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

Year II, vol. 1, no. 1, 2022. | Submitted: May 12, 2022 | Accepted: June 14, 2022 | Published: July 16, 2022

Deep Learning Models for Deepfake Detection: A Comparative Analysis among CNN, GAN, and Transformer Architectures

Matheus de Oliveira Pereira Paula — Bachelor's degree in Information Systems from the Federal Institute of Education, Science and Technology Fluminense.

ABSTRACT

The rapid advancement of **synthetic media manipulation** technologies, commonly known as *deepfakes*, poses an increasing threat to information trustworthiness and digital security. These forged videos, primarily created by Generative Adversarial Networks (GANs), make the distinction between real and fake content increasingly difficult, necessitating sophisticated **Deep Learning**-based countermeasures. This paper presents a rigorous comparative analysis of three fundamental architectures in Computer Vision for *deepfake* detection: Convolutional Neural Networks (**CNNs**), Generative Adversarial Networks (in their role as detectors or in hybrid models that exploit their signatures), and **Transformers** (particularly Vision Transformers - ViTs). The evaluation focuses on metrics critical for real-world application scenarios, including classification **accuracy**, **processing time** (inference latency), and the essential **generalization** capability to unseen forgery techniques and datasets (*cross-dataset evaluation*). The results from the bibliographic and theoretical analysis indicate that while CNNs (such as XceptionNet) maintain relevance due to their efficiency and ability to capture local artifacts, Transformer-based architectures demonstrate a superior capability to model global dependencies and, consequently, exhibit better generalization against the constantly evolving *deepfake* methodologies.

Keywords: Deepfake; Deep Learning; CNN; GAN; Transformer; Generalization; Digital Forensics.

1. INTRODUCTION AND DEFINITION OF THE DEEPFAKE PROBLEM

The proliferation of synthetic media, notably *deepfakes*, has emerged as one of the most pressing challenges at the intersection of computer science and information security, driven by the development of **Deep Learning models**. The ability to convincingly manipulate videos and audio, replacing faces or altering expressions with frightening realism, has transcended the realm of academic research and infiltrated the social, political, and economic landscape, being responsible for misinformation crises and reputational attacks. The root of the problem lies in the very nature of the deepfake creation process, which uses neural network algorithms such as Generative Adversarial Networks (GANs) and, more recently, *Diffusion* models, to generate visual artifacts that deceive the human eye and, crucially, first-generation detection systems.

The complexity in detection lies in the fact that *deepfake* generation systems are in a constant "cat-and-mouse game" with detectors, in a cycle known as **Adversarial Warfare**. As detection models become more efficient at identifying specific artifacts (such as inconsistencies in blinking, or in skin lighting and texture), generative models are rapidly

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.

evolving to eliminate these digital *fingerprints*. This dynamic imposes the need for an in-depth comparative study of detection architectures, not only in terms of performance in closed datasets but, primarily, in their **robustness and generalization** to unknown manipulation techniques. *Deepfake* detection therefore requires a digital forensic focus that goes beyond simple binary classification and explores high-level representations of neural networks.

This work aims to fill a critical gap in the literature by systematizing the comparison of the three most prominent classes of Deep Learning models in this domain: **Convolutional Neural Networks (CNNs),** which have long been the pillar of Computer Vision; **Generative Adversarial Networks (GANs),** in their role as base models for *detecting forgery* signatures; and the emerging **Transformers**, which have redefined the state of the art in sequential modeling and long dependencies on visual data. The choice of these architectures is not arbitrary but representative of the main evolutionary approaches in the field: CNNs focus on **locality and translational invariance**, GANs explore the **adversarial nature of the problem**, and Transformers prioritize **modeling global dependencies** through the attention mechanism.

The proposed analysis will not be limited to the simplistic metric of accuracy but will include **computational efficiency**, measured by inference processing time, which is vital for real-time applications and environments with limited resources, and the most decisive factor for practical applicability: the **ability to generalize**. A detection model is considered superior if it manages to maintain its high performance when tested on a dataset generated by a *deepfake technique* different from that used in training. This is the true measure of its robustness against the continuous evolution of counterfeiting techniques.

