Year III, v. 1, 2023. | submission: 02/10/2023 | accepted: 03/12/2023 | publication: 04/14/2023

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

Matheus de Oliveira Pereira Paula
Bachelor's degree in information systems from the Federal Institute of Education, Science and Technology of Fluminense. Master's student in Data Science and Artificial Intelligence (MSc) at Université Côte d'Azur.

ABSTRACT

The growing sophistication of *deepfakes* has heightened the urgency of robust Deep Learning-based 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 legal**. This article explores the application of **Explainable Artificial Intelligence (XAI)** techniques in the context of deepfake detection, investigating how to achieve model transparency and interpretability. *This article discusses post-hoc and intrinsic methodologies, such as CAMs (Class Activation Maps), SHAP, and LIME, and analyzes* their ability to generate visual and logical evidence about the classification process, specifically identifying the regions of the image or video (artifacts) that are decisive for the falsity decision. The primary objective is to demonstrate that integrating XAI is essential to build the trust needed in detection systems, transforming algorithmic decision-making into verifiable expert evidence and supporting the validity and admissibility of these technologies in courts and investigations.

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

1. INTRODUCTION: THE IMPERATIVE OF EXPLAINABILITY IN FORENSIC DIGITAL

The rapid advancement of *deepfake* creation methods has created new challenges for society, requiring **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 authentic from counterfeit media. However, this high performance is achieved at the expense 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 integrating

This is an article published in open access (Open Access) under the Creative Commons Attribution license, which allows use, distribution, and reproduction in any medium, without restrictions as long as the original work is properly cited.

deepfakes detection into high-responsibility contexts, such as **forensic investigation and legal litigation**, where a mere declaration of "false" by an algorithm is insufficient to be considered proof.

In a legal environment, forensic 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 video characteristics led the model to classify the content as forged. If CNN is focusing on a compression artifact rather than an inconsistency in facial geometry, the decision may be wrong, but the model's opacity prevents this critical distinction. Therefore, **Explainable Artificial Intelligence (XAI)** emerges not only as an enhancement 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 the main explainability methodologies can be adapted to the forensic domain, transforming the binary detector 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 absent in traditional opaque models.

The research will differentiate between model-agnostic and model-specific explainability techniques, assessing their applicability to identifying deepfake artifacts that vary in subtlety and location (from micropixel artifacts to global illumination inconsistencies). The purpose is to strengthen detection by ensuring that the model is not "cheating" by learning spurious correlations (such as watermarks from training datasets or metadata) rather than true manipulation signatures. Ultimately, adopting XAI goes beyond mere algorithmic interpretation; it is the foundation for building a responsible, **ethically grounded digital defense system**.

2. THE NATURE OF THE "BLACK BOX" AND THE CHALLENGE OF FORENSIC RELIABILITY

The rise of **Deep Learning** in Computer Vision tasks marked a paradigm shift: exponential gains in accuracy came at the expense 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 nonlinear transformations make the mapping from input (video) to output (classification) inherently opaque. This opacity, known as the **"black box"** problem, generates fundamental distrust in any environment where system failure can have serious and irreversible consequences, with the legal system being the most prominent example.

This is an article published in open access (Open Access) under the Creative Commons Attribution license, which allows use, distribution, and reproduction in any medium, without restrictions as long as the original work is properly cited.

In the forensic context, the **reliability** of digital evidence depends on its **traceability and justification**. When an expert presents a report claiming that a video is a *deepfake*, this conclusion cannot rest solely on a 98% probability-of-falsehood result from an unknown algorithm. The judge, lawyers, and jury need to understand ***why***; it is necessary for the model to point out, with spatial and temporal precision, the characteristics (or their absence) that corroborate the manipulation thesis, turning probabilistic prediction into **causal evidence**. The absence of this explicit justification prevents adversarial proceedings and technical criticism of evidence, violating basic principles of due process of law.

The challenge of the black box in *detecting deepfakes* is compounded 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 "false" 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 under minimal variation, making the model useless for generalization and totally inadmissible as reliable proof, since the cause of classification is not *deepfake*, but a dataset production artifact.

Therefore, implementing XAI in *deepfake detection* is an **ethical-legal necessity** that goes beyond 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 turning the black box into a **glass box**, where the protrusions of the decision are visible, XAI not only validates the high accuracy of Deep Learning models but also elevates your decision from a technical suggestion to a **high-value expert evidence**, essential for tackling misinformation in the era of synthetic media.

