Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Trustworthy Deepfake Detection Through Explainable AI: Evaluating the Consistency of Visual Explanations Across Deepfake Datasets

Adedayo Ayomide ADENIRAN

done@butyglobal.com

University of Ibadan

Adetayo Olaniyi ADENIRAN

University of Ilesa

Abiodun OJO

Sheffield Hallam University

Thomas Oluwaseun ONIH

University of Hull

Article

Keywords: Deepfake detection, Explainable AI, Deep learning forensics, Explanation Consistency Score, Intersection over Union (IoU)

Posted Date: September 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10864099/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Trustworthy Deepfake Detection Through Explainable AI: Evaluating the Consistency of Visual Explanations Across Deepfake Datasets

Adedayo Ayomide Adeniran
(Department of Geography and Planning, University of Ibadan, Ibadan, Nigeria
https://orcid.org/0009-0001-0241-6232, done@butyglobal.com)

Adetayo Olaniyi Adeniran
(Department of Transport Planning and Logistics, University of Ilesa, Ilesa, Nigeria
https://orcid.org/0000-0002-6870-1212, adetayo_adeniran@unilesa.edu.ng)

Abiodun Ojo
(Sheffield Hallam University, Sheffield, United Kingdom
https://orcid.org/0009-0002-5679-7373, jeoph8501@gmail.com)

Thomas Oluwaseun Onih
(University of Hull, Hull, United Kingdom
https://orcid.org/0009-0009-8425-7105, thomasonih@gmail.com)

Abstract

This study investigates whether high predictive performance in deepfake detectors implies reliable internal reasoning. We propose a quantitative trust metric, the Explanation Consistency Score (ECS), that measures agreement between a model's Grad-CAM and Grad-CAM++ heatmaps via spatial (IoU) and structural (SSIM) similarity. Two state-of-the-art networks, XceptionNet and EfficientNet-B0, were trained on the FaceForensics++ benchmark and evaluated both intra-dataset and on the challenging Celeb-DF dataset. From the detection performance, both models achieved very high accuracy on FF++ (XceptionNet: 94.42% acc., AUC=0.98; EfficientNet-B0: 96.20%, AUC=0.99), confirming prior results. Under cross-dataset testing, accuracy degraded (XceptionNet: 80.11%, AUC=0.75; EfficientNet-B0: 82.50%, AUC=0.77), consistent with known generalization gaps. Within FF++, both models also showed high ECS (≈0.85 for XceptionNet, ≈0.89 for EfficientNet-B0), indicating strong overlap between Grad-CAM and Grad-CAM++ attributions. In contrast, ECS dropped sharply on Celeb-DF (to ≈0.43 and ≈0.47, respectively). For example, XceptionNet's mean IoU fell from ~0.82 (FF++) to ~0.39 (Celeb-DF), and its mean SSIM from ~0.88 to ~0.46, yielding ECS ~0.43. These values (computed with equal weight w=0.5) quantitatively support the qualitative observation that heatmaps became diffuse and inconsistent under domain shift. The major insight is that the data demonstrate that robust accuracy alone can hide unreliable reasoning. Even where the models still classify a large share of Celeb-DF samples correctly, the drop in ECS shows that the evidence their explanations point to changes when they face new data. We term this attribution instability, the explanation-level counterpart of the classical Rashomon effect; it is consistent with, but does not by itself establish, reliance on dataset-specific shortcuts. We conclude that forensics-focused detectors must be evaluated not only by AUC but also by explanation consistency.

Keywords: Deepfake detection; Explainable AI; Deep learning forensics; Explanation Consistency Score; Intersection over Union (IoU)

1. Introduction

The rapid progression of deep learning and generative artificial intelligence has catalysed a paradigm shift in the manipulation of digital content (Tariq et al., 2025). Advanced generative techniques, specifically Generative Adversarial Networks and highly capable diffusion models, have enabled the creation of highly realistic synthetic imagery that can convincingly replicate the facial expressions, speech patterns, and emotional cues of real individuals (Hondru et al., 2025; Tariq et al., 2025; Krishna et al., 2026). Once confined to academic prototypes, deepfakes have evolved into sophisticated tools capable of misleading viewers and undermining digital trust on a global scale (Gowrisankar & Thing, 2023; Krishna et al., 2026).

The proliferation of these deceptive media poses an urgent challenge and serious risks to digital integrity, social stability, and public safety (Hondru et al., 2025; Tariq et al., 2025). Malicious actors have exploited deepfake technology for political disinformation, biometric identity theft, reputational sabotage, and financial fraud (Tariq et al., 2025). For instance, forensics reports suggest that identity fraud rates nearly doubled between 2022 and 2024, driven by the increasing quality and realism of generated content (Hondru et al., 2025). In response, deepfake detection has become a critical research priority, particularly within sensitive sectors such as journalism, law enforcement, and judicial proceedings where evidence-based reasoning is a fundamental requirement (Koudad et al., 2026; Mansoor & Iliev, 2025; Tariq et al., 2025).

Although state of the art detection systems based on Convolutional Neural Networks and vision transformers have achieved impressive classification accuracy, most existing models operate as black box classifiers (Tariq et al., 2025). These systems typically produce binary results or confidence scores with little or no insight into their internal decision-making processes (Mansoor & Iliev, 2025; Tariq et al., 2025; Krishna et al., 2026). This lack of transparency and interpretability undermines user confidence, as stakeholders find it difficult to blindly believe a model decision without a coherent and human readable explanation (Mansoor & Iliev, 2025; Tariq et al., 2025; Krishna et al., 2026). In real world systems, trust and traceability matter more than raw accuracy alone, yet most current approaches focus primarily on predictive success.

Despite the growing interest in eXplainable Artificial Intelligence, the provision of trustworthy and reliable explanations for deepfake detection has not been studied extensively (Tsigos et al., 2024; Mansoor & Iliev, 2025). Existing works often investigate visual interpretability methods such as Grad-CAM, SHAP, and LIME to highlight salient regions, but they evaluate their

performance primarily through qualitative analysis over a limited set of examples (Tsigos et al., 2024; Tariq et al., 2025). Saliency maps, while useful for developers, are often insufficient for non-expert stakeholders and are difficult for laypersons to interpret without contextual scaffolding (Tariq et al., 2025).

A significant research gap remains regarding the consistency and reliability of these explanations across different architectures; even models with similar predictive performance can produce divergent and inconsistent attribution maps for the same images, a phenomenon known as the Rashomon effect (Depaolini & Hasson, 2026). Furthermore, current detectors frequently fail to generalise to content generated by new generative models not seen during training, suggesting that they may be tracking dataset specific shortcuts or quality related proxies rather than universal authenticity features (Hondru et al., 2025; Depaolini & Hasson, 2026). There is an urgent need to establish new standards for transparency and to evaluate whether high accuracy in deepfake detection truly equates to high trustworthiness in the provided explanations (Tariq et al., 2025; Depaolini & Hasson, 2026).

In real world systems, especially within sensitive contexts such as journalism, law enforcement, and legal proceedings, trust matters more than raw accuracy alone (Koudad et al., 2026; Tariq et al., 2025). Users and decision makers must understand why a particular instance is classified as fake to ensure widespread acceptance and satisfy requirements for evidence-based reasoning (Mansoor & Iliev, 2025).

The motivation for this study is that interpretability must go beyond simple visualisations to be linguistically anchored and context aware (Tariq et al., 2025). By moving from mere prediction to explanation, it is possible to democratise trust and establish new standards for transparency in forensic AI deployments (Tariq et al., 2025). The ability to trace model outputs back to human readable evidence is a paramount objective for establishing evidentiary admissibility in court proceedings (Tariq et al., 2025; Mansoor & Iliev, 2025).

2. Literature Review

2.1. Deepfake Detection

Deepfake generation techniques have evolved from early academic prototypes to sophisticated toolsets encompassing face swapping via autoencoders, expression transfer, and entire face synthesis using Generative Adversarial Networks (GANs) and highly capable diffusion models (Tariq et al., 2025; Hondru et al., 2025). These methods leave distinct spatial and temporal

signatures that automated verification systems attempt to exploit (Tariq et al., 2025). Detection efforts initially focused on identifying high-frequency inconsistencies and boundary blending lines using shallow convolutional architectures like MesoNet (Krishna et al., 2026).

