RCMOS - Multidisciplinary Scientific Journal The Knowledge.
ISSN: 2675-9128. São Paulo-SP.
Year III, v.2 2023. | Submission: 10/21/2023 | Accepted: 10/23/2023 | Publication: 10/25/2023

Deepfake Forensics: Integrating Computer Vision and Machine Learning Techniques Supervised Training for Video Authenticity Analysis

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 an MSc in Data Science and Artificial Intelligence from the Université Côte d'Azur.

SUMMARY

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

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

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 GENERATION ZERO

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

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.), while the domain of *deepfakes* (the "fake" data) is **unlimited and mutable.** Supervised Learning fails because it attempts to learn the boundaries between a finite set of known forgeries and the real thing, leaving the system blind to any forgery that was not seen during training. For **Deepfake Forensics** systems aimed at **cybersecurity** and **information integrity,** this reliance on prior knowledge is a fatal flaw.

This article proposes exploring **Unsupervised Learning (USL)** as the fundamental solution to mitigate the vulnerability of generation zero. Instead of learning what is "fake," USL focuses on **modeling the distribution of "normality"** —that is, what is inherently **authentic** and statistically consistent. Any input sample that **deviates significantly** from this normality distribution (in terms of reconstruction, latency, or statistical patterns) is classified as **an anomaly** and, consequently, as a *deepfake.*
This approach reverses 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 focus on integrating advanced **Computer Vision** techniques to extract relevant forensic *features ,* which are then processed by ANS algorithms such as *Autoencoders* (AEs) and variations of Generative Adversarial Networks (GANs) trained for anomalies. The goal is to demonstrate the feasibility of building detection *pipelines* that are **inherently resilient** to the evolution of *deepfakes,* with immediate application in **verifying them.**

Content creation in social media and cybersecurity **forensics ,** where speed and robustness against the unknown are critical.

2. Autoencoders and the Modeling of the Normal Distribution

Autoencoders **(AEs),** a prominent class of neural networks in **Unsupervised Learning (ANS),** are the mainstay of anomaly detection in Computer Vision data and represent a powerful tool for *deepfake forensics.* The operating principle of 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 that data and then reconstruct it with the highest possible fidelity. The AE effectively learns the **statistical distribution of normality.**

The architecture of Advanced Experiments (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, AE becomes highly efficient at reconstructing **authentic faces and scenes.** However, when presented with a **manipulated video *(deepfake),*** which contains visual and statistical patterns that AE has never encountered (forgery *artifacts),* the network fails to accurately map the *deepfake* to its latent space and, consequently, fails in its reconstruction. The model, in attempting to reconstruct what it does not understand, generates an image with **a high reconstruction error.**

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

The great strength of AE in *Deepfake Forensics* is its **inherent robustness against zero-generation attacks.** Since it doesn't learn what is false, but rather 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 AE, however, lies in its susceptibility to *deepfakes* that are **extremely convincing,** that is, that fit very well into the normal distribution. 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 manifest.

RCMOS – Multidisciplinary Scientific Journal The Knowledge.
ISSN: 2675-9128. São Paulo-SP.

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 (ANS)** for anomaly detection. This approach, called *GAN-Based Anomaly Detection,* exploits the ability of a well-trained GAN's **Discriminator** to model the **normal data space** with high accuracy, becoming a statistical deviation detector with superior sensitivity compared to traditional AEs.

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

A powerful variation is **AnoGAN (Anomaly Detection with GANs),** which attempts 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 normal distribution, should have a corresponding *z-* point in the latent space of the Generator that reconstructs it with high fidelity. An anomalous image (the *deepfake)* should not have a *z-point* that reconstructs it well. The anomaly score is therefore a combination of the **reconstruction error** and the **latent mapping distance (the** distance between the found *z-point* and the training latent space).