3. POST-HOC *LAMB TECHNIQUES*: ACTIVATION AND RESOURCE ALLOCATION MAPS

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 the need to modify its internal architecture. These techniques focus on measuring the **contribution and influence** of each input *feature* (pixels or pixel regions) in the model's final prediction. In computer vision, 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 generate a heatmap over the input image, highlighting the regions most relevant for classification. In deepfake detection, a successful Grad-CAM should highlight areas with manipulation artifacts, such as the face-fusion **line, inconsistencies in the eyes/mouth, or discrepancies in lighting and skin texture**. If the model is properly focused, the heat map should focus on *the forgery points*. This visualization is extremely valuable for

This is an article published in open access (Open Access) under the Creative Commons Attribution license, which allows use, distribution, and reproduction in any medium, without restrictions as long as the original work is properly cited.

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 real handling failure or an irrelevant background noise.

In parallel, **Resource Allocation 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* for the probability of falsehood. It provides Shapley values that quantify each region's contribution to the final classification, ensuring the explanation is **globally consistent and locally accurate**. LIME, in turn, builds a locally interpretable linear model (near the data point being explained) to approximate the complex behavior of the black box model.

The applicability of these *post-hoc* methods to *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 been *overfitted* to a spurious *dataset artifact* . The limitation of these methods is that they provide only a local explanation (valid only for the video under analysis, not for the overall behavior of the model) and can be **computationally expensive** (like SHAP), which impacts real-time analysis of long videos. However, for *offline* forensic reports, the richness and depth of 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 to 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*, not by retrospective analysis. The search for intrinsically interpretable models is a direct response to the inherent fragility 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 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 faces and false 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 **justification by highly intuitive analogy** for humans.

This is an article published in open access (Open Access) under the Creative Commons Attribution license, which allows use, distribution, and reproduction in any medium, without restrictions as long as the original work is properly cited.

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 Transformer, which indicates how the model weights different *patches* of the image to build its representation, can be directly interpreted as a **relevance map**. Although it is not a complete causal explanation in the sense of SHAP, 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 weight concentration.

The great advantage of intrinsically interpretable models is the **inherent confidence** they provide, since 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, designing these architectures is a challenge, as the restriction of interpretability can, paradoxically, **limit the model's predictive capacity**, forcing a shift from performance to transparency. The field of research continuously seeks to mitigate this shift by 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 detecting *deepfakes* goes beyond merely justifying a single classification; it becomes a critical tool for validating **generalization** capacity and **combating algorithmic bias**. As discussed in comparative studies, the main weakness of *deepfake detection models* is their tendency **to fail in unseen data** (*cross-dataset generalization*), which is often a symptom that the model has learned spurious correlations in training data. XAI provides the necessary microscope to identify and correct this unwanted behavior.

By applying techniques like Grad-CAM or SHAP on a large-scale validation dataset, experts can audit the model's features. If the model, in all videos classified as "false," consistently focuses on metadata or a uniform *compression artifact*, this reveals dataset **bias**. The detector is not learning to manipulate the concept of facial falsehood, but rather to detect the fingerprint of the training dataset creation process. This discovery, facilitated by XAI, allows researchers to **retrain the model** using more robust **data augmentation** techniques (such as introducing different levels of compression) or force the model to ignore areas with spurious artifacts through *attention masking*.

Validation of **generalization capability** also greatly benefits XAI. When a model trained on a *dataset* (e.g., FaceForensics++) shows a drop in accuracy in 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 cannot produce a coherent *heatmap* for the fake videos in the new *dataset*, the evidence is clear: the *deepfake* of

This is an article published in open access (Open Access) under the Creative Commons Attribution license, which allows use, distribution, and reproduction in any medium, without restrictions as long as the original work is properly cited.

the new dataset does not have 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, becomes a **feedback *and correction* mechanism**. It ensures that detection models are trained to focus on **falsehood invariants** – the logical and physical flaws that are difficult for any generator to eliminate, such as inconsistency in global illumination or error in mapping the 3D geometry of the face – instead of easily eliminatable 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 MODEL TRANSPARENCY

The transition from opaque detection to **Explainable Artificial Intelligence (SAI)** has profound and transformative implications for the forensic and legal spheres. The main barrier to adopting AI systems in courts is not a lack of accuracy, but the inability to meet the standard **of scientific evidence**, which requires the methodology to be auditable, understandable, and rebuttable by an expert from the opposing party. XAI is the tool that enables this admissibility, turning the algorithm's decision into **robust expert evidence**.

In a legal process, the **model's explanation** becomes the **expert report** itself. A Grad-CAM heatmap, for example, can be attached to the report to visually demonstrate that the model based its classification of "false" on the interpolation failure around the jaw (a common artifact area) and not on a random spot in the background of the video. Similarly, Shapley values can be used to quantify the **magnitude of the artifact of falsity** in terms of the probability of manipulation, providing a numerical measure of decision certainty that is transparent and logically grounded. The absence of this explicit makes 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 focuses on spurious or irrelevant artifacts, allowing technical **criticism** of the evidence presented. This ability to audit the algorithm's bias is essential to ensure the justice system's impartiality. Without explainability, the only way to refute it is to claim that the model "simply got it wrong," which is not a valid scientific criticism.

