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

Year III, v.1 2023. | Submission: April 10, 2023 | Accepted: April 12, 2023 | Publication: April 14, 2023

Explainable AI (XAI) in Deepfake Detection: Transparency and Interpretation in Computer Vision Models

Explainable AI (XAI) in Deepfake Detection: Transparency and Interpretation in Computer Vision Models

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 increasing sophistication of *deepfakes* has heightened the urgency for robust **Deep Learning** detection systems. However, the "black box" nature of Computer Vision models, such as Convolutional Neural Networks (CNNs) and Transformers, represents a significant obstacle to their acceptance in critical domains such as **forensics and law.** This article explores the application of **Explainable Artificial Intelligence (XAI)** techniques in the context of deepfake detection , investigating how the **transparency and interpretability** of the models can be achieved. *Post-hoc* and *intrinsic methodologies,* such as **CAMs (Class Activation Maps), SHAP, and LIME,** will be discussed , analyzing their ability to generate visual and logical evidence about the classification process, specifically identifying the regions of the image or video (artifacts) that are determinant for the falsity decision. The primary objective is to demonstrate that XAI integration is indispensable for building the necessary trust in detection systems, transforming algorithmic decisions into **verifiable expert evidence,** essential for establishing the validity and admissibility of these technologies in courts and investigations.

Keywords: XAI; Deepfake; Interpretability; Transparency; Digital Forensics; Computer Vision; Trust.

ABSTRACT

The increasing sophistication of *deepfakes* has escalated the urgency for robust **Deep Learning** detection systems. However, the "black-box" nature of Computer Vision models, such as Convolutional Neural Networks (CNNs) and Transformers, poses a significant barrier to their acceptance in critical domains such as **forensics and law.** This paper explores the application of **Explainable Artificial Intelligence (XAI)** techniques within the context of deepfake detection, investigating how model **transparency and interpretability** can be achieved. We discuss *post-hoc* and *intrinsic* methodologies, such as **CAMs (Class Activation Maps), SHAP, and LIME,** analyzing their capacity to generate visual and logical evidence regarding the classification process, specifically identifying the regions of the image or video (artifacts) that are decisive for the forgery decision. The primary objective is to demonstrate that the integration of XAI is indispensable for building the necessary **trust** in detection systems, transforming the algorithmic decision into **verifiable expert evidence,** which is essential for establishing the validity and admissibility of these technologies in courts and investigations.

Keywords: XAI; Deepfake; Interpretability; Transparency; Digital Forensics; Computer Vision; Trust.

1. INTRODUCTION: THE IMPERATIVE OF EXPLANABILITY IN FORENSICS DIGITAL

The rapid advancement of *deepfake* creation methods has set a new standard of challenges for society, demanding equally sophisticated **Artificial Intelligence (AI)** countermeasures. **Computer Vision models,** such as complex deep Convolutional Neural Networks (CNNs) and emerging *Vision Transformers* (ViTs), have demonstrated high accuracy in distinguishing between authentic and falsified media. However, this high performance is achieved at the cost of **transparency,** resulting in a "black box" architecture where the final decision is a mathematical mystery, inaccessible even to its developers. This opacity is at the heart of the problem when *deepfake* detection is applied in high-responsibility contexts, such as **forensic investigations and legal litigation,** where a simple declaration of "false" by an algorithm is insufficient to be considered evidence.

In a legal setting, expert evidence must be **verifiable, replicable, and, above all, logically justified.** The reliability of a *deepfake* detection system is called into question if the expert cannot clearly and intuitively present to the judge *which* characteristics of the video led the model to classify the content as forged. If CNN is focusing on a compression artifact instead of an inconsistency in facial geometry, the decision may be erroneous, but the opacity of the model prevents this critical distinction. Therefore, **Explainable Artificial Intelligence** (XAI) emerges not only as an improvement but as a **methodological imperative** to ensure the **admissibility and validity** of *deepfake* detection results in any process that requires a rigorous standard of...

proof.

The central objective of this article is to provide a theoretical and practical framework for integrating XAI techniques into the deepfake detection pipeline . We will analyze how key explainability methodologies can be adapted to the forensic domain, transforming the detector's binary output *(real* or *fake)* into an **expert report with visual and weighted evidence.** The relevance of this topic lies in the need to build a bridge between the **predictive capacity** of advanced AI and the **requirement for rational justification** in legal systems, promoting a level of **trust and traceability** that is currently nonexistent in traditional opaque models.