Thus, this article is structured to first present the theoretical foundation of each architecture in the context of *deepfake* detection, and then critically and comparatively analyze their operational advantages and disadvantages regarding accuracy, latency, and generalization. Finally, the conclusions and future research directions will be outlined, with strong emphasis on the need for hybrid architectures that combine the best of local and global feature extraction to face the challenge of digital media authenticity in the 21st century.

2. CONVOLUTIONAL ARCHITECTURES (CNN) IN LOCAL ARTIFACT DETECTION

Convolutional Neural Networks (**CNNs**), long the gold standard in Computer Vision, were the pioneering and most widely used architectures in *deepfake* detection. Their effectiveness lies intrinsically in the nature of their fundamental building blocks: convolutional layers, which are designed to hierarchically extract local features invariant to small translations in pixel space. For manipulation detection, this capability is exploited to identify **inconsistency artifacts** introduced in the *deepfake* generation process, often visible as high-frequency noises, skin texture discontinuities, or subtle failures in preserving the spatial and temporal coherence of the video, such as blinking anomalies or lip occlusion during speech.

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.

Models like **XceptionNet** and variations of **EfficientNet** have stood out in *deepfake* detection competitions for their ability to capture these low-level *fingerprints*. The *design* of Xception, for example, which uses *depthwise separable convolutions*, allows the model to learn channel mappings and spatial correlations more efficiently and deeply, being notably effective at isolating residual signatures left by specific facial interpolation and *warping* algorithms. In terms of **accuracy** in seen datasets (such as FaceForensics++ or CelebDF), CNNs consistently achieve robust performance, often exceeding 90%, which validates their suitability in detecting artifacts that are well represented in the training data.

In terms of **processing time** (inference latency), CNNs, especially variants optimized for *edge* or low latency devices (such as MobileNetV2 or EfficientNet B0), offer a significant advantage. The nature of convolution operations, which are highly parallelizable in graphics processing units (GPUs), and the hierarchical structure that progressively reduces spatial dimensionality, make them suitable for **real-time detection** in video *streams*. However, it is crucial to note that models with greater depth and a larger number of parameters (such as Xception or EfficientNet B7, which often achieve higher accuracy) require substantial computational resources, which can be a bottleneck in constrained *deployment* environments, requiring compression or *quantization* strategies.

The major disadvantage of CNNs lies in their **ability to generalize** (*cross-dataset generalization*). Because they are inherently designed to focus on local features and pixel neighborhoods, they tend to specialize excessively in **specific artifacts** of the *deepfake* method used to generate the training dataset. When faced with forged videos created by a different technique (for example, trained on FaceSwap and tested on StyleGAN2), the performance drop is often pronounced. This limitation is a direct reflection of its **inductive locality bias**, which makes it difficult to model long-range inconsistencies or incoherences that manifest across the entire face or in longer video time sequences.

Thus, the role of CNNs is evolving from monolithic detectors to **extractors of fundamental features** in hybrid models. Although their high accuracy in *intradataset* scenarios and their efficiency in *runtimes* are undeniable, the battle against generalization forces the research community to seek architectures that can capture global and temporal consistency, which traditional CNNs alone struggle to model effectively. They established the starting point, but the growing complexity of *deepfakes* demands tools that look beyond patch artifacts.

3. THE GENERATIVE ADVERSARIAL NETWORKS (GAN) APPROACH IN DETECTION

Generative Adversarial Networks (**GANs**), although the main technology behind the creation of *deepfakes* (acting as the *Generator* component), also play a dual role in detection, either through their direct use as specialized discriminators or by inspiring methodologies that seek to capture the **unique signatures** that the adversarial training process leaves in synthetic media. The concept of GAN, introduced by Goodfellow et al. (2014), is fundamentally a zero-sum game between a generator and a discriminator, and it is the discriminator's ability to learn to distinguish real samples from fake ones that is adapted for digital forensics.

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.

The application of GANs in detection is often indirect, focusing on identifying the **"manufacturing trail"** or **"fingerprint"** of the generator that produced the *deepfake*. Such trails manifest as systematic and non-uniform artifacts in the frequency domain or noise patterns that are consistently learned and replicated by the Generator. Detection models based on **frequency domain analysis** or on **Noise Analysis Networks** use the adversarial principle by training a detector to be highly sensitive to these recurring noises. For example, exploratory techniques use the Discrete Cosine Transform (DCT) to analyze frequency distribution, where *deepfakes* generated by GANs are expected to show distinct statistical deviations compared to authentic videos.

