffusion* models). The limitation of the AE, however, lies in its susceptibility to *deepfakes* that are **extremely convincing**, that is, that fit the distribution of normality very well. Refining this technique involves the use of **Variational Autoencoders (VAEs)** and the integration of attention modules to focus the reconstruction on sensitive areas of the face, such as skin texture and eye reflections, where artifacts are most likely to appear.

3. UNSUPERVISED LEARNING WITH ADVERSARIAL ARCHITECTURES (GAN-BASED ANOMALY DETECTION)

Although Generative Adversarial Networks (GANs) are the driving force behind the creation of *deepfakes*, they can also be creatively adapted in the field of **Unsupervised Learning (UL)** for anomaly detection. This approach, called *GAN-Based Anomaly Detection*, exploits the ability of the **Discriminator** of a well-trained GAN to model the **space of normal data** with high precision, turning it into a detector of statistical deviations with greater sensitivity than traditional AEs.

The most common method involves training a GAN on a dataset that is **exclusively authentic** (the domain of normality). The Generator (G) learns to produce realistic samples, and the Discriminator (D) learns to distinguish generated (synthetic) images from real (authentic) ones. For *deepfake* detection, the focus shifts to the **Discriminator**. At inference time, a new video (which may be a *deepfake*) is introduced. If the Discriminator classifies it as "real" with high certainty, the video is considered authentic, since it fits well within the distribution that the Discriminator learned as normal. If, however, the Discriminator classifies it as "fake" or, in a more refined way, maps it to a region of the latent space with a **high mapping distance** (*mapping distance*), the sample is flagged as anomalous.

A powerful variation is **AnoGAN (Anomaly Detection with GANs)**, which tries to map the input image (the potential *deepfake*) back to the **latent space (z)** of the trained Generator. The principle is that an authentic image, belonging to the distribution of normality, must have a corresponding point *z* in the Generator's latent space that reconstructs it with high fidelity. An anomalous image (the *deepfake*) should not have a *z* that reconstructs it well. The anomaly score is therefore a combination of the **reconstruction error** and the **latent mapping distance** (the distance between the *z* found and the training latent space).

This adversarial UL approach offers two crucial advantages: **sensitivity and robustness**. The adversarial nature of GAN training forces the model to learn sharper decision boundaries and more detailed representations of normality than a simple AE, which results in greater sensitivity in detecting subtle deviations. Moreover, as with AEs, GAN-based anomaly detection is **intrinsically resistant to zero-generation attacks**, since the model is not looking for known *deepfakes* but for any sample that violates the statistical law of authenticity

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

it has learned, making it extremely valuable for **cybersecurity** in a context of rapidly evolving threats.

4. INTEGRATING COMPUTER VISION: EXTRACTING FORENSIC *FEATURES*

The success of **Unsupervised Learning (UL)** in anomaly detection depends critically on the quality of the *features* extracted from videos by **Computer Vision (CV)**. Applying UL techniques to *Deepfake Forensics* is not limited to feeding the network raw pixels; it is necessary to preprocess and isolate the areas of the video where manipulation artifacts are most likely to occur, ensuring that UL is modeling the normality of the most relevant forensic *features*, and not noise or irrelevant context.

The first CV step involves **Facial Detection and Alignment**. Algorithms such as **MTCNN** (*Multi-task Cascaded Convolutional Networks*) or **RetinaFace** are used to identify and crop the facial region with high precision. The subsequent alignment, based on reference points (such as the eyes, nose, and mouth), standardizes the pose of the face, minimizing non-manipulative variation and allowing the UL model to concentrate on the internal anomalies of the face, such as skin texture and micro-expressions.

Next, CV is used to isolate critical **temporal and physiological signals**. One of the hardest things for *deepfakes* to fake is the **Remote Photoplethysmography (rPPG)** signal, which is the variation in skin color caused by blood pulsation (the heartbeat). CV techniques can be used to extract this subtle color time series from the face. By feeding the AE or the GAN with the **rPPG time series** instead of only the static frame, UL learns the distribution of normality of human heart rates. Any break in temporal coherence or an unrealistic rPPG frequency will be classified as an anomaly, regardless of the visual artifact of the *deepfake*.

