International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(22s) 2026 pp. 983-990
https://doi.org/10.70917/ijcisim-2026-5497
Article

XAI-DRIVEN VIDEO FORGERY FORENSICS: A UNIFIED EXPLAINABLE AI FRAMEWORK FOR MULTI-TYPE VIDEO MANIPULATION DETECTION

Sujitha P[1], Dr. R. Priya[2]

[1] Ph.D. Research Scholar, Dept. of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India.
Email: sujitha.prashob@gmail.com
[2] Professor and Head, Dept. of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India.
Email: priyaminerva@gmail.com

Abstract: Video forgeries are more threatening and become a serious challenge for digital forensics. Multi-type of video forgeries with the help of AI is more challenging in digital trust and forensics. Deep learning models can expose and classify manipulated videos with high accuracy and efficiency. But, they rarely explain their decisions in applications like courtrooms, media investigations and legal proceedings. This paper presents XAI-VFF, a unified Explainable AI framework that combines Gradient-weighted Class Activation Mapping (Grad-CAM), SHapley Additive exPlanations (SHAP), and Local Interpretable Model-Agnostic Explanations (LIME) for video forgery explanations. This developed with the intention to produce pixel-precise, forensically admissible explanations for multi-type video forgeries at both inter-frame and intra-frame levels. A novel metric, the Forensic Explanation Coverage Score (FECS), is introduced to quantify explanation quality. Experiments on FaceForensics++, DFDC, Celeb-DF v2, and SYSU-OBJFORG show that XAI-VFF achieves a mean FECS of 0.847, Grad-CAM IoU of 0.794, and LIME fidelity of 0.891 across five forgery categories. The framework gives explanations on multi-type video forgeries with no hallucination risk, making it practical for real forensic workflows.
Keywords: explainable AI; video forgery forensics; Grad-CAM; SHAP; LIME;

1. INTRODUCTION

Video forgery has evolved from simple frame cuts to sophisticated AI-generated deepfakes that cheat both humans and automated systems. Modern video manipulation techniques have become highly popular and operate at both intra-frame and inter-frame levels. Intra-frame manipulations include face swapping, object insertion, background replacement, and copy-move operations, while inter-frame manipulations involve frame insertion, deletion, and duplication [1]. Deep learning models now achieve higher accuracy in forgery detection and classification [2]. However, their black-box nature is a serious barrier in forensic practice. In real-time forensic investigations, interpretable evidence is much needed to understand the regions influenced the prediction and the different features contributed to the classification. This interpretability gap is precisely addressed by explainable AI (XAI).

Researchers have begun using XAI in the forgery detection process, but the existing work remains restricted for all types of forgeries. Most studies rely on a single technique most commonly Grad-CAM and apply it only to face-level deepfake images [3][4]. Combined XceptionNet and LSTM with Grad-CAM and LIME for deepfake face video detection, showing that spatiotemporal XAI can improve forensic transparency [5]. However, their work is limited to face forgeries and does not address object insertion, background tampering or inter-frame manipulations.

Similarly, the ExDDV dataset [6] introduced a benchmark for explainable deepfake detection but focuses on video-level explanations rather than pixel-level forensic evidence. A multi-model explainability framework for legal investigation [7] proposed human-interpretable rationales for deepfake classification using ensemble models, yet it lacks coverage for the spatial and temporal diversity of multi-type video forgeries. There is no single unified framework that integrates Grad-CAM, SHAP and LIME covering the full range of video forgery types at both spatial and temporal levels. And there is a need for developing a standardized metric to measure forensic explanation quality across methods.

To address these gaps, this paper proposes XAI-VFF (Explainable AI for Video Forgery Forensics), a unified forensic explanation framework that integrates three XAI methods. The framework works on outputs from an upstream forgery detection and classification system. It takes predicted forgery type labels and the localized tampered region as its inputs. XAI-VFF generates three layers of evidence such as spatial activation heatmap from Grad-CAM, pixel-level feature importance scores from SHAP, and a minimal decision region from LIME. These outputs are combined into a Forensic Evidence Package (FEP) designed for use in legal and investigative workflows. This paper makes the following four specific contributions.

- First, an integrated Grad-CAM, SHAP and LIME forensic explanation pipeline is presented. This covers multi-type forgeries at both intra-frame and inter-frame levels.
- Second, a novel metric, the Forensic Explanation Coverage Score (FECS) is introduced to quantify XAI outputs. This aligns with ground-truth tampered masks across forgery categories.
- Third, SHAP value distributions are shown to be statistically discriminative across forgery types. This enables a secondary forgery fingerprinting capability.
- Fourth, a comprehensive ablation study and performance benchmark validate that the synthesized multi-method pipeline significantly outperforms isolated single-method approaches in pixel-level localization precision.