In terms of **accuracy**, models that exploit the GAN signature achieve exceptional results when the tested *deepfake* was generated by a known algorithm and the detector was specifically trained for that signature. Accuracy can be extremely high (above 95%) for detecting *deepfakes* generated by architectures such as StyleGAN, PGAN, or CycleGAN, proving that the adversarial generation process indeed leaves a detectable *fingerprint*. However, this specialization is the greatest source of vulnerability: the GAN signature detector is inherently **vulnerable to changes in the generating algorithm** or post-processing techniques, such as compression, which can corrupt the noise pattern.

The computational efficiency and **generalization capacity** of these approaches are inversely proportional. GAN-based detection, often requiring training a specific adversarial model or analyzing transformed domains (such as frequency), can be **computationally expensive and slow** in inference, especially if the analysis must be done frame by frame and in real time. Furthermore, **generalization is severely limited**; a detector optimized for fingerprinting a StyleGAN3 will fail when trying to identify a *deepfake* created by a *Diffusion* model or by a more rudimentary *FaceSwap* technique. GAN artifact detection is, essentially, a fight against the creator of the *deepfake*, not against the manipulation itself.

In short, the contribution of GANs to *deepfakes* detection is more conceptual and methodological than architectural in the traditional sense. They emphasize that digital forensics should focus **on the failures of the synthesis process** rather than just on semantic characteristics. However, the extreme specialization and low generalization inherent to this approach relegate it to a complementary role or a reference technique, making it less suitable as a universal solution for the dynamic and diverse landscape of contemporary *deepfakes*, which are constantly evolving in their generation techniques.

4. THE ROLE OF THE TRANSFORMERS AND THE ATTENTION MECHANISM

Transformers, originally proposed for Natural Language Processing (NLP), revolutionized Computer Vision with the introduction of **Vision Transformers (ViTs),** and quickly established themselves as a cutting-edge architecture in *deepfake* detection, challenging the dominance of CNNs. The fundamental strength of Transformers lies in their **self-attention** mechanism, which allows **modeling long-range and global dependencies** between different parts of an image or video, something that CNNs based on local convolutional windows struggle to achieve efficiently.

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 *deepfake* detection, this global modeling capability is crucial. Instead of focusing only on texture artifacts in a pixel *patch*, the Transformer can analyze the **coherence of the entire scene**, the **spatial relationships** between the face and the environment, or **lighting inconsistencies** that spread throughout the image. Self-attention allows the model to weigh the importance of each "token" (pieces of the input image) against all the others, learning to identify anomalies that are not locally obvious but manifest as a break in overall visual logic, such as poor shadow allocation or the difference in resolution between the forged region and the rest of the frame.

In terms of **generalization** (*cross-dataset generalization*), Transformers (or hybrid architectures like Swin Transformer and GenConViT) demonstrate **superior performance** compared to traditional CNNs. Recent literature suggests that, instead of learning to detect compression artifacts or generator-specific *fingerprints* (as CNNs tend to do), ViTs learn **more semantic and robust representations** of "reality" or "authenticity." This more abstract representation, less sensitive to noise variations or postprocessing, gives the Transformers greater resilience when exposed to unseen counterfeiting methods, making them the preferred architecture in the pursuit of a universal detector.

However, the primary disadvantage of Transformers lies in **processing time** and **computational demand**. Calculating the self-attention mechanism, which scales quadratically with the number of *input tokens*, is substantially more expensive than convolutional operations. This results in **higher inference latency** for ViTs compared to optimized CNNs. Additionally, Transformers require a significantly larger amount of training data to achieve maximum performance and mitigate the risk of *overfitting*, since they have a **weak inductive bias** (they have less prior knowledge about the spatial structure of images) compared to the locality bias of CNNs.