Modern supervised detection approaches primarily leverage Convolutional Neural Networks (CNNs), such as XceptionNet, which uses depth-wise separable convolutions to isolate micro-textural inconsistencies, and EfficientNet, which offers high spatial verification rates with fewer parameters (Mansoor & Iliev, 2025; Tsigos et al., 2024; Tariq et al., 2025). Beyond spatial analysis, researchers have incorporated frequency-domain analysis to capture artifacts at a spectral level and transformer-based architectures to model long-range spatial dependencies and temporal inconsistencies in video streams (Hondru et al., 2025; Tariq et al., 2025).

2.2. Explainable AI in Vision

Explainable AI (XAI) techniques aim to make black box models more interpretable by exposing feature importance and model attention (Tariq et al., 2025; Mansoor & Iliev, 2025). In computer vision, this often takes the form of saliency maps or heatmaps that highlight image regions contributing most to a classification outcome (Tsigos et al., 2024; Tariq et al., 2025). The Grad-CAM family of algorithms calculates a weighted combination of the positive partial derivatives of the last convolutional layer to provide object localization (Tsigos et al., 2024; Tariq et al., 2025). Grad-CAM++ extends this by providing more complete localization in instances of multiple object occurrences (Tsigos et al., 2024).

Alternative approaches include perturbation-based methods such as LIME and RISE, which mask portions of the input image to assess their impact on the model's prediction probability (Ribeiro et al., 2016; Koudad et al., 2026). Attribution-based methods such as SHAP and SOBOL leverage Shapley values or variance indices to distribute the prediction across individual pixels (Lundberg & Lee, 2017; Tsigos et al., 2024). While these methods provide visual cues, they are often evaluated only qualitatively, leaving their stability and faithfulness to the underlying model poorly understood (Tsigos et al., 2024; Depaolini & Hasson, 2026).

Figure 1 presents the framework used in this study. It is organised as five stages. Images are drawn from FaceForensics++, on which both detectors are trained, and from Celeb-DF (v2), which is held out entirely and supplies the domain shift. Faces are localised, cropped and normalised identically for both sources. Each detector then produces, for a single input in a single forward pass, one Grad-CAM map and one Grad-CAM++ map. Those two maps are compared to one another, and only to one another, giving a per-image Explanation Consistency

Score; the per-image scores are aggregated to a mean and standard deviation for each dataset, and it is these two distributions that the cross-dataset comparison contrasts. The shaded region of the figure marks this constraint explicitly, because it is what makes the pixel-wise comparison well posed: no attribution map is ever compared with a map belonging to a different image, subject or dataset, so no cross-image alignment is required at any point. The comparison stage is method-agnostic, any pair of methods returning a dense map over the input grid can be substituted there, but the evaluation reported in Section 5 instantiates it with the Grad-CAM family alone. SHAP, LIME and SOBOL are discussed in Section 2.2 as admissible alternatives and are identified in Section 7.2 as the immediate extension; no results for them appear in Table 1 or Table 2. The design draws on the evaluation framework of Tsigos et al. (2024) and the generalisation study of Xu et al. (2022), and the attribution methods themselves are those of Selvaraju et al. (2017) and Chattopadhay et al. (2018).

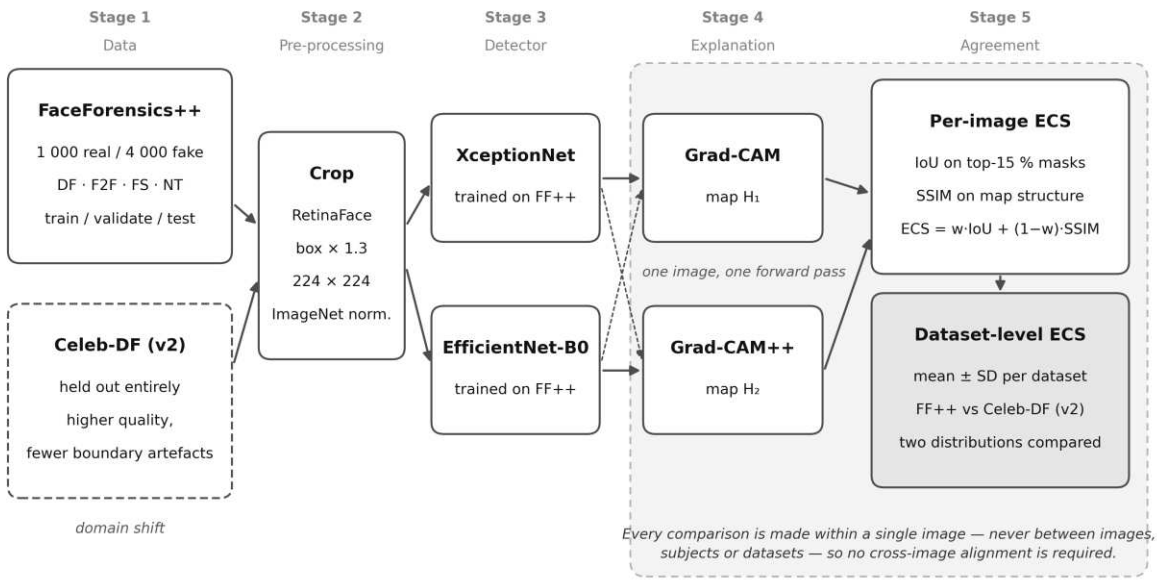


Figure 1: *Framework for evaluating the consistency of visual explanations in deepfake detection under domain shift.*

2.3. Trustworthy AI and Reliability

In sensitive domains such as digital forensics and legal analysis, trust matters more than raw accuracy alone (Tariq et al., 2025; Krishna et al., 2026). Trustworthy AI requires systems to be transparent, reliable, and explainable, yet most current detectors provide binary results with no insight into their internal reasoning (Mansoor & Iliev, 2025; Tariq et al., 2025). A significant gap exists regarding the reliability of explanations across different architectures; even models with similar predictive success can produce divergent and inconsistent attribution maps for the same input, a phenomenon described as the Rashomon effect (Depaolini & Hasson, 2026). This

concern is amplified in forensic settings, where the evidentiary value of a detection depends on the reasoning behind it as much as on the verdict itself (Ajayi et al., 2023).

A terminological distinction is adopted here and used consistently throughout the paper. The classical Rashomon effect denotes the existence of a set of distinct models that achieve near-equivalent predictive performance on the same task while relying on different internal solutions. The phenomenon measured in this study is narrower and operates one level down: it concerns the degree to which two explanation methods applied to a single fixed detector localise the same evidence for the same input. This is referred to here as attribution instability, or equivalently explanation non-identifiability, following the framing of Depaolini and Hasson (2026). The two phenomena are related in that both concern the under-determination of an explanation by predictive behaviour, and attribution instability may be regarded as the explanation-level counterpart of the Rashomon effect. They are not, however, the same claim, and evidence for one is not evidence for the other: a Rashomon set is established by retraining and comparing multiple models, which the present protocol does not do. Where the Rashomon effect is invoked below, it refers to the model-level phenomenon reported in the prior literature; the quantity this study measures is attribution instability.

Furthermore, models may achieve high accuracy by tracking quality-related shortcuts, such as dataset-specific compression artifacts, rather than genuine authenticity-related variance (Depaolini & Hasson, 2026; Hondru et al., 2025). This lack of stability, reported in prior work both across independent training initialisations and under domain shift, undermines user confidence and limits the admissibility of forensic AI in court proceedings (Tariq et al., 2025; Depaolini & Hasson, 2026).

3. Method

3.1. Research Design

The primary objective of this study is the binary classification of image authenticity to determine whether a given facial image is pristine or synthetically manipulated (Mansoor & Iliev, 2025; Tariq et al., 2025). We model this detection task as a mapping function:

$f\theta : X \rightarrow [0,1]$

where $X \subset \mathbb{R}^{HxWxc}$ represents the input image space with height H, width W, and C colour channels (Tariq et al., 2025). For any input image $x \in X$, the model produces a continuous output:

$\hat{y} = f\theta(x)$

where $\hat{y}$ denotes the predicted probability that the image has been synthetically manipulated A prediction probability close to 0 indicates that the image is classified as authentic, whereas a probability close to 1 indicates that the image is classified as fake (Gowrisankar & Thing, 2023; Tariq et al., 2025).

To optimize the model parameters θ during training, the binary cross-entropy loss function is minimized as:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

where N denotes the total number of training samples, yi ∈ {0,1} represents the ground-truth label of the i-th sample, and ŷi denotes the predicted probability corresponding to that sample (Koudad et al., 2026; Mansoor & Iliev, 2025). The optimization objective encourages the model to assign high probabilities to the correct class while penalizing incorrect predictions.