The research will delve into the differentiation between *model-agnostic* and model-specific explainability techniques , evaluating their applicability in identifying *deepfake* artifacts.

which vary in subtlety and location (from pixel micro-artifacts to global illumination inconsistencies). The goal is to strengthen detection, ensuring that the model is not "cheating" by learning spurious correlations (such as watermarks from training datasets or metadata) instead of the true signatures of manipulation. Ultimately, the adoption of XAI transcends mere algorithmic interpretation; it is the foundation for building a **responsible and ethically grounded digital defense system.**

2. The Nature of the "Black Box" and the Challenge of Reliability FORENSIC

The rise of **Deep Learning** in Computer Vision tasks has been marked by a paradigm shift: exponential gains in accuracy have been achieved at the cost of **interpretability.** Deepfake detection models , such as deep CNNs (e.g., ResNet, XceptionNet) and ViTs, are composed of millions or even billions of parameters arranged in complex layers, where non-linear transformations make the mapping from input (the video) to output (the classification) inherently opaque. This opacity, known as the **"black box" problem,** generates fundamental distrust in any environment where system failure could have serious and irreversible consequences, with the legal system being the most prominent example.

In a forensic context, the **reliability** of digital evidence depends on its **traceability and justification.** When an expert presents a report alleging that a video is a *deepfake,* this conclusion cannot be based solely on the 98% probability of falsity result issued by an unknown algorithm. The judge, lawyers, and jury need to understand **why ;** the model must point out, with spatial and temporal precision, the characteristics (or lack thereof) that corroborate the manipulation thesis, transforming the probabilistic prediction into **causal evidence.** The absence of this explicit justification prevents the adversarial process and technical criticism of the evidence, violating basic principles of due process.

The black-box challenge in *deepfake* detection is exacerbated by the adversarial nature of the problem itself. Without XAI, it is impossible to guarantee that the model is not focusing on **spurious artifacts** that accidentally correlate with the "fake" class in the training dataset. Examples of this include models that learn to identify the **invisible watermark** of a *codec* or the **edges of the bounding box** used in the face swapping phase , instead of the flaw in skin texture or shadows. Such correlations are fragile and break down under minimal variation, rendering the model useless for generalization and completely inadmissible as reliable evidence, since the cause of the classification is not the *deepfake,* but a production artifact of the *dataset.*

Therefore, the implementation of XAI in *deepfake* detection is an **ethical and legal necessity.** which transcends mere academic curiosity. It is the only way to mitigate opacity, allowing the forensic community and the judiciary to audit the algorithm's decision-making process. By transforming the black box into a **glass box,** where the nuances of the decision become apparent.

As these findings are evident, XAI not only validates the high accuracy of Deep Learning models, but also elevates their decision from a technical suggestion to a piece of **expert evidence with high probative value,** essential for combating disinformation in the age of synthetic media.

3. ***Post-hoc*** **XAI Techniques : Activation Maps and Resource Allocation**

Post-hoc **XAI** techniques are the most common group of explainability methodologies, as they can be applied to any already trained *deepfake* detection model without needing to modify its internal architecture. These techniques focus on measuring the **contribution and influence** of each input *feature* (pixels or pixel regions) on the model's final prediction. In the Computer Vision domain, two main categories dominate: **Class Activation Maps (CAMs)** and Feature **Attribution** methods .

Class **Activation Maps (CAMs)** and their variations (such as Grad-CAM and Grad-CAM++) are visual tools that provide a ***heatmap*** **over** the input image, highlighting the regions most relevant for classification. In a *deepfake detection context,* a successful Grad-CAM should focus on areas containing manipulation artifacts, such as **facial fusion lines, inconsistencies in the eyes/mouth, or discrepancies in lighting and skin texture.** If the model is correctly focused, the heatmap should concentrate on the *forgery points.* This visualization is extremely valuable for forensics, as it transforms an abstract numerical decision into **intuitive and spatially localized visual evidence,** allowing the expert to identify whether the model is detecting the actual manipulation flaw or irrelevant background noise.

In parallel, **Feature Assignment** methods , such as **SHAP (Shapley Additive exPlanations)** and **LIME (Local Interpretable Model-agnostic Explanations),** provide a **more numerical and causal justification.** SHAP, based on Shapley's game theory, calculates the marginal contribution value of each *pixel* or image *patch* to the probability of falsity. It provides a set of **Shapley values** that quantify the exact weight of each region in the final classification, ensuring that the explanation is **globally consistent and locally faithful.** LIME, in turn, builds a locally interpretable linear model (close to the data point being explained) to approximate the complex behavior of the black-box model.