In addition, CV is crucial in **frequency-domain analysis**. Using transforms such as the **Discrete Cosine Transform (DCT)** or the **Fourier Transform**, it is possible to map the video from the spatial domain to the frequency domain. *Deepfakes*, being products of interpolation and neural networks, frequently exhibit **predictable statistical anomalies** in certain frequency bands that real videos do not have. The UL model is then trained on the **statistical characteristics of the frequency domain**, making it sensitive to anomalies that are invisible in the pixel domain but highly revealing of manipulation. This integration of CV and UL is what gives *Deepfake Forensics* an **analytical depth** that first-generation supervised detectors cannot replicate.

5. CHALLENGES AND LIMITATIONS OF UNSUPERVISED LEARNING IN DEEPFAKE FORENSICS

Despite the promise of **Unsupervised Learning (UL)** in overcoming the zero-generation vulnerability, its application in *Deepfake Forensics* faces significant practical and conceptual challenges that must be addressed for large-scale *deployment*. The main challenge is the **rigorous definition of the anomaly threshold**. The threshold that separates a “normal” video from an “anomaly” (the *deepfake*) is typically determined statistically from the distribution of reconstruction error (in AEs) or of mapping distance (in GANs). However, natural variations

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

in authentic videos, such as camera noise, compression, different skin tones, and lighting conditions, can by themselves produce a high reconstruction error.

This sensitivity to **non-adversarial variations** (noise) and to **natural variations** (diversity of pose and environment) can lead to a high rate of **false positives**. An AE may classify a real but heavily compressed video (and therefore with many JPEG compression artifacts) as a *deepfake*, simply because the compression noise does not belong to the distribution of normality learned from high-quality data. This is a serious problem on social media platforms, where most videos are heavily compressed. To mitigate it, the "normality" training dataset must be **carefully diversified** to include all the forms of degradation and environmental variation expected in the real environment.

Another critical limitation is the **difficulty of identifying high-fidelity *deepfakes***. As generative technology (such as *Diffusion* models and advanced GANs) evolves, *deepfakes* become **almost indistinguishable** from real videos, fitting perfectly within the distribution of normality of the authentic data. A very-high-quality *deepfake* will result in a reconstruction error as low as that of a real video, making it invisible to the anomaly detector. In such cases, detection must rely on **higher-order frequency-domain artifacts** or on **temporal signals** (such as rPPG), which are harder to eliminate even for the most sophisticated generators, requiring a more complex integration of *features*.

Finally, UL does not provide the **explainability** that Supervised Learning with XAI can offer. Although it says "this video is an anomaly," it does not explain *what* the nature of that anomaly is. The AE may show a high reconstruction error, but the forensic examiner does not know whether the failure lies in the rPPG mapping or in the texture of the hair. Therefore, the next frontier for *Deepfake Forensics* is the integration of **Unsupervised Learning for Detection** with **Supervised Learning for Explanation**, combining the robustness of anomaly detection with the capacity for visual and forensic justification.

6. PRACTICAL APPLICATION IN CYBERSECURITY AND SOCIAL MEDIA

The integration of **Computer Vision** and **Unsupervised Learning (UL)** in *Deepfake Forensics* has transformative implications for **cybersecurity** and **content moderation** on social media platforms, offering solutions to challenges that supervised models cannot solve. The main value lies in the ability to establish a **proactive and adaptive defense mechanism** against the proliferation of synthetic media.

In the context of **cybersecurity**, real-time anomaly detection is crucial for the **verification of authenticity in sensitive communications**. In a *spear-phishing* attack using a voice or video *deepfake* (such as *CEO fraud*), a UL system trained on the normality of the target's voice or image can immediately flag the content as an anomaly, even if the forgery technique is new. The AE, having learned the person's **biometric fingerprint**, can detect the statistical deviation of the zero-generation *deepfake*, functioning as a **biometric authenticity filter** that is vital for protecting assets and confidential information.

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

For **social media platforms**, the challenge is the **massive volume and speed** of video dissemination. The UL *pipeline* becomes a **first-line screening layer** of very high efficiency. *Autoencoders* and GAN-Based Anomaly models can process millions of videos, identifying the *deepfake* candidates that present a high anomaly score. This is especially useful for dealing with the volatile nature of *deepfake generation*, where techniques change constantly. Once a video is flagged as anomalous by UL, it can be forwarded to a **secondary supervised system** (with XAI) and to a **human review** for forensic confirmation.

The use of UL on social media is not limited to binary classification. The model can be used to **quantify the degree of anomaly** (the reconstruction error), allowing the platform to prioritize the removal or flagging of the *deepfakes* that deviate most from normality (the most bizarre or lowest-quality ones, which cause the greatest confusion) or those that pose the greatest threat (the most convincing ones, which deviate only slightly, indicating a high level of sophisticated *forgery*). By modeling normality continuously, platforms can develop a system of **adaptive surveillance** that is inherently more resilient and less dependent on blacklists of forgery techniques.