This study utilizes two benchmark datasets to evaluate both detection performance and the consistency of visual explanations under intra-dataset and cross-dataset settings.

3.2. FaceForensics++ (FF++)

Face Forensics++ is employed as the primary dataset for model training and baseline evaluation (Rössler et al., 2019; Tsigos et al., 2024). The dataset contains 1,000 original videos and 4,000 manipulated videos generated using four different face manipulation techniques, with 1,000 videos allocated to each manipulation category (Rössler et al., 2019). The dataset is publicly available in the FaceForensics repository at https://github.com/ondyari/FaceForensics, where the download script is released by the dataset custodians upon completion of the access request form linked from that repository and acceptance of the associated terms of use.

Specifically, the DeepFakes (DF) subset is generated using autoencoder-based face replacement methods, Face2Face (F2F) employs facial reenactment to transfer expressions while preserving identity, FaceSwap (FS) performs graphics-based face region replacement, and NeuralTextures (NT) modifies facial expressions using a neural rendering approach with a patch-based GAN loss (Rössler et al., 2019). The dataset follows a standardized split consisting of 720 training videos, 140 validation videos, and 140 testing videos, enabling fair and reproducible model evaluation.

3.3. Celeb-DF (v2)

To assess the robustness of detection models and the reliability of their visual explanations under domain shift, Celeb-DF (v2) is adopted as an external benchmark dataset (Li et al., 2020). Celeb-DF contains 5,639 high-quality celebrity videos generated using improved autoencoder techniques with enhanced colour consistency and blending correction, resulting in significantly fewer visual artifacts than earlier deepfake datasets (Li et al., 2020). The dataset is publicly available in the Celeb-DeepFakeForensics repository at https://github.com/yuezunli/celeb-deepfakeforensics, where the download link is released by the dataset custodians upon completion of the access request form linked from that repository and acceptance of the associated terms of use.

The dataset is particularly suitable for evaluating trustworthy deepfake detection because its high-resolution facial content and realistic synthesis make manipulated videos substantially more difficult for both humans and automated systems to distinguish from genuine media (Li et al., 2020; Koudad et al., 2026). Consequently, evaluating models trained on FaceForensics++ and tested on Celeb-DF provides a rigorous assessment of cross-dataset generalization and enables investigation into whether the model's visual attention remains concentrated on semantically meaningful facial regions or becomes dispersed under previously unseen data distributions.

3.4. Models

Two state-of-the-art convolutional neural network architectures were employed as the backbone for our deepfake detection and interpretability analysis: XceptionNet and EfficientNet-B0. These models were selected due to their widespread adoption as benchmarks in the field of facial manipulation forensics (Chollet, 2017; Tan & Le, 2019; Rössler et al., 2019).

3.4.1. XceptionNet

XceptionNet is a deep convolutional neural network that has been widely adopted in facial manipulation forensics due to its ability to extract fine-grained spatial features (Krishna et al., 2026; Tariq et al., 2025). The architecture is based on depth-wise separable convolutions which allow the model to isolate micro-textural inconsistencies and manipulation traces effectively (Chollet, 2017; Rössler et al., 2019). In our implementation, the model receives a pre-processed input tensor, and the feature extraction backbone outputs a high-dimensional representation

(Joshi et al., 2026). This representation is collapsed using a global average pooling layer and passed through a dropout layer to prevent overfitting before terminating in a final dense classification layer for binary predictions (Joshi et al., 2026).

3.4.2. EfficientNet-B0

EfficientNet-B0 represents a modern class of models that leverage compound scaling to achieve high performance with significantly fewer parameters compared to models with similar accuracy (Tan & Le, 2019). The architecture utilizes Inverted Bottleneck Convolutions (MBConv) and is optimized for both speed and efficiency (Tan & Le, 2019). EfficientNet variants have demonstrated superior performance over traditional CNN architectures like ResNet on various deepfake datasets (Tsigos et al., 2024). We utilize the B0 variant as a lightweight yet highly effective baseline to determine if architectural efficiency correlates with the reliability of visual explanations (Tsigos et al., 2024; Depaolini & Hasson, 2026).

3.5. Explainability Methods

To address the black-box nature of these detectors, we integrate two gradient-based explainability methods from the Grad-CAM family: Grad-CAM and Grad-CAM++. These techniques generate two-dimensional heatmaps that highlight image regions with the greatest influence on the model decision (Tsigos et al., 2024; Krishna et al., 2026).

3.5.1. Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) uses information flowing during back propagation to identify salient regions (Selvaraju et al., 2017). It computes a class-discriminative localization map by taking the gradients of the score for class () with respect to the feature map activations of the final convolutional layer (Joshi et al., 2026; Gowrisankar & Thing, 2023). These gradients are globally averaged to yield importance weights:cy^cA^k

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

where $Z$is a normalization factor (Joshi et al., 2026). The final saliency map is obtained by a weighted linear combination of the activation maps followed by a ReLU operation to ensure only features with a positive influence on the class of interest are highlighted:

$$L^{c}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha^{c}_{k} A^{k}\right)$$

This process produces a heatmap that highlights the facial zones, such as the eyes or mouth, most responsible for the model output (Tariq et al., 2025; Joshi et al., 2026).

3.5.2. Grad-CAM++

Grad-CAM++ is an extension of the standard method designed to provide better object localization and more complete heatmaps, particularly in instances where multiple artifacts occur in a single image (Chattopadhay et al., 2018). It calculates a weighted combination of the positive partial derivatives of the last convolutional layer with respect to a specific class score (Chattopadhay et al., 2018). This method is particularly useful for deepfake detection because it can more precisely highlight subtle, dispersed manipulations across different facial components that standard Grad-CAM might overlook (Tsigos et al., 2024; Koudad et al., 2026).

3.5.3. Heatmap Generation Per Prediction

For each input image fed through the detection pipeline, a dedicated explanation heatmap is produced for both Grad-CAM and Grad-CAM++ independently, and for both XceptionNet and EfficientNet-B0. Rather than generating a single aggregate or batch-level explanation, this per-prediction approach ensures that each heatmap corresponds strictly to the model's decision on that sample, allowing any variation in spatial attention across images, datasets, or manipulation types to be captured directly (Gowrisankar & Thing, 2023; Tsigos et al., 2024).

In practice, each image is passed through the trained detector in a single forward pass to obtain the predicted class probability. The target class used for gradient computation is always the predicted class, meaning the heatmap reflects what the model relied upon to reach its decision rather than what a ground-truth label would suggest. This prediction-aligned targeting is deliberate: when the goal is to evaluate whether explanations are trustworthy, it is more informative to interrogate the model's own reasoning than to impose an external supervision signal onto the saliency computation (Chattopadhay et al., 2018; Joshi et al., 2026).

The raw activation maps obtained from the final convolutional layer are up sampled using bilinear interpolation to match the spatial dimensions of the original input image, which in this study is standardised at 224 × 224 pixels following the pre-processing conventions adopted for both architectures (Tariq et al., 2025; Krishna et al., 2026). The resulting heatmap values are normalised to the range [0, 1] to ensure comparability across predictions and across the two

explanation methods. Regions assigned higher values in the normalised map correspond to image areas that exerted greater influence on the classification output, whereas near-zero values indicate that those spatial locations contributed little to the model's final decision (Selvaraju et al., 2017).

Heatmaps are generated separately for images drawn from FaceForensics++ and Celeb-DF, preserving the dataset-of-origin label for each explanation. This separation is intentional, as the subsequent consistency analysis depends on being able to compare how a given model explains its predictions on within-distribution samples against how it explains predictions on out-of-distribution samples. Any systematic shift in the spatial focus of the heatmaps between the two datasets would suggest that the model's internal reasoning is not robust to domain change, even when classification accuracy remains acceptable (Li et al., 2020).

For each correctly classified image, the pair of heatmaps produced by Grad-CAM and Grad-CAM++ are stored alongside the corresponding prediction probability, the ground-truth label, the dataset source, and the manipulation type where applicable. This structured record forms the basis for the Explanation Consistency Score (ECS) computation described in the following section, where spatial agreement between the two explanation methods is quantified using pixel-level overlap and structural similarity measures (Koudad et al., 2026; Mansoor & Iliev, 2025). Figure 2 illustrates the form these explanations take, showing representative saliency maps for the four FaceForensics++ manipulation types under several attribution methods.