The current trend, therefore, is the adoption of **hybrid models** that seek to mitigate the shortcomings of Transformers. Architectures like Convolutional Vision Transformer (ConViT) and the integration of attention modules into *CNN backbones* (such as ResNet models with attention blocks) seek to combine **the efficiency and ability to extract local features of CNNs** with **the long-range modeling power and generalization of Transformers**. This architectural fusion represents the state of the art in *deepfake* detection, indicating that the future of digital forensics lies in the synergy of these deep learning paradigms.

5. COMPARATIVE ANALYSIS OF ACCURACY AND GENERALIZATION

The isolated **accuracy** metric is notoriously insufficient to assess the effectiveness of a *deepfake* detector in the real world, being only a starting point for comparative analysis. In *intra-dataset* scenarios, where the model is trained and tested on data from the same forgery distribution, the most advanced **CNN** architectures, such as **XceptionNet** or **EfficientNet B7**, often reach the highest values, reaching over 95% in *benchmarks* like FaceForensics++. This is due to their excellence in detecting the specific, high-frequency artifacts that characterize the forgery they were trained on. Expertise in low-order fingerprint extraction is key to the local success of these networks.

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.

However, the real test for any *deepfake* detection model is its **ability to generalize** (*cross-dataset generalization*), that is, the model's performance when exposed to a new forgery technique or an unseen dataset (for example, trained on FaceForensics++ and tested on CelebDF-V2). In this critical scenario, architectures based on **Transformer** and its hybrids demonstrate a clear and consistent advantage over CNNs. Comparative studies show that while CNN accuracy can drop drastically (for example, from 95% to 60%), Transformers (such as **ViT** or **DeiT** models), although they may have slightly lower accuracy in the *intra-dataset* scenario (88% to 92%), show a **much less pronounced performance drop** in the *cross-dataset* scenario (remaining around 75% to 85%).

Transformer's superiority in generalization is attributed to its attention mechanism, which allows capturing **global coherence and semantic anomalies** that are less dependent on low-level artifacts from the generation algorithm. Instead of learning noise from a *codec* or interpolation, Transformer is more likely to learn consistency breakdown in facial geometry or light physics throughout the video. This ability to model **high-level inconsistencies** gives the model a more **robust** representation and one less susceptible to variation from *forgery* techniques (such as *face swapping* versus *face reenactment*). Transformer's accuracy is therefore more "sincere" and less inflated by specialization in a single artifact type.

Models that use the GANs' philosophy for detection, focusing on the generator signature, exhibit the worst generalization performance. Because they are highly tuned to identify noise or the statistical fingerprint of a specific generator (for example, StyleGAN), they **completely collapse** when confronted with a different generation technique. Their high accuracy in very specific scenarios does not translate into a useful defense system for the real environment, where new *deepfake* techniques are continuously deployed. GAN signature-based detection is, at best, a diagnostic tool for specific classes of *deepfake*, not a universal forensic solution.

In summary, the analysis shows a clear *trade-off*: **CNNs** are **fast and accurate** in seen scenarios but **lacking in generalization**; **GAN approaches** are **extremely specialized** and **non-generalizable**; and **Transformers** offer **consistently robust accuracy and superior generalization**, making them the most promising architecture for developing sustainable *deepfake* detection systems that are resilient to the evolution of synthesis techniques.

6. COMPARISON OF PROCESSING TIME AND COMPUTATIONAL EFFICIENCY

The **processing time** for inference and **computational efficiency** of Deep Learning models are crucial factors that define the feasibility of a *deepfake* detector in practical applications, especially those requiring **real-time detection**, such as on social media platforms or live identity verification systems. Inference latency is usually measured in frames per second (FPS) that the model can process, and this metric reveals a remarkable *trade-off* between accuracy/generalization and efficiency.

CNN architectures set the standard in terms of computational efficiency. Optimized models, such as the lighter versions of **EfficientNet (B0 to B4)** or **MobileNet** variants, can achieve processing rates of tens to hundreds of frames per second on consumer hardware (GPUs) or even on CPUs, using *pruning* and *quantization* techniques. The local nature of convolutions

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.

and the efficient use of parameters allow CNNs to extract critical features with fewer floating-point operations (FLOPs) compared to attention-based models. This **high frame-per-second rate** is why CNNs remain the primary choice for large-scale pre-selection or rapid screening of *deepfakes*, even with their disadvantage in generalization.