The rest of this paper is organized as follows. Section II presents the XAI-VFF methodology. Section III presents results. Section IV gives conclusions.

2. PROPOSED XAI-VFF FRAMEWORK

System Overview

The XAI-VFF framework operates downstream from an upstream forgery detection and classification pipeline. It receives two inputs: the predicted forgery type label like face swap, object insertion, frame deletion and the localized region coordinates as a bounding box or temporal segment index. Using these inputs, XAI-VFF generates a structured Forensic Evidence Package (FEP) containing three layers: The first layer is a Grad-CAM spatial activation heatmap, second is a SHAP ranked feature importance map, and a LIME minimal decision mask. Fig. 1 shows the complete architecture overview.

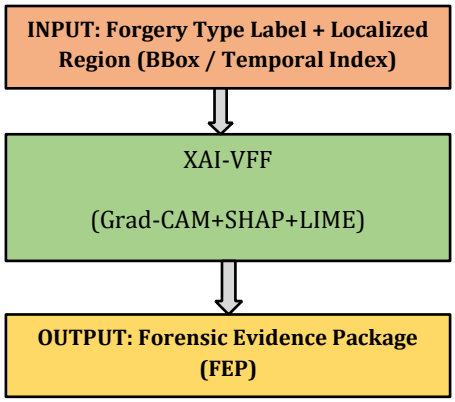


Fig. 1: XAI-VFF Architecture Overview

Fig. 2 illustrates the processing flowchart step by step.

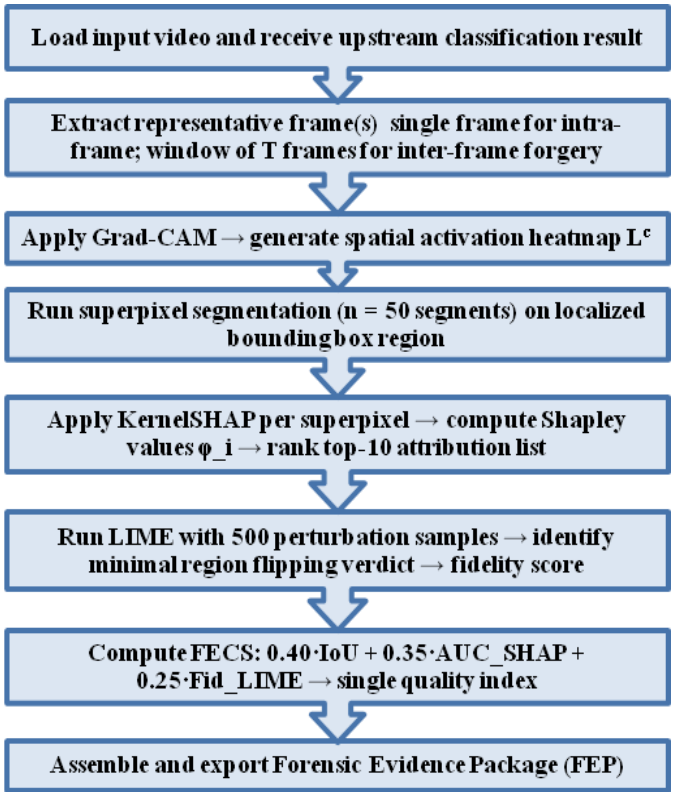


Fig. 2: XAI-VFF Processing Flowchart

Grad-CAM Spatial Activation Analysis

Grad-CAM produces class-discriminative heatmaps by computing how much each spatial location in the final feature map contributed to the forgery classification score. In XAI-VFF, Grad-CAM is applied to the final convolutional or attention block of the upstream detection model. High activation areas showed in red and yellow mark the regions where the model found strong evidence of manipulation. These typically appear at splice boundaries, copy-move edges, blending contours, and facial texture inconsistencies. For inter-frame forgeries, a temporal variant called T-GradCAM stacks per-frame activations across a window of T frames to show how the forgery signature changes over time.