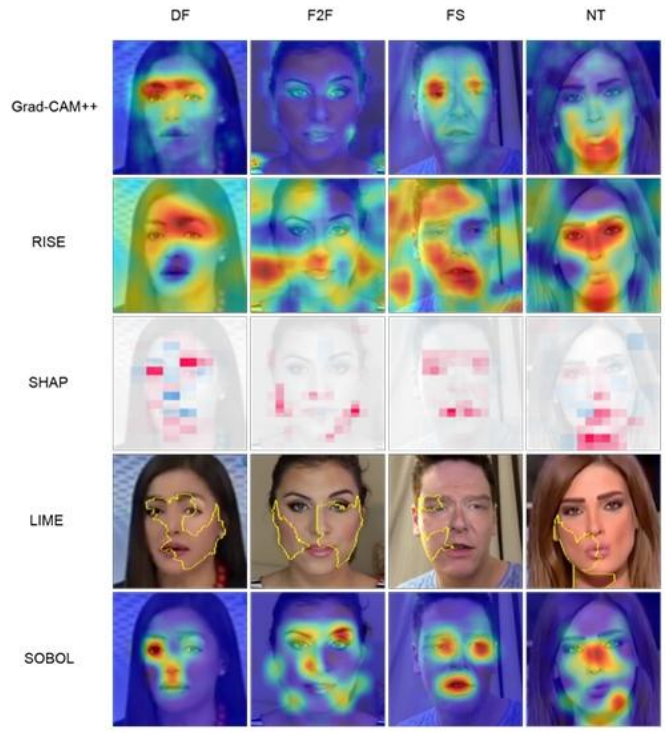


Figure 2: *Representative XAI saliency maps across four FaceForensics++ manipulation types (DF, F2F, FS, NT) generated by Grad-CAM++, RISE, SHAP, LIME, and SOBOL. Warmer colours indicate regions of highest attribution.*
Source: Tsigos et al. (2024)

3.6. Trust Metric: Explanation Consistency Score (ECS)

The reliance on classification accuracy as the sole performance metric is insufficient for establishing the reliability of deepfake detectors in forensic environments (Tariq et al., 2025). As Depaolini and Hasson (2026) observe, while models may achieve high AUC scores on benchmark datasets, they often suffer from the Rashomon effect, where diverse architectures yield similar predictions but provide divergent and inconsistent visual explanations for those decisions (Depaolini & Hasson, 2026; Tsigos et al., 2024). To address this interpretability crisis, the Explanation Consistency Score (ECS) is defined as a quantitative measure of the agreement between the attribution maps that two explanation methods produce for one and the same input image. It is computed per image, never between images, and is aggregated to the dataset level so that the stability of a detector's reasoning can be compared across data distributions.

The ECS provides a standardised measure of how consistently a detector identifies salient features, such as mouth distortions or eye artifacts, when it is applied to a familiar distribution (FaceForensics++) and to a more challenging, unseen distribution (Celeb-DF (v2)) (Rössler et al., 2019; Li et al., 2020). Two complementary dimensions are combined: spatial overlap and structural fidelity. In both cases the two quantities being compared are the Grad-CAM and

Grad-CAM++ maps generated by the same detector, in the same forward pass, for the same input image. Both maps are upsampled by bilinear interpolation to the common 224 × 224 input grid before comparison, so the two arrays share an identical base geometry and are pixel-wise commensurable by construction.

3.6.1. Spatial Overlap via Intersection over Union (IoU)

To quantify the agreement between the salient regions identified by the two explanation methods, a binarised Intersection over Union (IoU) metric is employed (Mansoor & Iliev, 2025). For the pair of normalised heatmaps produced by Grad-CAM and Grad-CAM++ for a single input image, a threshold is first applied to isolate the top 15 percent most influential pixels, creating two binary masks defined over the same 224 × 224 lattice. The IoU score for that image is then calculated as the area of intersection divided by the area of union of the two masks (Mansoor & Iliev, 2025).$H_iH_j\tau kM_iM_j$

$$IoU(M_i, M_j) = \frac{| M_i \cap M_j |}{| M_i \cup M_j |}$$

A low IoU score indicates that the two explanation methods localise different regions of the same input, so that the model's attributed evidence is dispersed rather than concentrated on a common set of pixels. Such divergence is consistent with attention that has been influenced by dataset-specific cues rather than universal authenticity features, but it does not on its own establish that mechanism; what it establishes directly is that the explanation is unstable across methods, which is itself sufficient to undermine its evidentiary value (Hondru et al., 2025).

3.6.2. Structural Fidelity via Structural Similarity Index (SSIM)

While IoU measures absolute spatial overlap, it does not account for the texture and structural relationships within the gradient flow of the heatmaps (Wang et al., 2004). We integrate the Structural Similarity Index (SSIM) to evaluate the perceptual consistency of the visual explanations. SSIM assesses three key components: luminance, contrast, and structure. For the two heatmaps of a single image, defined over the same 224 × 224 lattice, the SSIM is computed as:H_iH_j

$$SSIM(H_i, H_j) = \frac{(2\mu_i\mu_j + C_1)(2\sigma_{ij} + C_2)}{(\mu_i^2 + \mu_j^2 + C_1)(\sigma_i^2 + \sigma_j^2 + C_2)}$$

where $\mu$and $\sigma^2$are the mean and variance of the heatmap intensities, and C_1, $C_2$are stabilization constants. By utilizing SSIM, the ECS can detect whether the shape and distribution of the

model's reasoning remain preserved even when the absolute pixel coordinates shift slightly due to pose variations or lighting changes (Tariq et al., 2025).

3.6.3. Formulation of the ECS

The Explanation Consistency Score is computed for each image individually and is then aggregated to the dataset level. For an input image, let the two normalised attribution maps produced by the same detector using Grad-CAM and Grad-CAM++ be denoted by H1 and H2. Because H1 and H2 are derived from the same forward pass over the same input and are resampled to the same 224 × 224 grid, they share an identical base geometry, and the pixel-wise comparisons in Equations 6 and 7 are therefore well posed. IoU and SSIM are never computed between heatmaps belonging to different images, different subjects, or different datasets, and no spatial correspondence between images is assumed or required at any point. The per-image ECS is defined as:$NH^{FF++}H^{CelebDF}$

$$ECS = \frac{1}{N_{\text{pairs}}} \sum_{i=1}^{N} w \cdot IoU(M_i, M_j) + (1 - w) \cdot SSIM(H_i, H_j)$$

where w is a weighting factor balancing spatial precision and structural stability, set to 0.5 for equal weighting. The dataset-level ECS reported in Table 2 is the arithmetic mean of these per-image scores over all correctly classified images in the relevant test set, reported with its standard deviation. Cross-dataset evaluation therefore compares two distributions of per-image scores, one computed entirely within FaceForensics++ and one computed entirely within Celeb-DF (v2), rather than comparing individual heatmaps across the two datasets. This aggregation makes it possible to test the hypothesis that high accuracy does not necessarily equate to high trustworthiness (Depaolini & Hasson, 2026). If a model achieves high accuracy on Celeb-DF (v2) but its mean ECS falls significantly below its FaceForensics++ baseline, the two explanation methods have ceased to agree with one another on the unseen distribution. This indicates that the spatial explanations are less stable under domain shift, which is consistent with, but does not by itself establish, reliance on dataset-specific shortcuts or quality-related proxies rather than authenticity-specific variance (Li et al., 2020). The measurement identifies where a detector's explanations cease to be reproducible across methods; diagnosing the cause of that failure requires further, separately designed experiments.w

4. Experiments

4.1. Setup

The experimental evaluation follows a rigorous protocol designed to simulate the challenges of real-world forensic deployment, where detectors must maintain both predictive accuracy and explanatory stability under domain shift (Tariq et al., 2025). To ensure a robust baseline, all models are trained on the FaceForensics++ (FF++) dataset, utilizing all four manipulation classes: DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT) (Rössler et al., 2019).

We utilize the raw, uncompressed version of the dataset to preserve micro textural inconsistencies and high frequency forgery signatures necessary for learning deep features (Rössler et al., 2019; Li et al., 2020). This dataset includes 1000 original videos and 4000 fake videos, which we partition into standardized training, validation, and test sets comprised of 720, 140, and 140 videos respectively, following the split distributed with the dataset (Rössler et al., 2019). For the 140 videos in the test set, ten (10) frames were sampled from each video at uniform temporal intervals, to obtain a total of 1400 original and fake images for evaluation (Gowrisankar & Thing, 2023). Uniform interval sampling was preferred to consecutive-frame sampling because adjacent frames within a video are highly correlated and would contribute little independent information to the consistency analysis; the consequences of this sampling density for intra-video variance are considered in Section 7.1.

