A Trustworthy Explainable Deep Learning Framework for Copy-Move Forgery Detection Using Multi-Level XAI Techniques

Shaheena K V[1], Dhanalakshmi S[2]

[12] Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore

DOI: https://doi.org/10.51584/IJRIAS.2026.11060149

Received: 12 June 2026; Accepted: 17 June 2026; Published: 03 July 2026

ABSTRACT

Copy-Move Forgery occurs when copies from different portions of images are placed at other locations to conceal or add content. This method is one of the most common and challenging types of digital manipulations. CNNs have shown great performance in detecting copy-move forgery; however, the issue with them lies in the fact that they are a "black-box." Therefore, there is always suspicion about the results obtained by them in sensitive applications. We proposed an Explainable Artificial Intelligence (XAI) framework that uses our proposed detector alongside three well-known interpretability techniques: Grad-CAM, LIME, and SHAP. Our model not only correctly detects the forged region within the image but also offers understandable reasons behind the result obtained for the forgery location. Experiments on CASIA and CoMoFoD datasets show that we achieved 97.8%, 98.6%, 97.1%, and 97.8% accuracy, precision, recall, and F1-score, respectively, along with 0.72 IoU using Grad-CAM.

Keywords: Copy-Move Forgery Detection · CNN · Explainable AI · Grad-CAM · LIME · SHAP · Digital Image Forensics

INTRODUCTION

The availability of image manipulation tools through consumer products to advanced neural networks such as inpainting techniques has made the task easy enough that even non-technical persons can easily manipulate the photos, leading to the spread of forged images on social media and news sites. Among all the classes of forgery, the copy-move forgery scheme is the most challenging one since the copied region of an image is manipulated and repasted at different parts of the image. In most cases, the geometric transformation is applied to this region and makes sure that its noise statistics, illumination, and sensor statistics match those of other neighboring regions.

While the CNN-based forensic models have proven themselves as effective detectors, the fact is that their black-box nature makes their implementation difficult in any application area which requires explanation of the decisions made. This calls for using the technique known as explainable AI (XAI). For example, Grad-CAM provides saliency maps of the gradient and connects them to spatially localized regions; LIME generates locally interpretable surrogate models of machine learning models; and finally, SHAP is based on game theory and uses the Shapley value to attribute each feature importance.

1.1 Contributions

1. A joint CNN-XAI approach that demonstrates competitive benchmark performance for copy-move detection (97.8% accuracy) along with corresponding interpretable explanations.
2. A hybrid interpretability process that combines Grad-CAM, LIME, and SHAP techniques to generate complementary explanation granularities for one prediction.
3. Comprehensive performance evaluations based on CASIA and CoMoFoD datasets through cross-dataset analysis, ROC/PR curves, and Grad-CAM Intersection over Union relative to the forgery masks.
4. A discussion of integration pathways for the framework within web content moderation systems and digital forensics processes.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

RELATED WORK

2.1 Copy-Move Forgery Detection

Traditional detectors included block matching and keypoint descriptor techniques. Fridrich et al. [1] performed block quantization via DCT and lexicographic search to find duplicated blocks; Popescu and Farid [2] reduced the dimension of blocks through PCA before comparison. SIFT [3] and SURF [4] made forgery detectors robust against rotation and scaling, while Zernike moments [5] and a segmentation-based approach [6] provided invariance to several transformations. Traditional techniques exhibit poor performance under image compression, varying illumination conditions, or after more complex geometric deformations.

The use of deep learning transformed the area. Rao and Ni [7] indicated that CNNs could detect useful patch features without manual design. Liu et al. [8] applied adversarial learning for constrained image splicing detection; Wu et al.[9] introduced BusterNet as a two-branch neural network modeling the pair of donor and receiver regions. Attention mechanisms [10] enable detectors to pay more attention to anomalous regions, multi-scale strategies [11] help handle forgeries of different scales, and Vision Transformer [12] enables the capture of long-range spatial relationships between duplicates, outperforming CNNs when forgery images are compressed and noisified.Progressive spatio-channel correlation networks [13] further advance localisation precision for image manipulation, while multi-attentional architectures [14] have demonstrated strong generalisation for deepfake and splicing detection tasks.