7. INTEGRATING TEMPORAL SIGNALS AND SPATIO-TEMPORAL AUTOENCODERS

Deepfake Forensics requires that **Unsupervised Learning (UL)** models look beyond the static frame, integrating **temporal coherence** as a fundamental aspect of normality. Real videos maintain a logical consistency over time that is notoriously difficult for *deepfake* generators to replicate without introducing artifacts. The architectural solution to this problem is the development of **Spatio-Temporal Autoencoders (STAEs)**.

STAEs extend the functionality of traditional AEs by incorporating layers that model the temporal dimension. This is usually done by replacing 2D convolutional layers with **3D convolutional layers** (which operate on time × width × height *volume* spaces) or by integrating **recurrent networks (LSTMs or GRUs)** or **temporal attention modules** after the spatial encoding stage. When trained on real videos, these models learn the normality not only of the appearance of a single frame but also of the **dynamics of movement** between frames, including the coherence of lighting across a sequence, the rate of eye blinking, and the continuity of optical flow.

Applying STAEs to anomaly detection is particularly effective at identifying **artifacts of *face swapping*** in videos. Such manipulations frequently introduce **inconsistencies in the transition of the facial mask** between frames, resulting in *flickering* or unstable edges that are hard to see with the naked eye but that violate the normality of smooth, continuous movement. In reconstructing the video, the STAE will generate a **high reconstruction error in the areas and moments of incoherent transition**, flagging the temporal anomaly.

In addition, modeling **physiological signals** through temporal integration is the key to robustness. The rPPG, the heartbeat signal, is a time series. An STAE fed with rPPG information over time will learn the normal pattern of human heart-rate variability. A manipulation that fails to simulate this variability, or that presents a fixed or statistically

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

unrealistic frequency (a common failing of basic *deepfakes*), will be immediately classified as a **temporal and physiological anomaly**, providing evidence of falsity that is almost impossible to replicate with current *forgery* techniques. Spatio-temporal modeling is therefore the most robust path to ensuring the **resilience** of *Deepfake Forensics* against the growing sophistication of video manipulation.

8. CONCLUSION AND FUTURE IMPLICATIONS

This study has demonstrated, on the basis of a rigorous architectural analysis, that defending against the next generation of *deepfakes* requires a **strategic abandonment** of the **Supervised Learning (SL)** paradigm in favor of robust **Unsupervised Learning (UL)** approaches, aimed at mitigating the **zero-generation** vulnerability. The dependence on data labels for known *deepfakes* has proved to be an unsustainable point of failure in an environment of rapid and continuous evolution of media synthesis technologies. The use of models such as **Autoencoders (AEs)** and **GAN-Based Anomaly Detection** systems establishes an inherently more resilient methodology, since its focus is on **modeling the statistical distribution of normality**, what is authentic and statistically coherent, and classifying as an anomaly any significant deviation, regardless of the *forgery* technique employed. This inversion of the problem, from searching for known flaws to tracking breaks in the coherence of reality, is the foundation of truly adaptive *Deepfake Forensics* systems.

The effectiveness of UL is inseparable from the **efficient integration of Computer Vision (CV)** as an extractor of **highly relevant forensic characteristics**. The success of the *pipeline* lies not in feeding the model raw pixels but in providing **isolated signals that are hard to manipulate**, such as **temporal coherence**, **frequency-domain anomalies**, and **physiological signals (rPPG)**. This intelligent extraction ensures that the UL model is learning the normality of **physical and biological invariants**, and not the cosmetic variations of appearance. The development of architectures such as **Spatio-Temporal Autoencoders (STAEs)** represents the state of the art of this synergy, enabling the detection of anomalies in the dynamics of movement and in the logical continuity of a video, which are artifacts notoriously hard to eliminate even for the most sophisticated generators. The STAE's ability to identify failures in the rPPG, for example, offers a layer of biometric authenticity that is crucial.

However, the transition to UL is not free of challenges, the most critical for practical application being the **calibration of the anomaly threshold**. The inherent sensitivity of UL models to non-adversarial variations (such as camera noise, different compression *codecs*, and environmental variations) can lead to an unacceptable rate of **false positives**, which would seriously compromise the credibility of the system on high-volume social media platforms. Future research should therefore focus on **training methods and filtering of normality data** that are rigorously representative of the real *deployment* environment, including a vast range of video degradations. The modeling of normality must be robust enough to disregard irrelevant noise but sensitive enough to capture the subtle deviations introduced by high-fidelity *deepfakes*.