The applicability of these *post-hoc* methods in *deepfake* detection lies in their ability to **audit and validate model bias .** If a SHAP or Grad-CAM consistently assigns high importance to areas outside the manipulated facial region (e.g., the watermark in the corner of the video), this indicates that the model has *overfitted* to a spurious *dataset* artifact . The limitation of these methods is that they provide only a **local explanation** (the explanation is only accurate for the video under analysis, and not for the overall behavior of the model) and can be **computationally expensive** (like SHAP), which impacts the real-time analysis of long videos. However, for the preparation of *offline* expert reports , the richness and depth of these methods are significant.

These causal explanations are an invaluable resource, establishing the basis for technical evidence in court.

4. Intrinsic Techniques and the Transition to Inherently Interpretable Models

Although *post-hoc* methods are essential for auditing existing models, the research trend in XAI points towards the creation of **inherently interpretable models** or **intrinsic techniques.** These approaches aim to incorporate the explainability mechanism directly into the neural network architecture, ensuring that the final decision is logically traceable by *design,* and not by retrospective analysis. The search for intrinsically interpretable models is a direct response to the inherent weakness of *post-hoc* methods , whose explanation is always an approximation and can be misleading if the underlying model is extremely complex.

In the field of Computer Vision for deepfake detection , this manifests itself in the use of architectures that explicitly learn to **separate semantic representation from artifact representation.** A theoretical example is the use of **Prototype-Based Models,** where classification is not done by an opaque decision boundary, but by similarity to visual "prototypes" stored in a latent layer, representing canonical examples of real and fake faces. The explanation for classifying a fake video would then be the simple presentation of the most similar *deepfake* prototype to which the video resembles, providing a highly intuitive **justification by analogy** for humans.

The greatest contribution to intrinsic interpretability in *deepfake* detection comes, ironically, from the **Transformer** architecture and its **self-attention** mechanism . The attention map naturally generated by the Transformer, which indicates how the model weights different *patches* of the image to construct its representation, can be directly interpreted as a **relevance map.** While not a complete causal explanation in the SHAP sense, the attention map is an intrinsic tool that reveals where the model is "looking." If the detection model is focused on lighting anomalies or the edges of the facial mask applied in the *deepfake,* the attention matrix will reflect this weighting concentration.

The great advantage of intrinsically interpretable models is the **inherent confidence** they provide, since the explanation is not a byproduct but an integral part of the inference process. This simplifies *deployment* in forensic environments, as the expert report can be based on a process that is, by definition, transparent. However, the design of these architectures is a challenge, as the constraint on interpretability can, paradoxically, **limit the predictive capacity** of the model, forcing a trade-off between performance and transparency. The research field relentlessly seeks to mitigate this trade-off, developing models that are simultaneously **highly accurate and perfectly traceable** in their decision-making logic.

5. XAI IN COMBATING BIAS AND VALIDATING GENERALIZATION

The strategic use of **Explainable Artificial Intelligence (XAI)** in *deepfake* detection . It transcends the mere justification of a single classification; it becomes a critical tool for **validating generalization** and **combating algorithmic bias.** As discussed in comparative studies, the main weakness of *deepfake* detection models is their tendency to **fail on unseen data** *(cross-dataset generalization),* which is often a symptom that the model has learned spurious correlations in the training data. XAI provides the necessary microscope to identify and correct this undesirable behavior.

By applying techniques like Grad-CAM or SHAP on a large scale to a validation dataset, experts can perform a feature **audit .** If the model, across all videos classified as "fake," is consistently focusing on metadata or a uniform *compression artifact ,* this reveals a dataset **bias .** The detector is not learning to manipulate the concept of facial falsification, but rather to detect the fingerprint of the training dataset creation process. This discovery, facilitated by XAI, allows researchers **to retrain the model** using **data augmentation** techniques.

more robust methods (such as introducing different levels of compression) or forcing the model to ignore areas of spurious artifacts through **attention** *masking.*