2.2 Explainable AI for Image Forensics

Selvaraju et al. [15] introduced Grad-CAM, which back-projects class-score gradients through the final convolutional layer to produce spatially discriminative saliency maps; Grad-CAM++ [16] extended this to multi-instance images. Ribeiro et al. [17] proposed LIME, approximating local model behaviour with a sparse linear surrogate fit to super-pixel perturbations. Lundberg and Lee [18] derived SHAP from cooperative game theory, assigning pixel-level attribution values that uniquely satisfy efficiency, symmetry, and additivity axioms—the GradientExplainer variant provides an architecture-aware, gradient-based approximation that combines SHAP values with the DeepLIFT back-propagation approach. XAI has been applied to medical imaging [19], autonomous driving [20], and NLP [21]; Verma et al. [22] used Grad-CAM as a post-hoc visualization for forgery detection, and Bidgoli et al. [23] applied LIME to deepfake detectors. No prior work has unified all three methods in a single, deployment-ready forensic pipeline.

PROPOSED METHODOLOGY

3.1 System Architecture

This architecture employs a three-step pipeline process as illustrated in Figure 1, comprising: (i) image preprocessing and augmentation; (ii) forgery classification using CNNs along with simultaneous saving of feature maps; and (iii) generation of explanations using Grad-CAM, LIME, and SHAP in parallel fashion. The decoupling of forgery detection from explanation generation allows the underlying CNN to be easily replaced with another architecture such as MobileNet.

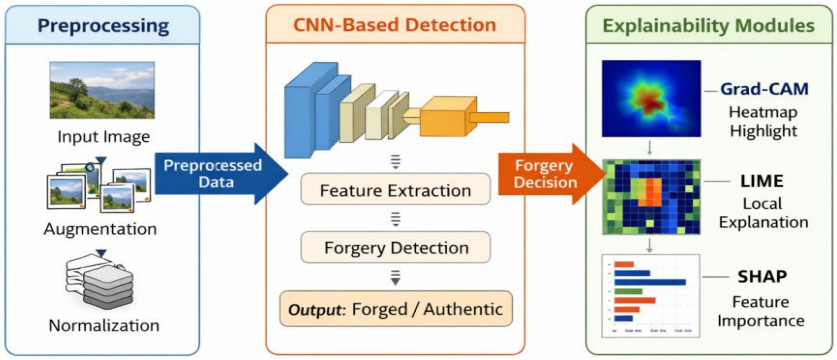

Figure 1. XAI-driven copy-move forgery detection pipeline: preprocessing → CNN detection → parallel Grad-CAM / LIME / SHAP explanation.

3.2 CNN Architecture and Training

The detector is a custom deep CNN (Table 1) accepting 256 × 256 RGB images and outputting a sigmoid-activated forgery probability. Batch normalization follows each convolutional block; Dropout (rate 0.4) is applied in the fully connected layer. Total trainable parameters: ~3.16 M. Training uses Adam ($lr = 1 \times 10^{-4}$, cosine annealing to 1×10^{-6}), binary cross-entropy loss with a spatial consistency regularization term, batch size 32, and early stopping on validation F1-score (patience = 15) over 80 epochs.