This adversarial approach to ANS 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, resulting in greater sensitivity in detecting subtle deviations. Furthermore, just like AEs, GAN-based anomaly detection is **inherently resistant to zero-generation,** as the model is not looking for known *deepfakes ,* but rather for any sample that violates the statistical law of authenticity it has learned, making it extremely valuable for **cybersecurity** in a rapidly evolving context.

threats.

4. COMPUTER VISION INTEGRATION: *FEATURE* EXTRACTION FORENSIC

The success of **Unsupervised Learning (ANS)** in anomaly detection critically depends on the **quality of the *features*** extracted from the videos by **Computer Vision (CV).**

The application of ANS techniques in *Deepfake Forensics* is not limited to feeding the network with raw pixels; it is necessary to pre-process and isolate the areas of the video where manipulation artifacts are most likely to occur, ensuring that the ANS is modeling the normality of the most relevant forensic *features ,* and not noise or irrelevant context.

The first step in VC involves **Facial Localization 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. Subsequent alignment, based on reference points (such as eyes, nose, and mouth), standardizes the facial pose, minimizing non-manipulative variation and allowing the ANS model to focus on internal facial anomalies, such as skin texture and micro-expression.

Next, VC is employed to isolate critical **temporal and physiological signals .** One of the most difficult *deepfakes* to fake is the **Remote Photoplethysmography (rPPG)** signal , which is the variation in skin color caused by blood pulsation (the heartbeat). VC techniques can be used to extract this subtle temporal series of facial color. By feeding the AE or GAN with the **rPPG time series,** instead of just the static frame, the ANS learns the normal distribution 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.*

Furthermore, VC 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 possess. The ANS 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 are highly revealing of manipulation. This integration of VC and ANS 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 (ANS)** in overcoming the vulnerability of generation zero, its application in *Deepfake Forensics* faces significant practical and conceptual challenges that need to be addressed for its 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 reconstruction error distribution (in AEs) or the mapping distance (in GANs). However, natural variations in authentic videos – such as camera noise, compression, different skin tones, and lighting conditions – can, by themselves, generate a high reconstruction error.

RCMOS - Multidisciplinary Scientific Journal The Knowledge.
ISSN: 2675-9128. São Paulo-SP.

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

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

Finally, ANS does not provide the **explainability** that Supervised Learning with XAI can offer. Although it says "this video is an anomaly," it does not explain *the* nature of that anomaly. The AE may have a high reconstruction error, but the expert does not know if the flaw is in the rPPG mapping or in the hair texture. 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 ability to provide visual and expert justification.

6. Practical Application in Cybersecurity and Social Media

The integration of **Computer Vision** and **Unsupervised Learning (ANS)** in *Deepfake Forensics* has transformative implications for **cybersecurity** and **content control** on social media platforms, offering solutions to challenges that supervised models cannot address. 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 verifying the **authenticity of sensitive communications.** In a *spear-phishing* attack...
Using a *deepfake* voice or video (like the *CEO fraud),* an ANS 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 *deepfake* generation .

Zero, functioning as a **biometric authentication filter** that is vital for the protection of assets and confidential information.

For **social media platforms,** the challenge is the **massive volume and speed** of video dissemination. The ANS *pipeline* becomes a highly efficient **first-line screening layer .** *Autoencoders* and GAN-based Anomaly Models.

They can process millions of videos, identifying ***deepfake* candidates** that exhibit a high anomaly score. This is especially useful for dealing with the volatile nature of *deepfake generation,* where techniques are constantly changing. Once a video is flagged as anomalous by the ANS, it can be forwarded to a **secondary supervised system** (with XAI) and **human review** for expert confirmation.

The use of ANS in 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 *deepfakes* that deviate most from normality (the most bizarre or lowest quality, causing the most confusion) or those that pose the greatest threat (the most convincing ones that deviate only slightly, indicating a high level of *forgery)* .

(sophisticated). By continuously modeling normality, platforms can develop an **adaptive surveillance** system that is inherently more resilient and less dependent on blacklists of spoofing techniques.