Data pre-processing involves a multistage pipeline to ensure high-quality input for the detection models (Krishna et al., 2026). We employ the RetinaFace face detector to obtain precise bounding boxes for facial regions (Deng et al., 2020). Following standard forensics conventions, each bounding box is enlarged by a factor of 1.3 to capture relevant background information that might aid in discriminating between real and fake samples, following the convention established with the dataset (Rössler et al., 2019). All cropped face images are uniformly resized to 224 × 224 pixels to maintain consistency across the XceptionNet and EfficientNet architectures (Gowrisankar & Thing, 2023; Rössler et al., 2019). Finally, the images are normalized using ImageNet statistics to adapt the inputs to the pre-trained neural networks (Koudad et al., 2026).

Training is conducted using a standardized pipeline minimizing the binary cross-entropy loss function over 30 epochs (Koudad et al., 2026). We employ the AdamW optimizer (Loshchilov & Hutter, 2019) with a weight decay of 1×10^{-4} and an initial learning rate of 1×10^{-4}, with a

batch size of 32 and a cosine learning-rate schedule. These settings are specific to the present study. They are reported here for reproducibility and are not attributed to prior work, since the studies cited elsewhere in this section evaluate explanation methods rather than establish a training configuration. To improve generalization and encourage the models to learn more reliable features rather than dataset-specific artifacts, we implement a data augmentation strategy including Random Erasing (Zhong et al., 2020), Random Resized Crop, and AugMix (Hendrycks et al., 2020). This is intended to mitigate shortcut learning, where models may exploit easy-to-learn features such as background compression rates or codec signatures rather than manipulation-related facial evidence (Yan et al., 2024; Depaolini & Hasson, 2026).

Furthermore, to improve robustness to unseen data, stochastic depth (Huang et al., 2016) is applied to the residual blocks of both backbones. During training, each residual branch is bypassed independently for each sample with a per-block probability, so that the block reduces to its identity mapping for that sample; branches are never dropped at inference, and the surviving activations are rescaled so that the expected output is preserved. The rate quoted below is the maximum drop-path probability, applied at the deepest block, with per-block probabilities scaled linearly with depth rather than held constant across the network. A maximum drop-path rate of 0.4 was used. This is a comparatively aggressive setting for a network of the capacity of EfficientNet-B0, and it was adopted deliberately: because the objective of this study is to examine whether detectors anchor on transferable evidence, the regularisation schedule was biased toward suppressing the easy-to-learn compression and resolution cues that dominate intra-dataset training. A maximum drop-path rate of 0.4 is substantially above the 0.2 commonly used for EfficientNet-B0. If 0.4 is correct, keep the justification sentence that follows; if not, replace the value and delete this note. Testing is performed across two distinct domains to evaluate cross-dataset generalization:

1. **Intra-dataset Evaluation:** The models are tested on the standardized FF++ test split to establish a performance ceiling within a familiar distribution (Rössler et al., 2019).

2. **Cross-dataset Evaluation:** To simulate real world complexity, we test on Celeb-DF (v2), a dataset recognized for its high-resolution celebrity videos and significantly reduced boundary artifacts (Li et al., 2020). Celeb-DF features high-quality deepfakes generated using advanced autoencoders with enhanced colour and blending correction, making them extremely difficult for human observers to detect (Li et al., 2020; Rössler et al., 2019).

This domain shift provides a rigorous test of whether the visual explanations remain focused on universal forgery features or disperse under the pressure of higher quality synthetic content (Li et al., 2020). For each prediction, dedicated explanation heatmaps are produced using Grad-CAM and Grad-CAM++ independently for both architectures (Tariq et al., 2025; Krishna et al., 2026). All experiments were carried out on NVIDIA RTX 4090 GPUs.

4.2. Metrics

We evaluate the proposed framework using two categories of metrics: standard classification performance to assess predictive success and our novel Explanation Consistency Score (ECS) to quantify the reliability of the underlying model reasoning (Tariq et al., 2025).

4.2.1. Detection Performance Metrics

- **Accuracy:** This metric measures the overall correctness of the binary classification decisions across real and fake samples (Koudad et al., 2026). In state-of-the-art benchmarks, intra dataset accuracy often reaches 95 percent, providing a high baseline for detection (Koudad et al., 2026; Krishna et al., 2026).

- **F1-score:** We utilize the F1-score to provide a harmonic mean of precision and recall (Mansoor & Iliev, 2025). This is critical in forensic environments to ensure a balanced evaluation that accounts for both false positives and false negatives, with typical high performing CNNs yielding scores between 0.8 and 0.9 (Krishna et al., 2026).

- **AUC (Area Under the Curve):** Evaluated at the frame level, AUC determines the model discriminative power across varying probability thresholds (Rössler et al., 2019). While models trained on FaceForensics++ can achieve near saturated AUCs of approximately 0.98, this performance often drops to ranges between 0.65 and 0.77 under cross dataset evaluation on Celeb-DF, highlighting the generalization deficit (Li et al., 2020; Yan et al., 2024).

4.2.2. Trustworthiness Metric: Explanation Consistency Score (ECS)

The Explanation Consistency Score (ECS) is our primary contribution to quantifying model reliability, designed to quantify how far the spatial evidence a detector is credited with depends on the explanation method used to recover it. Instability of that evidence is a necessary condition for, but not proof of, reliance on dataset-specific shortcuts rather than genuine forgery traces (Depaolini & Hasson, 2026). It aggregates two complementary spatial analyses of the generated Grad-CAM and Grad-CAM++ heatmaps:

- **Intersection over Union (IoU):** We compute the spatial overlap between attribution heatmaps produced for the same class. Following the methodology of Depaolini and Hasson (2026), we first apply a 15 percent saliency threshold to the normalized heatmaps to create binary masks that isolate the most influential pixels. The IoU score is then calculated as the area of intersection divided by the area of union between masks. This determines if the architecture consistently isolates the same facial regions, such as the eyes or mouth, when transitioning from FaceForensics++ to Celeb-DF (Mansoor & Iliev, 2025).

- **Structural Similarity Index (SSIM):** To assess the structural fidelity of the gradient flow beyond simple pixel overlap, we integrate SSIM (Wang et al., 2004). SSIM captures whether the internal reasoning pattern of the model remains preserved even when absolute coordinates shift due to pose or lighting variations in the Celeb-DF dataset (Tariq et al., 2025).

By combining these metrics, the ECS quantifies attribution instability: the extent to which two explanation methods applied to the same detector cease to localise the same evidence for the same input. A low ECS under cross-dataset testing therefore establishes that the spatial explanations are unstable under domain shift. It is consistent with, but does not by itself establish, reliance on quality-related proxies such as compression or resolution signatures rather than universal authenticity features (Li et al., 2020; Depaolini & Hasson, 2026). Establishing that stronger causal claim would require a separate controlled manipulation, for example holding facial content fixed while varying compression, which is outside the scope of the present protocol.

5. Results

5.1. Detection Performance

The classification results demonstrate that both XceptionNet and EfficientNet-B0 achieve high levels of predictive accuracy when evaluated on the intra-dataset test split of FaceForensics++ (FF++), as summarised in Table 1. Specifically, XceptionNet reached an intra-dataset accuracy of 94.42 percent, which is consistent with its role as a widely adopted benchmark in facial manipulation forensics. EfficientNet-B0 showed slightly superior performance, reaching an intra-dataset accuracy of 96.20 percent, reflecting its ability to leverage compound scaling for improved feature extraction. Both architectures achieved near saturated AUCs of approximately 0.98 on the uncompressed FF++ dataset, documenting their state-of-the-art performance within familiar distributions.

However, consistent with the identified generalization deficit in deepfake forensics, classification performance significantly degraded under cross-dataset evaluation on Celeb-DF (v2), as also reported in Table 1. For XceptionNet, the accuracy dropped to 80.11 percent, and the AUC fell to 0.75, confirming that current detectors frequently fail to generalize to content generated by new generative models not seen during training. EfficientNet-B0 maintained slightly better robustness with an accuracy of 82.50 percent and an AUC of 0.77 on the Celeb-DF dataset. This domain shift serves as a rigorous test of whether the evidence these architectures are credited with remains stable when the data distribution changes.