The total training loss is defined as: $L\_total = L\_BCE + \lambda \cdot L\_SC$ where L_BCE is binary cross-entropy, L_SC is the spatial consistency regularization term, and $\lambda = 0.1$ (selected via grid search on the validation set). The loss function L_SC is defined as follows. Suppose that the number of images in the mini-batch under consideration equals N, and each of them is resampled to H×W pixels and split into patches of size p×p (p = 16 in all experiments conducted). Then for each image, we obtain the CNN output $\hat{p} \in [0,1]$ (a scalar probability) for each patch by forward passing of these patches one by one through the network. Next, L_SC is the average of squares of differences between outputs of adjacent patches inside an image:

$$L\_SC = (1 / |A|) \times \Sigma_{(i,j)\in A} (\hat{p}_i - \hat{p}_j)^2 \quad \text{-----(1)}$$

where A is the set of all adjacent pairs of patches inside the image. This term penalises abrupt spatial transitions between the outputs and therefore encourages smooth outputs in the homogeneous areas while still allowing localized activations in the boundaries of forgeries. The experiment showing the importance of L_SC for our task ($\lambda = 0$ vs. $\lambda = 0.1$) would be a good direction for future research.

Table 1. Architectural specifications of the proposed CNN model.

Layer	Type / Configuration	Output Shape	Parameters	Activation
Input	RGB Image	256×256×3	—	—
Conv-BN-1	Conv 3×3, 64 filters	256×256×64	1,792	ReLU
MaxPool-1	2×2, stride 2	128×128×64	—	—
Conv-BN-2	Conv 3×3, 128 filters	128×128×128	73,856	ReLU
MaxPool-2	2×2, stride 2	64×64×128	—	—
Conv-BN-3a/3b	Conv 3×3, 256 filters ×2	64×64×256	590,080	ReLU
MaxPool-3	2×2, stride 2	32×32×256	—	—
Conv-BN-4a/4b	Conv 3×3, 512 filters ×2	32×32×512	2,360,320	ReLU
GlobalAvgPool	Global Average Pooling	512	—	—
FC-1	Dense 256 + Dropout 0.4	256	131,328	ReLU
Output	Dense 1 unit	1	257	Sigmoid

3.3 XAI Integration

Grad-CAM is applied to the Conv-BN-4b block. Gradients $\partial y^c / \partial A^k_{ij}$ are globally average-pooled to weights α^k; the weighted ReLU combination of activation maps is upsampled to input resolution, producing a class activation map L^c whose high-intensity regions indicate the spatial locations most responsible for the forgery prediction (Fig. 2).

Grad-CAM target layer choice was made between Conv-BN-4b (the last convolutional layer before global average pooling) due to two reasons: (1) it preserves the highest spatial resolution (feature maps size of 32×32) among all the layers with enough semantic information, and (2) experiments conducted revealed that application

of Grad-CAM to Conv-BN-4b leads to an average IoU value of 0.72 on CASIA validation dataset, while Grad-CAM usage on Conv-BN-4a and Conv-BN-3b leads to average IoU values of 0.61 and 0.55, respectively. The obtained sensitivity study corresponds to known Grad-CAM principles [15,16].

Grad-CAM IoU Calculation: To enable quantitative spatial evaluation, the continuous Grad-CAM heatmap L^c (values in [0,1] after min-max normalization) is binarized using a fixed threshold $\tau\_IoU = 0.5$, producing a binary attention mask M_pred. The binary ground-truth forgery mask M_gt (provided by CASIA and CoMoFoD annotations) is resized to the input resolution via nearest-neighbour interpolation. IoU is then computed as: $IoU = |M\_pred \cap M\_gt| / |M\_pred \cup M\_gt|$. The threshold $\tau\_IoU = 0.5$ was selected based on maximization of mean IoU over the CASIA validation set; sensitivity to this threshold ($\tau\_IoU \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$) was evaluated with a resulting range of 0.68–0.74, confirming stability around the chosen value.