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A fundamental future implication of this work is the imperative need to develop **hybrid, layered detection systems**. The ideal *Deepfake Forensics* model will be an architecture in which **UL acts as the first line of defense and screening**, rapidly processing millions of videos and isolating those that violate statistical normality. The *deepfake* candidates flagged by this UL module should then be automatically forwarded to a **secondary Supervised Learning (SL) module** that incorporates **Explainable Artificial Intelligence (XAI)** techniques. This cascade architecture combines the **robustness of zero-generation anomaly detection** with the **capacity for forensic justification and traceability** of XAI, overcoming the isolated limitations of each paradigm.

Research should also explore the use of **Unsupervised Conditional Adversarial Networks**, which can be trained to model not only the distribution of normal *features* but also to learn the **separation between the mapping of the normal latent space and the anomalous one** more efficiently than traditional AEs. The focus should be on refining the **anomaly score**, evolving from a simple reconstruction-error metric (MSE) to one that incorporates the **manifold distance** in the latent space. This makes detection more discriminative, distinguishing a genuine anomaly (*deepfake*) from a natural variation of the object (*outlier*).

In terms of **application in cybersecurity**, UL has a crucial role in **adaptive biometric identity verification**. A UL system can be continuously trained on the normality of an individual's biometric characteristics (voice, face, rPPG) to detect personalized *deepfake* attacks (such as *spear-phishing* carried out with a *deepfake* of the CEO). Robustness against zero-generation attacks ensures that, even if the attacker uses a state-of-the-art voice or face synthesis technique, the break in the statistical distribution of the target's physiological patterns will be detected as an anomaly, offering a layer of defense that adapts individually to the user.

The **quantification of risk and degree of anomaly** on social media platforms is another area of future impact. Instead of a binary classification, the output of UL, the anomaly score, can be used for **dynamic risk escalation**. Videos with the greatest deviation from normality could be flagged immediately for removal or priority human review, while videos with marginal deviations that still meet certain anomalous criteria could be flagged with a *disclaimer* notice. This approach gives platforms a **content moderation** tool based on statistical evidence and with superior decision granularity.

In summary, the evolution of *Deepfake Forensics* is no longer about achieving 100% accuracy on a fixed dataset; it is about building **resilience and adaptability** against the unknown. Unsupervised Learning provides the backbone of this resilience, allowing digital forensic science to focus on the **modeling and protection of statistical truth**. The strategic integration of Computer Vision and UL is the only sustainable path to ensuring that the defense system is always one step ahead of the curve of attack innovation, prioritizing the **capacity for adaptation** as the most critical performance metric for information security in the twenty-first century.

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

REFERENCES

Books and Articles

1. AN, Jihwan; CHO, Sungzoon. Variational Autoencoder based Anomaly Detection using Reconstruction Probability. Special Lecture on AI, 2015.

2. GOODFELLOW, Ian et al. Generative Adversarial Networks. In: Advances in Neural Information Processing Systems (NIPS), 2014.

3. LIDEN, Lars; TULBURE, Andrei; VAN DER HALLEN, Dries. Deepfake Detection using an Anomaly Detection Approach. In: International Conference on Digital Image Processing (ICDIP), 2021.

4. SCHLEGEL, Christian et al. f-AnoGAN: Fast Anomaly Detection with GANs. In: Medical Image Computing and Computer Assisted Intervention (MICCAI), 2019.

5. SUN, Shuheng et al. Deepfake Detection Based on Spatiotemporal Features and Unsupervised Learning. In: IEEE International Conference on Image Processing (ICIP), 2020.

6. WANG, Shiqing et al. Adversarial Training for Deepfake Detection: A Survey. arXiv:2104.09068, 2021.

7. XU, Weili et al. Deep Anomaly Detection: A Survey. arXiv:2003.02983, 2020.

8. YANG, Xin; SUN, Jianli; ZHOU, Songyuan. Learning to Detect Digital Forgeries with a Weakly Supervised Approach. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

9. ZHOU, Peng et al. Two-Stream Neural Networks for Tampered Face Detection. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

10. ROSSLER, Andreas et al. FaceForensics++: Learning to Detect Manipulated Facial Images. In: International Conference on Computer Vision (ICCV), 2019.

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.