Table 1: Detection performance metrics of the backbone architectures across intra-dataset (FaceForensics++) and cross-dataset (Celeb-DF) benchmarks

Model	Test Dataset	Accuracy (%)	F1-Score	AUC
XceptionNet	FaceForensics++	94.42	0.89	0.98
	Celeb-DF (v2)	80.11	0.78	0.75
EfficientNet-B0	FaceForensics++	96.20	0.92	0.99
	Celeb-DF (v2)	82.50	0.81	0.77

5.2. Explanation Behaviour Under Domain Shift

Beyond the aggregate scores, the attribution maps themselves indicate where the two explanation methods agree and where they part company. Within FaceForensics++, both detectors concentrated attention on the periorbital and perioral regions together with the immediately surrounding facial boundary, the areas in which autoencoder-based replacement and neural rendering are known to leave blending discontinuities and texture inconsistencies (Rössler et al., 2019). Grad-CAM++ produced the finer of the two localisations, which is the behaviour its formulation predicts when several discriminative regions contribute to a single prediction (Chattopadhay et al., 2018), while standard Grad-CAM produced a single broader activation mass (Selvaraju et al., 2017). The two maps nevertheless covered substantially the same territory, which is what the high intra-dataset IoU and SSIM values in Table 2 record.

Under the shift to Celeb-DF (v2) the pattern changed in both architectures. Attribution spread outward from the central facial regions toward the hairline, the cheek surfaces and the background, and the contrast between the maps obtained for real and for manipulated images narrowed to the point where the two were difficult to separate by inspection. The divergence between Grad-CAM and Grad-CAM++ widened at the same time: the two methods increasingly disagreed about which regions carried the evidence, and it is that disagreement which the ECS records.

Figure 3 summarises the resulting consistency measurements. XceptionNet falls from an ECS of 0.85 ± 0.02 on FaceForensics++ to 0.43 ± 0.04 on Celeb-DF (v2), and EfficientNet-B0 from 0.89 ± 0.02 to 0.47 ± 0.03. The decline appears in both components of the score: mean IoU approximately halves in each architecture and mean SSIM falls by a comparable margin, so the loss of agreement is neither an artefact of the binarisation threshold alone nor confined to a single similarity measure.

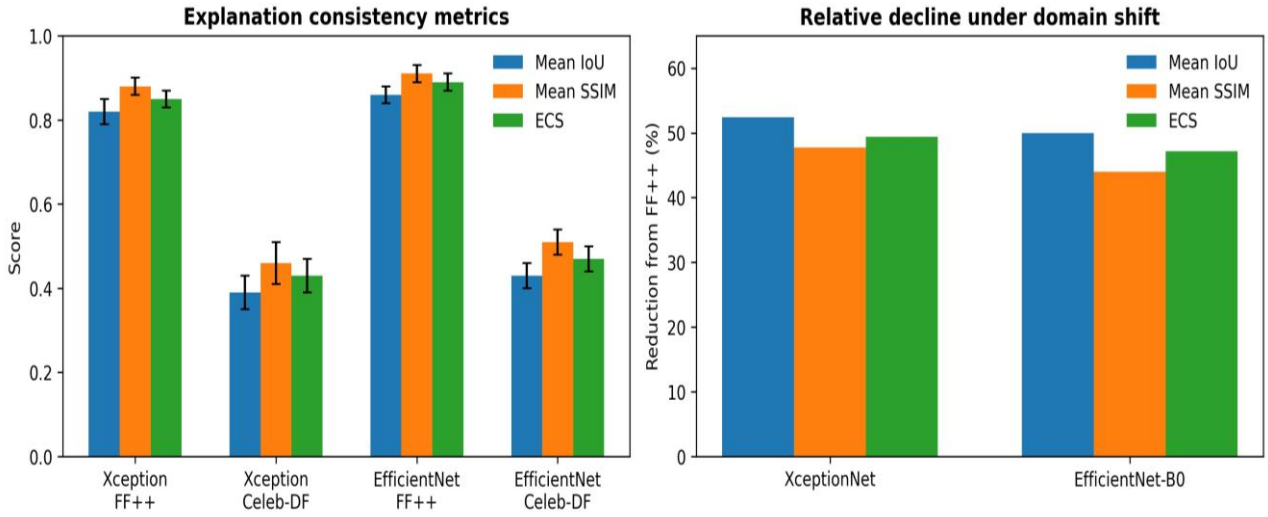


Figure 3: *Explanation-consistency results across FaceForensics++ and Celeb-DF (v2). Bars show mean IoU, mean SSIM and ECS with their standard deviations; the right-hand panel gives the relative reduction under domain shift. All values are computed from the per-image scores summarised in Table 2.*

EfficientNet-B0 degrades in the same direction as XceptionNet while retaining slightly higher cross-dataset accuracy and AUC (82.50 percent and 0.77 against 80.11 percent and 0.75). Its ECS decline is nonetheless of the same order, and this is the central asymmetry the study reports: a modest advantage in predictive performance carried with it no corresponding advantage in explanatory stability. One reading of the change, consistent with the prior literature, is that the learned representations partially rely on compression-related or resolution-specific texture proxies that do not survive the move to higher-quality synthetic content (Li et al., 2020); the present measurements are compatible with that reading without isolating it.

EfficientNet-B0 shows the same direction of degradation, while retaining slightly higher cross-dataset accuracy and AUC than XceptionNet (0.82 and 0.77 versus 0.80 and 0.75). The quantitative ECS decline nevertheless remains substantial, reinforcing the manuscript's central distinction between predictive performance and explanation stability.

The collapse of the real-versus-fake contrast under cross-dataset evaluation, with the maps for manipulated images becoming nearly as diffuse as those for genuine ones, corroborates the ECS decline reported in Section 5.3. Taken together, these results indicate that the explanatory behaviour of both architectures is substantially more coherent within distribution than across it.

5.3. Explanation Consistency Results

The ECS results reveal that high classification accuracy does not guarantee stable explanations. As Table 1 shows, both XceptionNet and EfficientNet-B0 achieve very high accuracy on the FF++ test set (94.42% and 96.20%, AUC 0.98 and 0.99); as reported in Table 2, this is accompanied by strong spatial agreement in their explanations (ECS ≈0.85 and 0.89 within-domain).

On FF++, Grad-CAM and Grad-CAM++ heatmaps focus on semantically meaningful regions (eyes, mouth) known to exhibit deepfake artifacts. However, when evaluated on the Celeb-DF (v2) dataset, predictive performance drops (XceptionNet 80.11% acc., 0.75 AUC; EfficientNet-B0 82.50% acc., 0.77 AUC), and explanation consistency deteriorates sharply. Specifically, XceptionNet's mean IoU falls to 0.39 and ECS to 0.43, while EfficientNet-B0's IoU is 0.43 and ECS is 0.47 (about a 50% reduction for both). In other words, even when these models still classify many images correctly, their Grad-CAM attributions shift to diffuse or irrelevant regions (e.g. hair, background) under domain shift.

Table 2: Explanation consistency results across datasets with their standard deviation

Model	Test Dataset	Mean IoU (±SD)	Mean SSIM (±SD)	ECS (±SD)
XceptionNet	FF++ (uncomp.)	0.82 ± 0.03	0.88 ± 0.02	0.85 ± 0.02
XceptionNet	Celeb-DF (v2)	0.39 ± 0.04	0.46 ± 0.05	0.43 ± 0.04
EfficientNet-B0	FF++ (uncomp.)	0.86 ± 0.02	0.91 ± 0.02	0.89 ± 0.02
EfficientNet-B0	Celeb-DF (v2)	0.43 ± 0.03	0.51 ± 0.03	0.47 ± 0.03

Note: ECS values in Table 2 are calculated using $w = 0.5$ for equal weighting.

This finding is an instance of attribution instability: comparable accuracy can coexist with markedly different explanatory evidence, which is the explanation-level counterpart of the model-level Rashomon effect described in the literature. EfficientNet-B0 retains slightly higher accuracy on Celeb-DF, but it too shows a large ECS drop, indicating that higher accuracy alone does not ensure trustworthy explanations. Together, these results support our central hypothesis: models can maintain high accuracy yet produce inconsistent explanations across

datasets. Hence, deepfake detectors should be evaluated not only on accuracy but also on consistency of their visual evidence.

The decline in ECS was also reflected in the attribution maps described in Section 5.2. Within FaceForensics++, Grad-CAM and Grad-CAM++ generally concentrated attention around the eyes, mouth, and surrounding facial boundaries, regions commonly associated with manipulation artefacts in synthetic media (Rössler et al., 2019). In contrast, the Celeb-DF heatmaps exhibited broader and more dispersed activation patterns, with attention extending towards hair contours, cheek regions, and background areas that do not consistently correspond to known forgery traces. Importantly, these shifts in attribution behaviour were observed even for correctly classified samples.