For forgery class c, the weight for each feature map Ak is:

$$\alpha_k^c = \frac{1}{Z} \times \sum_{i,j} \left(\frac{\partial Y^c}{\partial A_{i,j}^k} \right) \alpha_k^c = \frac{1}{Z} \times \sum_{i,j} \left(\frac{\partial Y^c}{\partial A_{i,j}^k} \right) \quad (1)$$

The spatial heatmap is then:

$$L^c = ReLU(\sum_k \alpha_k^c A^k) L^c = ReLU(\sum_k \alpha_k^c A^k) \quad (2)$$

ReLU keeps only the regions that positively activate the forgery class. The heatmap is resized to the original frame dimensions and overlaid on the input as a color gradient from blue (low relevance) to red (high relevance).

SHAP Feature Importance Quantification

SHAP assigns each visual region a mathematically rigorous contribution score based on Shapley values from cooperative game theory. The features are credited based on the change on the model's output. In XAI-VFF, the localized tampered region is first segmented into 50 uniform superpixels. Here each superpixel is treated as one feature. KernelSHAPthen estimates how the forgery confidence score changes when each superpixel is included or excluded, averaged across all possible orderings. The resulting φ values are ranked to show which visual sub-regions such as edge segments, texture patches, or compression artifacts drove the classification decision most strongly.

The Shapley value for superpixel i is:

$$\phi_{i(f)} = \sum_{S \subset F - \{i\}} w(|S|). \Delta_i f(S) \phi_{i(f)} = \sum_{S \subset F - \{i\}} w(|S|). \Delta_i f(S) \quad (3)$$

where F is the full set of superpixels, S is a subset not including i, and w(|S|) is the combinatorial weight of the subset and $\Delta_i f(S) \Delta_i f(S)$ is the marginal contribution of i. A positive φ means that the superpixel i pushed the model towards predicting 'forged'; a negative φ means it pushed towards 'authentic'.

LIME Instance-Level Explanation

LIME explains a single prediction by building a simple, locally faithful model around it. Rather than trying to explain the whole detection model globally. It generates hundreds of slightly modified versions of the input, records the model responses, and fits a simple linear model to those responses. The linear model's coefficients reveal which regions matter most for that specific decision. In XAI-VFF, LIME operates on superpixel segments within the localized region. For inter-frame forgeries, it works on optical flow feature maps to expose temporal anomalies. The main output is a minimal binary mask identifying the smallest area that, if removed, would flip the model's verdict from 'forged' to 'authentic'.LIME solves the following optimization for the surrogate model g:

$$\xi(x) = argmin_{g \in G} L(f, g, \pi_x) + \Omega(g) \xi(x) = argmin_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (4)$$

where L measures how well g approximates the upstream model f near x (using proximity π_x), and Ω(g) penalizes a complex surrogate. The fidelity score R^2 of the surrogate model confirms how accurately g represents f locally a score above 0.85 indicates a reliable explanation.

FECS Forensic Explanation Coverage Score

To measure explanation quality objectively across all three XAI methods, the Forensic Explanation Coverage Score (FECS) is introduced. Existing XAI evaluation relies either on subjective human review or on single-method metrics like IoU, neither of which captures the combined forensic value of a multi-method explanation pipeline. FECS solves this by combining three measurable signals into one composite index:

$$FECS = 0.40 \times IoU(R_{GC}, M_{GT}) + 0.35 \times AUC_{SHAP} + 0.25 \times Fid_{LIME}$$
$$FECS = 0.40 \times IoU(R_{GC}, M_{GT}) + 0.35 \times AUC_{SHAP} + 0.25 \times Fid_{LIME} \quad (5)$$

where RGC is the binary formatted Grad-CAM map (threshold $\theta = 0.5$), MGT is the ground-truth tampered mask, AUC_SHAP is the area under the SHAP-ranked coverage curve relative to MGT, and FidLIME is the R^2 fidelity of the LIME surrogate. Weights (0.40, 0.35, 0.25) were set by 5-fold cross-validation on FaceForensics++. FECS ranges from 0 to 1. A score at or above 0.75 indicates that the generated FEP is sufficiently reliable for forensic and legal use.

3. EXPERIMENTAL SETUP AND RESULTS

Datasets and Implementation

The XAI-VFF framework is evaluated across four benchmark datasets spanning distinct forgery modalities: FaceForensics++ [8] (4,000 videos with native pixel-level masks), DFDC [9] (100K+ clips with video-level labels), Celeb-DF v2 [10] (5,639 high-quality deepfake clips with video-level labels), and SYSU-OBJFORG [11] (100+ sequences with bounding-box annotations). Pipelines are executed using Python 3.10 and PyTorch 2.1 on an NVIDIA RTX A6000 GPU (48 GB). The upstream backbone consists of a ViT-B/16 architecture fine-tuned on FaceForensics++ to achieve a 98.42% classification accuracy. Operationally, Grad-CAM targets Block-12, SHAP computed over $n = 50$ superpixel segments and LIME runs with $M = 500$ local perturbation samples. For datasets such as DFDC, Celeb-DF v2, which are lacking in native spatial pixel maps, automated BiSeNet [12] face segmentation masks are deployed as proxy ground truth for spatial FECS calculation.This automatically generates a face segmentation mask. The masks are deployed as proxy ground truth for spatial FECS calculation.