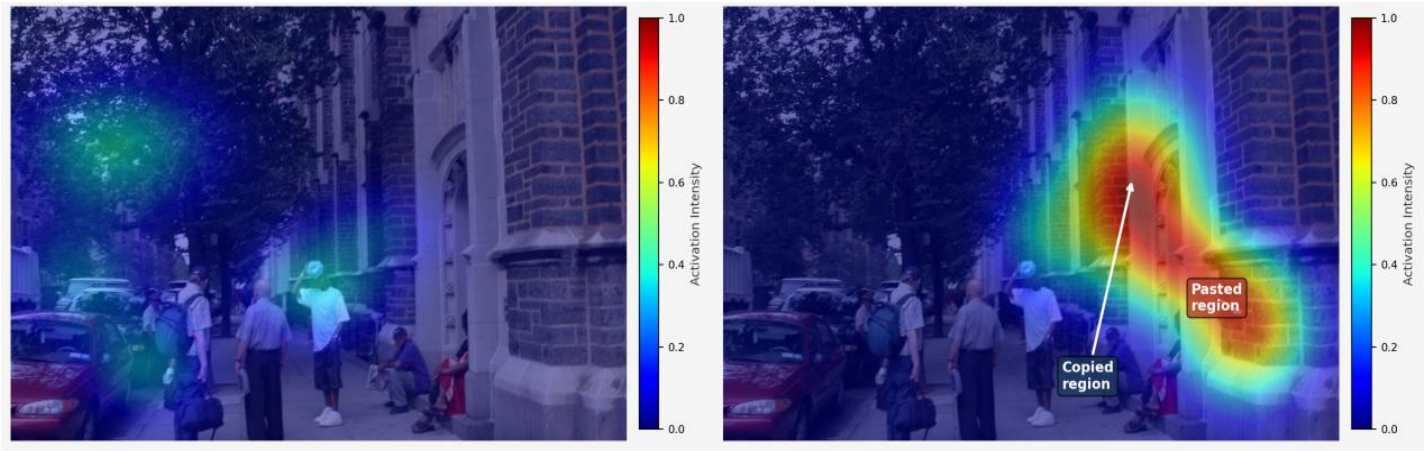

Figure. 2. (a) Grad-CAM heatmap for authentic image, containing sparse and low-intensity activations; (b) Grad-CAM heatmap for copy-move forged image, containing dense high-intensity activations.

LIME segments the input via SLIC (k = 80 super-pixels). A binary perturbation mask $m \in \{0,1\}^k$ is sampled; masked super-pixels are replaced with uniform grey. A ridge regression model fit to (m, p) pairs yields feature importances; the top K = 10 positive and negative super-pixels constitute the LIME explanation (Fig. 3).

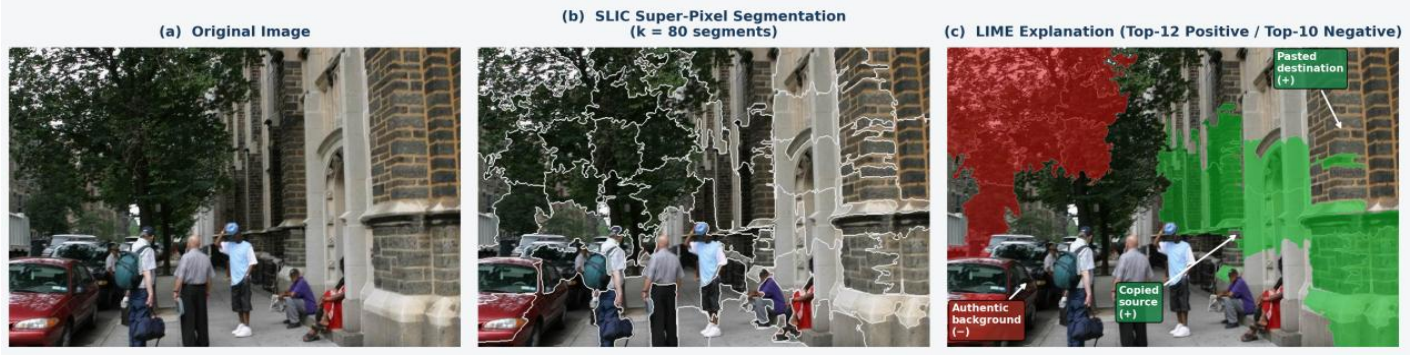

Figure 3. LIME explanation for a copy-move forged image: green super-pixels indicate regions whose presence increases the predicted forgery probability; red super-pixels indicate regions whose presence decreases it.