Validating **generalization capacity** also benefits immensely from XAI. When a model trained on a *dataset* (e.g., FaceForensics++) shows a drop in accuracy on an unseen *dataset* (e.g., CelebDF-V2), XAI can diagnose the cause of this failure. If the XAI explanations of the "fake" videos in the original *dataset* focus on a low-frequency artifact, but the model fails to produce a coherent *heatmap* for the fake videos in the new *dataset,* the evidence is clear: the *deepfake* in the new dataset does not possess the same artifact, and the old model failed to generalize to a more complex **semantic manipulation feature .**

The continued use of XAI in the development pipeline, therefore, transforms into a ***feedback* and correction mechanism .** It ensures that detection models are trained to focus on **falsity invariants** – the logical and physical flaws that are difficult to eliminate for any generator, such as inconsistency in global illumination or error in mapping the 3D geometry of the face – rather than easily eliminable implementation artifacts. This approach not only increases **robustness** against unseen *deepfakes* but also underpins the **scientific reliability** of the model.

6. Forensic and Legal Implications of the Model's Transparency

The transition from opaque detection to **Explainable Artificial Intelligence (XAI)** has profound and transformative implications for the forensic and legal spheres. The main barrier to the adoption of AI systems in courts is not a lack of accuracy, but the inability to...

To meet the **scientific proof standard,** which requires the methodology to be auditable, understandable, and capable of being refuted by an expert from the opposing party, XAI is the tool that unlocks this admissibility, transforming the algorithm's decision into **robust expert evidence.**

In a legal proceeding, the **model's explanation** becomes the **expert report itself.** A Grad-CAM heat map, for example, can be attached to the report to visually demonstrate that the model based its "false" classification on the interpolation failure around the jaw (a common area for artifacts) and not on a random blemish in the background of the video. Similarly, Shapley values can be used to quantify the **magnitude of the falsity artifact.**
In terms of the probability of manipulation, providing a numerical measure of the certainty of the decision that is transparent and logically sound. The absence of this explicit statement leaves the evidence vulnerable to being classified as "black box evidence" and therefore inadmissible.

The transparency generated by XAI also facilitates the **adversarial process and the *Daubert Standard*** (in jurisdictions that use it), where scientific evidence must be testable, peer-reviewed, and have a known error rate. With XAI, the opposing party's lawyer can analyze whether the model is focused on spurious or irrelevant artifacts, allowing for **technical criticism** of the evidence presented. This ability to audit the algorithm's bias is essential to ensure the **impartiality** of the justice system. Without explainability, the only way to refute it is to claim that the model "simply erred," which is not a valid scientific critique.

Furthermore, XAI is crucial in **assigning responsibility and in the ethics of AI.** By justifying classification, the XAI explanation can, in theory, help identify the **type of artifact** left by a specific generator (GAN, Autoencoder, Diffusion), aiding in traceability and tracing the origin of the manipulation. In a future where AI legislation will be more stringent, the ability of a *deepfake* detection system to **self-explain** will be a **normative requirement,** not just a desirable feature.

7. Integration of XAI in Real-Time *Deployment* Environments

Although **Explainable Artificial Intelligence (XAI)** techniques are academically robust, their integration into real - **time** *deepfake* detection environments (such as *streaming* platforms or social networks) presents significant computational challenges.
Most of the more informative *post-hoc* explainability methods , such as SHAP and LIME, require a **high computational cost** because they rely on evaluating multiple input perturbations or running a large number of iterations to estimate the contribution of each *feature.* This additional latency is often incompatible with the requirement to process tens or hundreds of frames per second.

The challenge lies in finding an optimized *trade-off* between **depth of explanation** and **speed of inference.** For real-time applications, **intrinsically interpretable** techniques and optimized variations of **CAMs** become the most viable. The use of attention maps naturally generated by **Transformers** (ViTs) can be explored as a...

This provides a "nearly free" explanation in terms of latency. Since attention calculation is inherent to Transformer inference, the attention matrix visualization can be extracted with minimal additional computational cost, offering a **spatially intuitive explanation** without significantly degrading the frame rate.

Another optimization strategy for large-scale *deployment* is **Temporal Sampling Explainability.** Instead of generating an XAI map for each frame of the video, the system can be configured to generate explanations only in **keyframes** (e.g., every 30 or 50 frames) or only when the probability of falsity exceeds a certain **uncertainty threshold.** This sampling reduces the overall computational cost and focuses explainability efforts on the most critical moments of the manipulation. This allows for *core* detection.
(Binary classification) continues to run at high speed, while the ***auditability* layer** (XAI) is selectively activated.