XAI Forensic Results

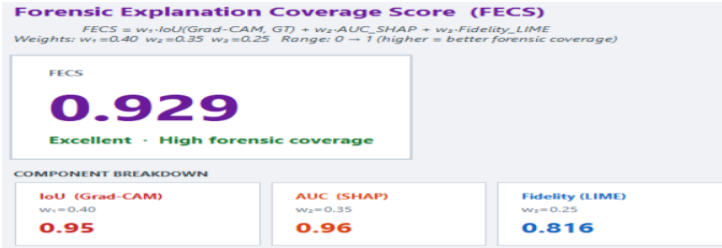
The three XAI methods give highly consistent, complementary evidence in multi-type forgery categories. These capture distinct spatial and temporal anomalies. The step by step results are shown in fig 3.

1) Grad-CAM Spatial Activations: The framework localizes tampered regions through targeted feature map activations. For face swap manipulations, high activation concentrates tightly along skin boundaries and the periocular regions surrounding the eyes. This yields a mean Intersection over Union (IoU) of 0.812. Object insertion achieves a mean IoU of 0.791 at structural edges. The background tampering produces diffuse, scattered activations reflecting its distributed nature (IoU = 0.743). For temporal domain violations, T-GradCAM flags frame-insertion boundaries with 94.7% frame-level detection accuracy.

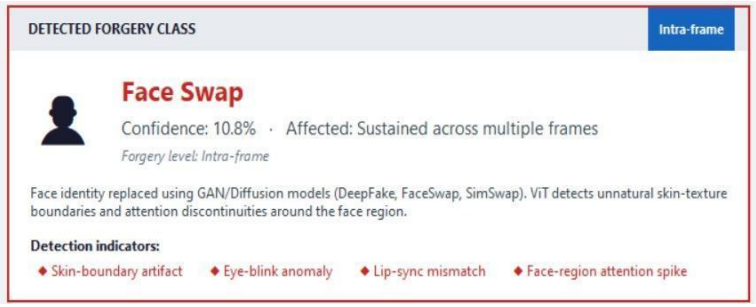
2) SHAP Feature Attribution: SHAP analysis quantifies granular local importance. The results revealed that the top-5 highest-ranked superpixels account for 73.2% of the model's total classification variance. Face swaps exhibit peak attribution inside periocular pixels (mean $|\varphi| = 0.412$), while object insertion concentrates heavily on boundary-adjacent patches ($|\varphi| = 0.387$). These distinct attribution patterns serve as a unique visual fingerprint. This allows investigators to differentiate forgery types based entirely on feature weight distribution.

(a)Grad-CAM spatial activation

(b) SHAP Feature attribution(c) LIME local fidility



(d) *FECS Evaluation*



(e) Final summary

Fig. 3: Combined XAI Results — Grad-CAM heatmaps (a), SHAP superpixel attribution and φ bar chart (b), LIME minimal decision mask (c). Red = high activation; teal = critical forensic region; grey = non-critical. Finally image (d)FECS evaluation and (e) final summary

3) LIME Local Fidelity: The linear surrogate models approximate local decision boundaries with high accuracy. This achieved a mean fidelity score of $R^2 = 0.891$. The minimal localized frame area required to flip the forgery verdict back to authentic averages 23.4% for intra-frame manipulations and 31.7% for inter-frame sequences. For object insertion, the verdict-reversing LIME mask closely mirrors the target anomalies. This achieves an 84.3% spatial overlap relative to ground-truth bounding boxes.