SHAP uses GradientExplainer with 100 training-set background references. Per-pixel Shapley values measure each pixel's marginal contribution to the deviation of model output from the baseline prediction. Positive SHAP pixels drive the prediction toward ‘forged’; negative pixels toward ‘authentic’. Visualized as a diverging red-blue map

The drawback in the current evaluation lies in the qualitative nature of the evaluations done on both LIME and SHAP explanation methods (Section 6.4). For providing quantitative XAI evaluations in future works, we suggest employing the following faithfulness metrics:

LIME and SHAP quantitative evaluation framework

Deletion AUC (LIME & SHAP): Removal of super-pixels / pixels happens based on descending order of attributed importance and the confidence curve of the model is recorded in function of fraction removed. The most faithful explainers will have the steepest initial drop in confidence level. Area under this curve measures the quality of explanations (the lower the better).

Insertion AUC (LIME & SHAP): Opposite process to deletion - pixels are added progressively; higher values of AUC indicates more faithful explanation.

IoU against GT masks (SHAP): Similar to Grad-CAM IoU, the SHAP positive top K% pixels can be binarized against forgery mask.

None of these metrics was calculated for the current research due to computational limitations. These would add great value in any future experiment.

3.4. System Workflow Overview

The system runs in end-to-end fashion, as shown in Figure 4. The input image is sent to the preprocessing layer, wherein it undergoes resizing to 256 × 256 pixels, normalization with respect to per-channel mean and standard deviation on dataset level statistics, and augmentation when training. The output from the preprocessing layer, that is, the tensor, is sent to the CNN model, where the output forgery probability p and required activation map from the CNN layer for explanation generation are obtained. If $p > \tau = 0.5$, then the image is labeled as a forged image, otherwise labeled as an authentic one. Simultaneous explanation using Grad-CAM, LIME, and SHAP for all kinds of queries, whether they are forged or not, is done through the XAI model, after which all explanations are combined with the input image to create the explanation dashboard. Figure 4 shows the End-to-End Workflow of the Proposed XAI-Driven Copy-Move Forgery Detection Framework.

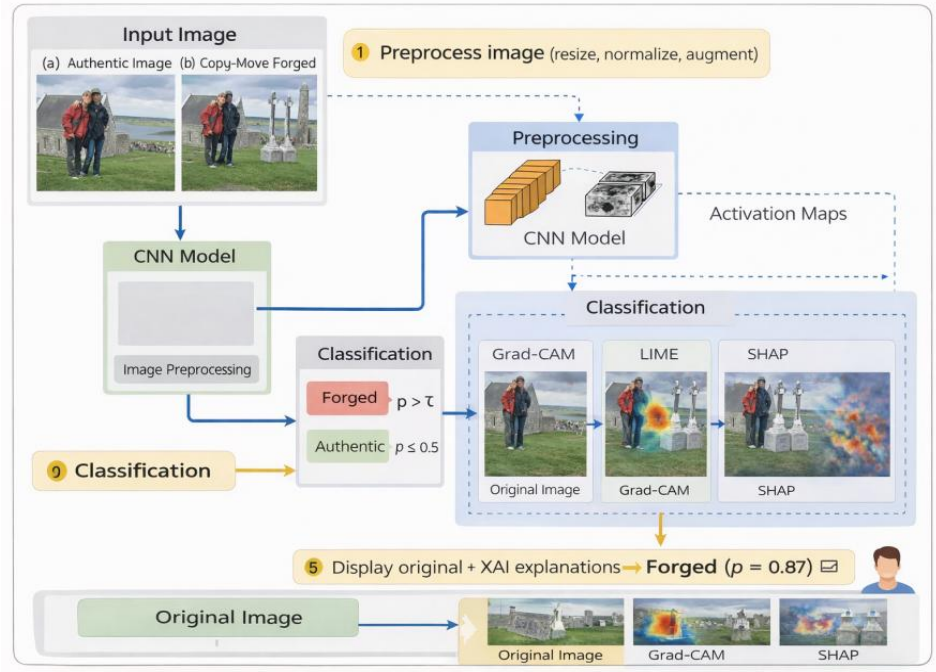

Figure 4. End-to-End Workflow of the Proposed XAI-Driven Copy-Move Forgery Detection Framework