Future research in XAI for real-time *deepfakes* should focus on developing **lightweight ,** approximate ***post-hoc* models .** This includes training smaller **explanatory neural networks** that learn to predict the heat map of a larger black-box model (knowledge distillation for XAI) or using optimized **adversarial interrogation** techniques to generate explanations in constant time. Successful integration of XAI into high-speed environments not only validates algorithmic decision-making for forensic purposes but also improves the **user and content moderator experience,** allowing them to quickly understand the source of manipulation and make informed decisions about removing or flagging the material.

8. CONCLUSION AND FUTURE IMPLICATIONS

This analysis confirmed that **Explainable Artificial Intelligence (XAI)** is the indispensable link connecting the high accuracy of Deep Learning models for detecting *deepfakes* with the stringent requirements of **trust, transparency, and causal justification.**
In the forensic and legal domains, simple predictive performance, inherent in black-box architectures like CNNs and Transformers, is not sufficient for admissibility in a court of law, where evidence must be auditable and refutable. XAI, through techniques such as **Grad-CAM and SHAP,** transforms opaque binary classification into a **detailed expert report,** presenting visual and quantitative evidence about the specific regions and *features* that motivated the falsity decision.

The main methodological *insight* lies in recognizing that XAI is not a mere accessory, but a **validation and diagnostic tool** that addresses the biggest flaw in *deepfake detection:* the **poor generalization** resulting from *overfitting* to spurious artifacts in the *dataset.* By auditing the model with XAI at scale, developers can ensure that the algorithm is focusing on **authenticity invariants** (physical coherence, lighting, geometry) instead of compression artifacts or low-level metadata. Integrating **intrinsic interpretability** through the analysis of Vision Transformers' (ViTs) self-attention maps represents the most promising path to achieving transparency with

Minimal computational overhead, balancing the need for high accuracy with the demand for traceability.

The future implications of this study are vast and outline a threefold research agenda. First, the development of **XAI-optimized hybrid models** is a priority, focusing on creating *post-hoc* methods that are **rapidly computable** and can be integrated into real-time detection *pipelines* without latency degradation. **Explainability knowledge distillation** techniques – where a smaller, faster model learns to mimic the explanations of a large black-box model – will be crucial for this purpose.

Secondly, research should focus on creating **metrics for forensic interpretability.** Currently, interpretability is assessed subjectively or by generic metrics; for the forensic domain, it is vital to create **XAI confidence scales** that quantify the degree of validity of the explanation, allowing the legal system to have an objective standard for accepting or rejecting algorithmic evidence. This standardization is fundamental for the future **regulation of AI** in security and justice contexts.

Finally, XAI should be expanded to the **multimodal domain,** offering coherent explanations that integrate evidence of visual, auditory, and physiological manipulation. By justifying why audio does not correspond to lip movement, or why the rPPG signal is inconsistent, the XAI explanation becomes more robust and difficult to challenge. **Transparency and interpretability** are no longer luxuries; they are the foundations upon which the next generation of defense against disinformation must be built, ensuring that **Artificial Intelligence is an ally of justice and not a source of opacity and distrust.**

REFERENCES

Books and Articles

1. BACH, Sebastian et al. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. **PLoS ONE,** v. 10, no. 7, p. e0130140, 2015.

2. GOODFELLOW, Ian et al. Generative Adversarial Networks. In: **Advances in Neural Information Processing Systems** (NIPS), 2014.
3. KINGMA, Diederik P.; BA, Jimmy. Adam: A Method for Stochastic Optimization. In: **International Conference on Learning Representations** (ICLR), 2015.
4. RIBEIRO, Marco Tulio; SINGH, Sameer; GUEST RINARD, Carlos. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: **ACM SIGKDD International Conference on Knowledge Discovery and Data Mining** (KDD), 2016.
5. SUNDARARAJAN, Mukund; TALY, Ankur; YAN, Vijay. Axiomatic Attribution for Deep Networks. In: **International Conference on Machine Learning** (ICML), 2017.

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

7. SELVARAIU, Ramprasaath R. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In: **IEEE International Conference on Computer Vision** (ICCV), 2017.

8. LUNDBERG, Scott M.; LEE, Su-In. A Unified Approach to Interpreting Model Predictions. In: **Advances in Neural Information Processing Systems** (NIPS), 2017.

9. FONG, Ruth; VEDANTAM, Shreya; LIM, Joo Hwee. Interpretable Deep Learning for Image Analysis. **arXiv:1901.07746,** 2019.

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