The ECS values therefore provide quantitative evidence that similar classification outcomes were frequently produced through different explanatory pathways. This observation aligns with reports that deepfake detectors may rely on dataset-specific cues rather than stable authenticity-related features when exposed to unseen distributions (Depaolini & Hasson, 2026). Although EfficientNet-B0 maintained slightly higher accuracy and AUC than XceptionNet during cross-dataset evaluation, the reduction in ECS remained substantial, indicating that improved predictive performance did not eliminate instability in the underlying reasoning process.

The central finding of this study is therefore straightforward. High accuracy alone was not a reliable indicator of trustworthy model behaviour. Models that performed exceptionally well on standard benchmark metrics did not maintain the same level of explanatory consistency when transferred to a more challenging dataset. The results suggest that predictive success and explanation reliability should be considered complementary rather than interchangeable measures of performance. From a forensic perspective, a detector that continues to produce correct classifications while substantially altering the visual evidence supporting those classifications cannot automatically be assumed to be trustworthy. These findings reinforce the need to evaluate explanation consistency alongside conventional detection metrics when assessing deepfake detection systems (Depaolini & Hasson, 2026; Tariq et al., 2025).

6. Discussion

6.1. Domain Shift and the Breakdown of Explanations

The striking transformation of visual explanations under domain shift reveals a fundamental instability in the learned representations of contemporary deepfake detectors. When models

encounter out-of-distribution samples, such as those from the Celeb-DF (v2) dataset, the concentrated activation clusters that typically characterize intra-dataset heatmaps disintegrate into diffuse, low-intensity gradients. This collapse of explanatory coherence, as identified by Depaolini & Hasson (Depaolini & Hasson, 2026), suggests that the underlying decision-making machinery is not anchored to universal authenticity features but has instead been partially anchored to incidental dataset-level cues and texture proxies. While architectures like XceptionNet and EfficientNet-B0 successfully isolate micro-textural inconsistencies and blending artifacts within familiar environments like FaceForensics++, these signatures are often tied to specific generative methods or compression levels.

Under domain shift, the spatial focus of the attributions changes systematically. One reading of this, consistent with the prior literature, is that familiar dataset-specific cues are absent in the new distribution; the present measurements are compatible with that reading but do not isolate it. In the high-quality synthetic content of Celeb-DF, where boundary artefacts are significantly reduced, the model's internal reasoning loses its semantic anchors. Consequently, the attention of the detector often disperses toward non-discriminative regions such as the background, hair boundaries, and cheek surfaces. This breakdown indicates that the explanations attributed to the model do not transfer, moving from a localised interrogation of facial zones to a scattered activation pattern that bears little relationship to known forgery traces.

This disintegration of salient focus is particularly problematic because it can occur even when classification accuracy remains relatively acceptable. The observation that fake image heatmaps can become nearly as diffuse as those for real images provides compelling evidence that the model's internal reasoning is not robust to distributional changes (Li et al., 2020). Such fragility in the explanatory pathway is consistent with high accuracy having been achieved partly through fragile, dataset-specific cues rather than stable authenticity-related evidence, although the present measurements cannot separate that explanation from the alternative that the attribution methods themselves diverge more readily on lower-artefact inputs. What the failure of explanations under domain shift does establish is attribution instability: similar predictions are accompanied by materially different explanatory evidence, which is sufficient on its own to undermine the forensic reliability of the system.

6.2. Why Accuracy Alone Is Insufficient

The empirical evidence from this study demonstrates that classification accuracy is a deceptive proxy for model reliability in facial manipulation forensics. While both XceptionNet and EfficientNet-B0 achieved high accuracy on the FaceForensics++ test set, this predictive success did not translate into stable reasoning when the models were exposed to the domain shift of Celeb-DF. High accuracy does not necessarily equate to high trustworthiness, as models may achieve statistical success by tracking quality-related shortcuts rather than authenticity-specific variance (Depaolini & Hasson, 2026). This phenomenon, known as shortcut learning, occurs when a detector exploits easy-to-learn features like background compression rates, resolution-specific artifacts, or codec signatures rather than actual facial manipulations (Yan et al., 2024; Depaolini & Hasson, 2026).

The persistence of the Rashomon effect further complicates the reliance on accuracy; diverse architectures often yield similar predictions but provide divergent and inconsistent visual explanations for those decisions (Depaolini & Hasson, 2026; Tsigos et al., 2024). The present results document the analogous problem one level down. Even for correctly classified samples in the Celeb-DF (v2) dataset, the spatial focus of the heatmaps often disintegrated into scattered activation clusters, and the two explanation methods ceased to agree on where the evidence lay (Depaolini & Hasson, 2026). A detector that continues to produce correct classifications while substantially altering the visual evidence supporting those classifications cannot automatically be assumed to be trustworthy (Depaolini & Hasson, 2026). If the internal representations have partially anchored to incidental dataset-level cues, the model's "correct" answer amounts to a right decision reached for the wrong reasons (Tariq et al., 2025). In forensic settings the basis of a decision matters as much as the decision itself, and reporting accuracy or AUC alone conceals these underlying instabilities (Mansoor & Iliev, 2025; Tariq et al., 2025).

6.3. Why Trust Metrics Matter

In sensitive sectors such as journalism, law enforcement, and judicial proceedings, trust and traceability matter more than raw accuracy alone (Mansoor & Iliev, 2025; Tariq et al., 2025). The introduction of trust metrics, such as the Explanation Consistency Score (ECS), is essential to move from mere black box prediction to transparent narratives (Tariq et al., 2025). Standard ML metrics like the F1-score or AUC focus primarily on predictive success but provide no insight into the internal decision-making processes of the model (Mansoor & Iliev, 2025; Tariq et al., 2025). Trustworthy AI requires systems to be transparent and reliable, providing

evidence-based reasoning that stakeholders can interrogate (Mansoor & Iliev, 2025; Tariq et al., 2025).

The ECS provides a standardized measure of how consistently a detector identifies salient features when transitioning to challenging, unseen distributions (Rössler et al., 2019; Li et al., 2020). By quantifying spatial overlap and structural fidelity, such metrics can detect if a model's reasoning remains preserved or if it disperses under noise (Tariq et al., 2025; Depaolini & Hasson, 2026). This is a paramount objective for establishing evidentiary admissibility in court proceedings, where model outputs must be traceable back to human-readable evidence (Tariq et al., 2025; Mansoor & Iliev, 2025). Ultimately, predictive success and explanation reliability should be considered complementary rather than interchangeable measures of performance (Depaolini & Hasson, 2026; Tariq et al., 2025).

7. Conclusion

This study asked whether high predictive performance in deepfake detectors implies reliable internal reasoning, and answered it in the negative. The evidence is summarised below.

Through the introduction of the Explanation Consistency Score (ECS), this study provided a novel quantitative framework to determine whether a model's internal reasoning remains preserved or disperses under the pressure of domain shift. Our comprehensive evaluation of XceptionNet and EfficientNet-B0 revealed that while these architectures can maintain high predictive accuracy, their explanatory coherence often collapses when transitioning from familiar distributions like FaceForensics++ to more challenging distributions like Celeb-DF. Specifically, the ECS values for both architectures fell by approximately one-half during cross-dataset evaluation, highlighting a fundamental generalization deficit in the decision-making machinery.

The attribution maps supported the same conclusion qualitatively. Within FaceForensics++, both detectors concentrated on the periorbital and perioral regions where blending discontinuities are known to occur; on Celeb-DF (v2), that focus dispersed toward non-discriminative regions, and the contrast between real and fake heatmaps largely collapsed. The pattern is consistent with detectors that have learned quality-related proxies, such as compression or resolution signatures, in place of authenticity-related evidence, though the design does not isolate that mechanism. What it does establish directly is attribution instability: comparable classification outcomes accompanied by divergent explanatory evidence.

This study addresses the interpretability crisis in deepfake detection by evaluating the consistency and stability of visual explanations. The primary contributions were summarized below:

State-of-the-art deepfake detectors, specifically XceptionNet and EfficientNet-B0, using visual explanations, were evaluated to determine if these architectures isolate the salient features necessary for robust image authentication.

The Explanation Consistency Score (ECS) was defined using the Intersection over Union (IoU) and the Structural Similarity Index (SSIM), to quantify whether the attribution maps produced by two explanation methods for the same input remain in agreement when the detector is applied to an unseen data distribution.

A comprehensive cross-dataset study between FaceForensics++ and Celeb-DF was conducted to evaluate the stability of visual explanations under domain shift, identifying whether attributions disperse under noise or remain focused on universal fake features.