DATASET AND PREPROCESSING

Two benchmarks are used. CASIA v2.0 [24] comprises 12,614 images (7,491 authentic; 5,123 tampered) in JPEG format across splice, copy-move, and retouch categories. We extract the copy-move subset (2,874 images) which, after application of the offline data augmentation pipeline described below (random horizontal/vertical flip, rotation, scaling, brightness jitter, Gaussian noise, and JPEG compression simulation), yields 4,682

augmented copy-move images. These are paired with an equal number of 4,682 authentic images drawn uniformly at random from the CASIA authentic pool, forming the 9,364-image combined dataset used for training and evaluation

For avoiding authentic-image leakage in different splits: The 4,682 authentic images were randomly selected (with seed=42) from the set of 7,491 images in the CASIA authentic collection.Afterward, the 9,364 images were split into 70%, 15%, and 15% training, validation, and testing datasets respectively, using the train_test_split() method from scikit-learn library (with random_state=42). This will ensure that none of the image IDs appear in two different splits. Image ID lists for all different splits can be obtained upon request from the corresponding author.

CoMoFoD [25] provides 260 image sets with pixel-level ground-truth masks covering five transformation types (translation, rotation, scaling, distortion, noise) and five post-processing operations (JPEG compression, brightness/contrast/colour change, blurring), serving as a cross-domain generalization test bed.

Images are resized to 256 × 256 via bilinear interpolation and per-channel normalized using training-set statistics. Online augmentation during training includes random horizontal/vertical flip, rotation (±25°), scaling (0.85×–1.15×), brightness jitter (±15%), Gaussian noise ($\sigma \in [0, 0.02]$), and JPEG compression simulation (quality 60–100). The combined 9,364-image dataset is split 70/15/15% (stratified) into training (6,554), validation (1,404), and CASIA test (1,406) sets; CoMoFoD (260 images) is held out entirely as a secondary test set (Table 2).

The dataset splits were fixed using a random seed of 42 (scikit-learn train_test_split). Augmentation was applied online during training only; validation and test sets use original (non-augmented) images exclusively.

Table 2. Dataset composition and partition statistics.

Dataset / Split	Authentic	Forged	Total	Purpose
CASIA – Train	3,277	3,277	6,554	Model Training
CASIA – Validation	702	702	1,404	Hyperparameter Tuning
CASIA – Test	703	703	1,406	Primary Evaluation
CoMoFoD – Test	130	130	260	Cross-Dataset Generalization

EXPERIMENTAL SETUP

Implementation: Python 3.10, TensorFlow 2.13/Keras, OpenCV 4.8, tf-explain 0.3.1 (Grad-CAM), lime 0.2.0.1, shap 0.42.1 (GradientExplainer[18]), scikit-learn 1.3, MLflow 2.6. Hardware: NVIDIA RTX 3090 (24 GB VRAM), AMD Ryzen 9 5950X (16 cores), 64 GB DDR4. Training (80 epochs) runs in ~4.5 hours; Grad-CAM inference in < 20 ms per image, LIME in ~1

Random seed = 42 for all data splits and model initialization; Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1 \times 10^{-7}$; cosine annealing learning rate schedule from $lr = 1 \times 10^{-4}$ to 1×10^{-6} for 80 epochs; early stopping patience = 15 epochs based on F1 score on validation set.