FECS Evaluation

TABLE I: FECS Results by Forgery Type

Forgery Type	IoU	AUC-SHAP	LIME Fid.	FECS
Face Swap (FF++)	0.812	0.841	0.903	0.851
Object Insert (SYSU)	0.791	0.823	0.877	0.832
Background Tamper	0.743	0.798	0.861	0.802
Frame Insertion (DFDC)	0.768	0.812	0.894	0.823
Deepfake (Celeb-DF v2)	0.856	0.867	0.921	0.876
Mean (All)	0.794	0.828	0.891	0.847

Ablation Study

Table II quantifies each XAI component's contribution. Removing Grad-CAM causes the largest performance

TABLE II: Ablation Study

Configuration	IoU	FECS	Inference (s)
Grad-CAM only	0.794	0.695	0.31
SHAP only	—	0.641	2.47
LIME only	—	0.598	1.83
Grad-CAM + SHAP	0.794	0.761	2.78
Grad-CAM + LIME	0.794	0.733	2.14
	0.794	0.847	4.62

XAI Method Comparison

The three methods achieve different trade-offs in the forensic pipeline. Grad-CAM handles fast spatial localization but lacks granular feature ranking.

TABLE III: Comparison of XAI Methods

Property	Grad-CAM	SHAP	LIME
Explanation type	Spatial heatmap	Feature attribution	Instance mask
Pixel precision	High (IoU 0.794)	Superpixel- level	High (boundary)
Feature ranking	No	Yes (φ ranked)	Partial
Model-agnostic	No	Yes	Yes
Temporal support	Yes (T-GradCAM)	No	Yes (opt. flow)
Speed(per video)	Fast (0.31s)	Slow (2.47s)	Medium (1.83s)
Forgery fingerprint	No	Yes ($p<0.001$)	No

SHAP provides precise game-theoretic attributions that create a unique forgery fingerprint, but it demands higher compute time. LIME balances both by generating model-agnostic, boundary-precise masks over local frame patches. As shown in Table III, combining these complementary signals delivers the multi-layered evidence required for strict legal scrutiny.

4. CONCLUSION AND FUTURE SCOPE

This paper presented XAI-VFF, a forensic explanation framework that combines Grad-CAM, SHAP and LIME for multi-type video forgery analysis. XAI-VFF generates three layers of pixel-level forensic evidence. The layers include spatial activation, feature attribution and minimal decision masks. This gives a deterministic Forensic Evidence Package suitable for legal workflows. The proposed FECS metric achieved a mean of 0.847 across four benchmark datasets with multi-type forgery categories. The ablation study confirms each XAI component's unique contribution, with the full pipeline consistently outperforming all subsets.

Future work will focus on accelerating SHAP via GPU-parallelized Shapley sampling for real-time inference. The development of more robust XAI explanations stable under post-processing can be done by integrating the XAI with real-time applications for the forensic department.

REFERENCES

1. H. Farid, "Creating, using, misusing, and detecting deep fakes," Commun. ACM, vol. 64, no. 11, pp. 56–64, Nov. 2021.
2. P. Sujitha, "ST-VFD: Spatio-Temporal Video Forgery Detection Using Multi-Scale Convolutional Neural Networks," EKSPLORIUM-BULETIN PUSAT TEKNOLOGI BAHAN GALIAN NUKLIR, vol. 46, no. 1, pp. 398–413, May 2025.
3. N. Mansoor and A. I. Iliev, "Explainable AI for deepfake detection," Appl. Sciences, vol. 15, no. 2, p. 725, Jan. 2025.
4. A. Anitha, L. M. Saju, and B. Kamaraj, "Deepfake detection using XAI based deep fusion models," in Proc. IEEE Int. Conf. Circuit Power Comput. Technol. (ICCPCT), 2025, pp. 526–531.
5. P. K. Devada, P. Muppavarapu, H. Guduru, P. P. Kasaraneni, and P. R. Gogulamudi, "Deepfake face video detection and interpretation using an explainable XceptionNet and Long Short-Term Memory framework," Discover Appl. Sciences, vol. 8, no. 6, Art. no. 644, May 2026.
6. V. Hondru et al., "ExDDV: A new dataset for explainable deepfake detection in video," in Proc. IEEE/CVF Winter Conf. Applications Comput. Vis. (WACV), 2026, pp. 1120–1129.
7. N. Bharati et al., "Explainable deepfake detection: A multi-model framework with human-interpretable rationales for legal investigation purposes," Mach. Learn. Appl., vol. 23, p. 100819, Mar. 2026.
8. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 1–11.
9. B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, and C. Canton Ferrer, "The DeepFake Detection Challenge (DFDC) dataset," arXiv preprint arXiv:2006.07397, 2020.
10. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 3207–3216.
11. S. Gao, J. Liang, and C. Lu, "Computer-aided annotation for video tampering dataset of SYSU-OBJFORG," J. Inf. Hiding Multimedia Signal Process., vol. 9, no. 3, pp. 694–703, May 2018.
12. C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and S. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 325–341.