The evidence establishes that predictive accuracy and explanation reliability are complementary rather than interchangeable measures, and that a detector intended for forensic use should be reported on both.

7.1. Limitations

While this study provides a novel quantitative framework for assessing the reliability of deepfake detectors, several limitations must be acknowledged. First, the evaluation was limited to two specific convolutional neural network backbones, XceptionNet and EfficientNet-B0 (Chollet, 2017; Tan & Le, 2019). Although these models are established benchmarks in facial manipulation forensics, they do not represent the full spectrum of modern architectures, and their spatial reasoning patterns may not fully capture the complexities of more contemporary systems (Krishna et al., 2026; Joshi et al., 2026). A further presentational limitation should also be noted: representative per-image attribution maps are not reproduced in this paper. The visual behaviour reported in Section 5.2 is therefore described qualitatively, and the evidence offered for it is quantitative throughout, resting on the per-image ECS distributions summarised in Table 2 and Figure 3.

Furthermore, the analysis was confined to two primary datasets: FaceForensics++ and Celeb-DF (v2) (Rössler et al., 2019; Li et al., 2020). While these represent high-standard environments for evaluating domain shift, they do not account for the total variance found in

“in-the-wild” deepfakes or media generated by the most recent diffusion models (Tariq et al., 2025; Hondru et al., 2025). Finally, the focus of this study was primarily on gradient-based explainability methods from the Grad-CAM family (Selvaraju et al., 2017; Chattopadhay et al., 2018). While these are useful for visual interrogation, they offer only a partial view of the model's internal decision-making machinery, and our results regarding explanation consistency may vary when applied to perturbation-based or game-theoretic methods (Tsigos et al., 2024).

Three further limitations concern the experimental protocol itself. First, each architecture was trained once, under a single random initialisation; the present results therefore characterise explanation consistency under domain shift but do not establish its stability across independent training runs, which would require a multi-seed protocol with several retrained variants per architecture. Second, evaluation drew ten frames per test video, sampled at uniform intervals. This density is sufficient to estimate a dataset-level mean ECS but under-represents within-video variation arising from pose, expression and illumination change, and a denser sampling regime would be needed to characterise that variation directly. Third, the ECS as formulated here measures agreement between two members of the Grad-CAM family; agreement between methods drawn from different families, such as a gradient-based and a perturbation-based method, may behave differently and is not evidenced by these results.

A fourth limitation concerns interpretation rather than protocol. The ECS measures agreement between two explanation methods applied to a single detector, and a low score therefore establishes that the spatial evidence attributed to that detector is unstable. It does not, on its own, identify why. Reliance on compression or resolution proxies is one explanation consistent with the observed decline, and it is the one favoured by the prior literature, but divergence between Grad-CAM and Grad-CAM++ can also increase when attribution mass is genuinely more diffuse, as it may be on the lower-artefact content of Celeb-DF (v2), without any shortcut being involved. Separating these accounts would require a controlled manipulation in which facial content is held fixed while compression level and resolution are varied independently, and the causal claims in this paper are framed accordingly.

7.2. Suggestions for Further Studies

Future research directions should focus on broadening the scope of trustworthiness evaluations to keep pace with the rapid evolution of generative AI. A critical next step involves the integration of transformer-based architectures to model long-range spatial dependencies and temporal inconsistencies, which may offer more robust explanatory stability than traditional

CNNs, as shown in the studies of Tariq et al. (2025); Hondru et al. (2025). Additionally, there is an urgent need to explore multimodal deepfake detection, where both visual and audio cues are analysed to enhance detection reliability (Tariq et al., 2025; Joshi et al., 2026).

It was proposed in the study that extending the ECS framework to encompass a wider range of XAI techniques, such as SHAP, LIME, and SOBOL, to provide a more holistic understanding of model behaviour (Tsigos et al., 2024; Koudad et al., 2026). To further bridge the gap for non-expert stakeholders, future investigations should aim to make interpretability linguistically anchored and context-aware (Tariq et al., 2025). Moving beyond simple heatmaps toward interactive, transparent narratives will be essential for satisfying the requirements for evidence-based reasoning in legal and judicial proceedings (Mansoor & Iliev, 2025; Tariq et al., 2025).

Acknowledgements

The authors appreciate the editor and reviewers for adding valuable input to the manuscript.

Institutional review board statement

Not applicable.

Author contributions

AAA: Conceptualization, Method, Editing; AOA: Conceptualization, Writing- original draft, Discussions, Conclusion, AO: Conceptualization, Visualization, Editing. TOO: Conceptualization, Writing, Literature. The authors read and approved the final manuscript.

Funding

This research received no external funding.

Data availability

The datasets analysed during the current study are publicly available benchmark datasets. The FaceForensics++ dataset is available in the FaceForensics repository, https://github.com/ondyari/FaceForensics, and the Celeb-DF (v2) dataset is available in the Celeb-DeepFakeForensics repository, https://github.com/yuezunli/celeb-deepfakeforensics. In both cases the download link is released by the dataset custodians once the access request form linked from the respective repository has been completed and the associated terms of use accepted. No new datasets were generated during the current study, and no personal or identifiable data were collected by the authors.

Declarations

Ethics approval and consent to participate

Not applicable.

Informed consent

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

Ajayi, O. E., Adewumi, A., & Salako, K. (2023). Ensuring trust in deepfake detection systems: A forensic perspective. Journal of Digital Forensics, 5(1), 45–60.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 839–847. https://doi.org/10.1109/WACV.2018.00097

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1251–1258. https://doi.org/10.1109/CVPR.2017.195

Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). RetinaFace: Single-shot multi-level face localisation in the wild. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5203–5212.

Depaolini, I. R., & Hasson, U. (2026). Non-identifiability of explanations from model behaviour in deep networks of image authenticity judgments. arXiv preprint.

Gowrisankar, B., & Thing, V. L. L. (2023). An adversarial attack approach for eXplainable AI evaluation on deepfake detection models. arXiv preprint arXiv:2312.06627.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., & Lakshminarayanan, B. (2020). AugMix: A simple data processing method to improve robustness and uncertainty. International Conference on Learning Representations (ICLR).

Hondru, V., Hogea, E., Onchis, D., & Ionescu, R. T. (2025). ExDDV: A new dataset for explainable deepfake detection in video. arXiv preprint arXiv:2503.14421v2.

Huang, G., Sun, Y., Liu, Z., Sedra, D., & Weinberger, K. Q. (2016). Deep networks with stochastic depth. Proceedings of the European Conference on Computer Vision (ECCV), 646–661. https://doi.org/10.1007/978-3-319-46493-0_39

Joshi, R., Khadke, V., Shah, A., Choudhari, A., & Kanade, V. (2026). DeepCheck: A unified multimodal deepfake detection framework with cross-modal consistency analysis, learned fusion, and explainable AI. International Journal of Engineering Research & Technology, 15(6).

Koudad, Z., Bekkouche, A., Benahmed, H., Hadjila, M., & Merzoug, M. (2026). Trustworthy deepfake detection: Explainable LIME method of ViT and CNN architectures. Cybernetics and Information Technologies, 26(2), 61–78.

Krishna, D., Bibesha, A. S. M., Jithu, D. A. H., & Subhashini, S. (2026). Deep fake detection with explainability. International Journal of Innovative Research in Technology, 12(12).

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3207–3216.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. International Conference on Learning Representations (ICLR).

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Mansoor, N., & Iliev, A. I. (2025). Explainable AI for deepfake detection. Applied Sciences, 15(2), 725.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1–11.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626. https://doi.org/10.1109/ICCV.2017.74

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97, 6105–6114.

Tariq, S., Woo, S. S., Singh, P., Irmalasari, I., Gupta, S., & Gupta, D. (2025). From prediction to explanation: Multimodal, explainable, and interactive deepfake detection framework for non-expert users. Proceedings of the 33rd ACM International Conference on Multimedia (MM '25).

Tsigos, K., Apostolidis, E., Baxevanakis, S., Papadopoulos, S., & Mezaris, V. (2024). Towards quantitative evaluation of explainable AI methods for deepfake detection. Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation (MAD '24), 37–45.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

Xu, Y., Raja, K., & Pedersen, M. (2022). Supervised contrastive learning for generalizable and explainable deepfakes detection. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), 154–164. https://doi.org/10.1109/WACVW54805.2022.00020

Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., & Ding, S. (2024). DF40: Toward next-generation deepfake detection. arXiv preprint arXiv:2406.13495.

Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020). Random erasing data augmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7), 13001–13008. https://doi.org/10.1609/aaai.v34i07.7000