7. INTEGRATION OF TEMPORAL SIGNALS AND SPATIO-TEMPORAL AUTOENCODERS

Deepfake *Forensics* requires **Unsupervised Learning (ANS)** 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 achieved by replacing 2D convolutional layers with **3D convolutional layers** (which operate in time x width x 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 in a sequence, the speed of blinking, and the continuity of optical flow.

The application of STAEs in anomaly detection is particularly effective in identifying ***face-swapping* artifacts** in videos. Such manipulations frequently introduce **inconsistencies in the transition of the facial mask between** frames, resulting in *flickering* or

Unstable edges that are difficult to see with the naked eye, but which violate the normality of smooth and continuous movement. STAE, when reconstructing the video, will generate a **high reconstruction error in the areas and moments of incoherent transition,** signaling the temporal anomaly.

Furthermore, modeling **physiological signals** through temporal integration is key to robustness. rPPG, the heart rate signal, is a time series. A STAE fed 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 unrealistic frequency (a common flaw in basic deepfakes) will be immediately classified as a **temporal and physiological anomaly,** providing evidence of falsity that is nearly impossible to replicate using current *forgery* techniques . Spatiotemporal modeling is therefore the most robust way to ensure the **resilience of *Deepfake Forensics*** against the increasing sophistication of video manipulation.

8. CONCLUSION AND FUTURE IMPLICATIONS

This study demonstrated, based on 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 (USL) approaches,** aiming to mitigate **zero-generation vulnerability.** The reliance on data labels for known *deepfakes* has proven to be an unsustainable point of failure in an environment of rapidly evolving media synthesis technologies. The use of models such as **Autoencoders (AEs)** and **GAN-based Anomaly Detection** systems establishes an inherently more resilient methodology, as its focus is on **modeling the statistical distribution of normality** – what is authentic and statistically consistent – and classifying any significant deviation as an anomaly, 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 for truly adaptive *Deepfake Forensics* systems .

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

However, the transition to ANS is not without challenges, with **anomaly threshold calibration** being the most critical for practical application. The inherent sensitivity of ANS 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 system's credibility on high-volume social media platforms. Therefore, future research should focus on **normality data training and filtering** methods that are rigorously representative of the real-world *deployment* environment , including a wide range of video degradations. Normality modeling must be robust enough to disregard irrelevant noise, yet sensitive enough to capture subtle deviations introduced by high-fidelity *deepfakes .*

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

The research should also explore the use of **Unsupervised Conditional Adversarial Networks** that can be trained to model not only the distribution of *features.*
The goal is not only to learn how to **distinguish between normal and anomalous latent space mapping** 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 latent space. This allows for more discriminative detection, distinguishing a genuine anomaly *(deepfake)* from a natural variation of the object *(outlier).*

In terms of **cybersecurity applications,** ANS plays a crucial role in **adaptive biometric identity verification.** An ANS system can be continuously trained on the normality of an individual's biometric characteristics (voice, face, rPPG) to detect **customized *deepfake* attacks** (such as *spear-phishing).*
(executed 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.

Quantifying **risk and the degree of anomaly** on social media platforms is another area of future impact. Instead of a binary classification, the ANS output – the anomaly score – could be used for **dynamic risk scaling.** Videos with the greatest deviation from normality could be immediately flagged for removal or priority human review, while videos with marginal deviations, but which still meet certain anomalous criteria, could be flagged with a disclaimer . This

This approach provides platforms with a **content moderation tool based on statistical evidence** and with greater decision-making granularity.

In short, the evolution of *Deepfake Forensics* is no longer about achieving 100% accuracy on a fixed dataset; it's about building **resilience and adaptability** against the unknown. **Unsupervised Learning provides** the backbone of this resilience, allowing digital forensics science to focus on **modeling and protecting statistical truth.** The strategic integration of **Computer Vision and ANS** is the only sustainable way to ensure that the defense system is always one step ahead of the attack innovation curve, prioritizing **adaptability** as the most critical performance metric for information security in the 21st century.

REFERENCES

Books and Articles

1. AN, Jihwan; CHO, Sunzoon. 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.