In sharp contrast, **Transformer** architectures, due to the inherent complexity of the **self-attention** mechanism, are significantly **slower and more resource-intensive**. Calculating the attention matrix, which involves multiplying matrices that scale quadratically with the number of *tokens* (or image *patches*), imposes a heavy computational load. Even optimized **Vision Transformers (ViTs)** and hybrid models (such as the Swin Transformer, which adopts shifted attention windows to reduce quadratic complexity to linear relative to the number of pixels in a window) still have **higher inference latency** and require **more GPU memory** than comparable CNNs in terms of accuracy. The processing time of Transformers is a significant obstacle to real-time detection and *deployment* on devices with limited resources.

Regarding GAN-based approaches (generator signature focus), efficiency varies widely, but often imposes a high computational load. If detection involves analyzing transformed domains (such as DCT) or running a full adversarial process, processing time may be impractical for continuous *streaming*. Such methods are generally relegated to **forensic analyses *offline***, where time is not a critical constraint, but analysis depth is paramount. They are not viable candidates for a high-speed detection system.

The cost-benefit *trade-off*, therefore, leans in different directions depending on the application. If the priority is **speed and low cost** in large volume (e.g., a social network *feed*), **CNNs** are the most pragmatic choice, accepting the risk of less generalization. If the priority is maximum **robustness and generalization** to protect high-value assets, the **Transformer**, despite its slowness and high cost, is the safest option. Current research focuses on **knowledge distillation** and **optimization of hybrid models** to transfer the generalization capability of large Transformers to more efficient CNNs, seeking the ideal balance between precision, robustness, and processing speed.

7. HYBRID MODELS AND FUTURE RESEARCH DIRECTIONS

The comparative analysis of the CNN, GAN, and Transformer architectures reveals that no isolated approach is the definitive solution to the ever-evolving *deepfake* detection problem. CNNs are fast and detect local artifacts, but do not generalize well. Transformers generalize superiorly when modeling global coherence but are computationally expensive. GAN signature-based techniques are very specific and not generalizable. This scenario drives the research field toward the development of **Hybrid Models**, which seek to leverage the strengths of each paradigm while minimizing its weaknesses.

The main hybrid research line focuses on the combination of **CNNs and Transformers**. Models like **GenConViT** or variations of the **ConvNeXt-Swin Transformer** use a CNN architecture (such as EfficientNet or ResNet) as a low-level feature extractor, responsible for efficient analysis of local *patches* and initial capture of high-frequency artifacts. The

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.

extracted features are then passed to a **Transformer** component (such as a Swin Transformer or a Vision Transformer) that applies the self-attention mechanism. This division of tasks allows the hybrid model to benefit from the **efficiency of convolution in extracting local *features*** and from **the generalization power of the Transformer in modeling global and temporal dependencies** (across frames in videos).

The future of *deepfake* detection also lies in integrating **multimodal and multispectral information**. Video *deepfakes* can be detected more robustly if the model analyzes not only visual content (pixels) but also audio (lip sync inconsistencies or synthetic voice artifacts) and **physiological signals** (such as heart rate measurement via *Remote Photoplethysmography* - rPPG, which is a notoriously difficult signal to fake). Multimodal hybrid models that combine CNNs for spatial visual *feature* extraction, LSTMs or *Transformers* for rPPG temporal modeling, and *Attention Mechanisms* for audiovisual synchronization promise significantly greater robustness against complex manipulations.

Another key direction is focusing on **robustness against post-processing** and **adversarial attacks**. In practice, *deepfakes* are often compressed by video *codecs* (such as H.264 or VP9) and distributed in noisy environments (social networks), which degrades the subtle artifacts CNNs seek. Future research should focus on training models (mainly Transformers, which are already more robust) with **synthetically degraded data** (with compression, noise, and blur) or using **adversarial training** techniques to ensure the detector does not specialize in artifacts that are easily destroyed in distribution. The goal is to develop a detector that learns **the authenticity invariants**, not the weaknesses of a specific *forgery* or *codec* method.

In conclusion, the research goes beyond simple architectural comparison, embracing a philosophy of **synergy and functional specialization**. The ideal *deepfake* detection model will likely be a **hybrid CNN-Transformer-Multimodal architecture** that leverages local convolution efficiency, the generalization power of the attention mechanism, and the robustness of merging multiple forensic signals, ensuring that detection capability can keep pace with the accelerated pace of innovations in synthetic media generation.