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

This is an article published in open access (Open Access) under the Creative Commons Attribution license, which allows use, distribution, and reproduction in any medium, without restrictions as long as the original work is properly cited.

7. XAI INTEGRATION 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 informative post-hoc explainability methods, such as SHAP and LIME, require high computational cost because they rely on evaluating multiple input perturbations or running many iterations to estimate each feature's contribution. This additional latency is often incompatible with the requirement to process tens or hundreds of frames per second.

The challenge lies in seeking an optimized *trade-off* between **depth of explanation** and **speed of inference**. For real-time applications, **intrinsically interpretable techniques** and optimized CAM variants are the most viable. The use of attention maps naturally generated by **Transformers** (ViTs) can be explored as an "almost free" explanation in terms of latency. Since attention calculation is inherent to Transformer inference, the attention matrix visualization can be extracted at minimal additional computational cost, offering a **spatially intuitive explanation** without significantly degrading the frame rate.

Another optimization strategy for large-scale deployment is **Time Sampling Explainability**. Instead of generating an XAI map for each video frame, the system can be configured to generate explanations only in **keyframes** (for example, every 30 or 50 frames) or only when the probability of falsehood exceeds an **uncertainty threshold**. This sampling reduces the total computational cost and focuses explainability effort on the most critical moments of manipulation. This allows *core* detection (binary classification) to continue running at high speed, while the **auditability *layer*** (XAI) is selectively enabled.

Future research in XAI for real-time *deepfakes* should focus on developing **lightweight *and* approximate** post-hoc **models**. This includes training smaller **explanatory neural networks** that learn to predict the heatmap of a larger black-box model (distillation knowledge for XAI) or using optimized **adversarial interrogation** techniques to generate explanations in constant time. The successful integration of XAI in high-speed environments not only validates algorithmic decision-making for forensic purposes but also improves **the user and content moderator experience**, enabling them to quickly understand the source of manipulation and make informed decisions about removing or flagging 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 *deepfake detection* with the rigorous requirements of **trust, transparency, and causal justification** in the forensic and legal domains. Simple predictive performance, inherent to black box architectures like CNNs and Transformers, is not sufficient for admissibility in a court, where evidence must be

This is an article published in open access (Open Access) under the Creative Commons Attribution license, which allows use, distribution, and reproduction in any medium, without restrictions as long as the original work is properly cited.

auditable and refutable. The XAI, through techniques such as **Grad-CAM and SHAP**, transforms the 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 failure in *deepfake* detection: **the misgeneralization resulting** from *overfitting to* dataset spurious artifacts. By auditing the model with XAI at scale, developers can ensure that the algorithm is focusing on **authenticity invariants** (physical coherence, lighting, geometry) rather than compression artifacts or low-level metadata. Integrating **intrinsic interpretability** through the analysis of Vision Transformers' self-attention maps (ViTs) 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 triple research agenda. First, the development of **hybrid XAI-optimized models** is a priority, focusing on creating post-hoc methods that are computationally efficient 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.

Second, research should focus on creating **forensic interpretability metrics**. Currently, interpretability is evaluated subjectively or by generic metrics; for the forensic domain, it is vital to create **XAI trust 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 future **AI regulation** in security and justice contexts.

Finally, XAI should be expanded into the **multimodal** domain, offering coherent explanations that integrate evidence of visual, auditory, and physiological manipulation. By justifying why the audio does not match lip movement, or why the rPPG signal is inconsistent, the XAI explanation becomes more robust and difficult to contest. Transparency **and interpretability are** no longer luxuries; they are the foundations on 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 Decisionsby Layer-Wise Relevance Propagation. PLoS ONE, v. 10, n. 7, p. e0130140, 2015.

This is an article published in open access (Open Access) under the Creative Commons Attribution license, which allows use, distribution, and reproduction in any medium, without restrictions as long as the original work is properly cited.

2. GOODFELLOW, Ian et al. Generative Adversarial Networks. In: Advances in NeuralInformation 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 TrustYou?": 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 forDeep Networks. In: International Conference on Machine Learning (ICML), 2017.
6. VASWANI, Ashish et al. Attention Is All You Need. In: Advances in Neural InformationProcessing Systems (NIPS), 2017.
7. SELVARAIU, Ramprasaath R. et al. Grad-CAM: Visual Explanations from DeepNetworks 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 forImage 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.

This is an article published in open access (Open Access) under the Creative Commons Attribution license, which allows use, distribution, and reproduction in any medium, without restrictions as long as the original work is properly cited.