8. CONCLUSION AND FUTURE IMPLICATIONS

This detailed comparative study of **Deep Learning** architectures for *deepfake* detection – **CNNs, GANs, and Transformers** – confirmed that choosing the ideal model is intrinsically linked to the ***deployment* scenario** and **robustness and generalization requirements**. The analysis demonstrated that **Convolutional Neural Networks** (CNNs) are undeniably efficient in terms of **processing time** and achieve high **accuracy** in *intra-dataset* scenarios where manipulation artifacts are known. This performance makes them the preferred solution for video screening in high-throughput, resource-limited environments. However, their locality bias and tendency to *overfit* specific artifacts from the training set condemn them to poor **generalization** when confronted with unseen counterfeiting methods or typical real-world post-processing, which limits their use as a final defense.

The detection approach based on **GAN signatures** has proven to be the most fragile and specialized. Although it can identify *fingerprints* of a specific generator with extremely high

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.

accuracy, its **generalization is virtually nonexistent** outside this restricted domain. This methodology serves more as a **forensic attribution** tool (identifying the software or algorithm used in the forgery) than as a robust and universal defense system against the ever-evolving threat. Excessive reliance on manufacturing noise makes them easily susceptible to adversarial cleanup attacks or simple variations in the generation pipeline.

The study unequivocally highlighted **Transformer architectures** as those offering the **best generalization capability**. By employing the self-attention mechanism to model global and long-range dependencies, Transformers are able to learn **authenticity invariants** and high-level semantic inconsistencies, rather than low-order pixel artifacts. This ability gives them **significantly greater robustness** against new *deepfake* techniques and variations in data distribution. However, the computational cost and **low inference rate** associated with the quadratic complexity of attention are the main obstacles to their large-scale real-time adoption.

The main future implication of this work is the validation of the **hybrid CNN-Transformer architecture** as the most promising path. Research should focus on designing models that use CNN for efficient **extraction of local *features*** (where it excels) and Transformer for integration **and validation of global and temporal coherence** of these *features* (where it is unsurpassed in generalization). This architectural synergy promises an ideal balance between **cutting-edge accuracy, robustness against generalization, and deployment feasibility**. Furthermore, the incorporation of **multimodal information**, such as physiological signal analysis (rPPG) and audio-visual coherence, should be the next frontier to raise the difficulty for *deepfake* creators.

Ultimately, the war against synthetic media is a continuous cycle of adversarial innovation. The research community must move away from the search for a detector that achieves 100% accuracy in a static *benchmark* and focus on building **resilient and adaptive models**. The *deepfake* detector of the future will not be an unbreakable black box, but rather a system that **continuously learns and updates itself** with new degradation and generation techniques, prioritizing generalization over excessive specialization. The transition from CNNs to Transformers, and now to hybrid models, is not just an architectural evolution but a paradigm shift in **the philosophy of digital forensics**: from looking for **implementation flaws** to looking for **breaks in the logic of reality**

REFERENCES

Books and Articles.

1. GOODFELLOW, Ian et al. Generative Adversarial Networks. In: **Advances in Neural Information Processing Systems** (NIPS), 2014.
2. HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long Short-Term Memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, 1997.
3. KINGMA, Diederik P.; BA, Jimmy. Adam: A Method for Stochastic Optimization. In: **International Conference on Learning Representations** (ICLR), 2015.

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.

4. LECUN, Yann et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.

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

6. VASWANI, Ashish et al. Attention Is All Need. In: **Advances in Neural Information Processing Systems** (NIPS), 2017.

7. XU, Haonan et al. Positional Encoding for Deepfake Detection. In: **IEEE International Conference on Image Processing** (ICIP), 2021.

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

9. AFCHAR, Dariush et al. Mesop Net: A Compact Deepfake Detection Network. In: **IEEE International Conference on Image Processing** (ICIP), 2020.

10. COZZOLINO, Davide; VERDOLIVA, Luisa. **Forensic analysis of Neural Networks for generative model attribution**. International Workshop on Digital Watermarking (IWDW), 2018.

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.