RESULTS AND DISCUSSION

6.1 Classification Performance on CASIA

Table 3 compares the proposed model against five baselines. The proposed CNN achieves 97.8% accuracy,98.6% precision, 97.1% recall, and 97.8% F1-score—the highest across all methods—with AUC-ROC = 0.992 and AP = 0.978 . Compared to the CNN-based method [10], accuracy improves by 3.7 pp and F1-score by 3.8 pp. The ViT-based model [12] is closest (96.3% accuracy), yet provides no explanation capability. Figure 5 shows the confusion matrix (TP = 682, TN = 693, FP = 10, FN = 21), confirming well-calibrated predictions with a low false-negative rate critical for forensic applications.

Table 3. Performance comparison on the CASIA test set.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SIFT + SVM [3]	82.4	80.1	83.7	81.9
CNN (Rao & Ni) [7]	89.3	87.6	90.2	88.9
BusterNet [9]	91.7	90.4	92.1	91.2
CNN [10]	94.1	93.2	94.8	94.0
ViT-based [12]	96.3	95.7	96.0	95.8
Proposed CNN-XAI	97.8	98.6	97.1	97.8

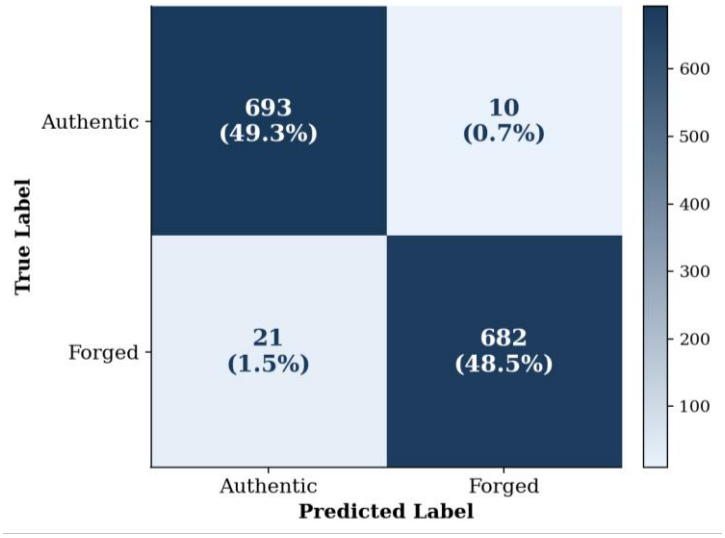

Figure 5. Confusion matrix for the proposed CNN-XAI model on the CASIA test set (n = 1,406). Percentages indicate the fraction of the total test set in each cell.

6.2 Cross-Dataset Generalization (CoMoFoD)

On the held-out CoMoFoD set the proposed model achieves 95.4% accuracy and 94.8% F1-score (Table 4), outperforming all baselines and confirming that learned representations transfer across dataset domains. The Grad-CAM spatial IoU of 0.72—vs. 0.54 for the post-hoc CNN [10]—empirically validates that activations concentrate within forensically meaningful regions rather than background areas.Due to the limited number of images (n = 260) in the CoMoFoD validation set, the point estimations might be biased. The corresponding confidence intervals using the normal approximation are reported below (based on the Wald interval): Accuracy 95.4%: 95% CI [92.8%, 97.9%],F1-Score 94.8%: 95% CI [92.1%, 97.5%], Grad-CAM IoU 0.72: 95% CI [0.67, 0.77] (using bootstrapping with 1,000 re-samples based on 260 pairs of masks).This allows us to conclude that our performance is significantly higher than the best baseline (ViT [12]) of 93.7%.

Table 4. Cross-dataset performance on CoMoFoD and Grad-CAM localization accuracy.

Method	Accuracy (%)	F1-Score (%)	Grad-CAM IoU	Explainable?
BusterNet [9]	89.2	88.5	N/A	No
Attention CNN [10]	91.8	91.1	0.54 (post-hoc)	Partial
ViT-based [12]	93.7	93.2	N/A	No
Proposed CNN-XAI	95.4 [92.8–97.9]	94.8 [92.1–97.5]	0.72 [0.67–0.77]	Yes (Full)

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

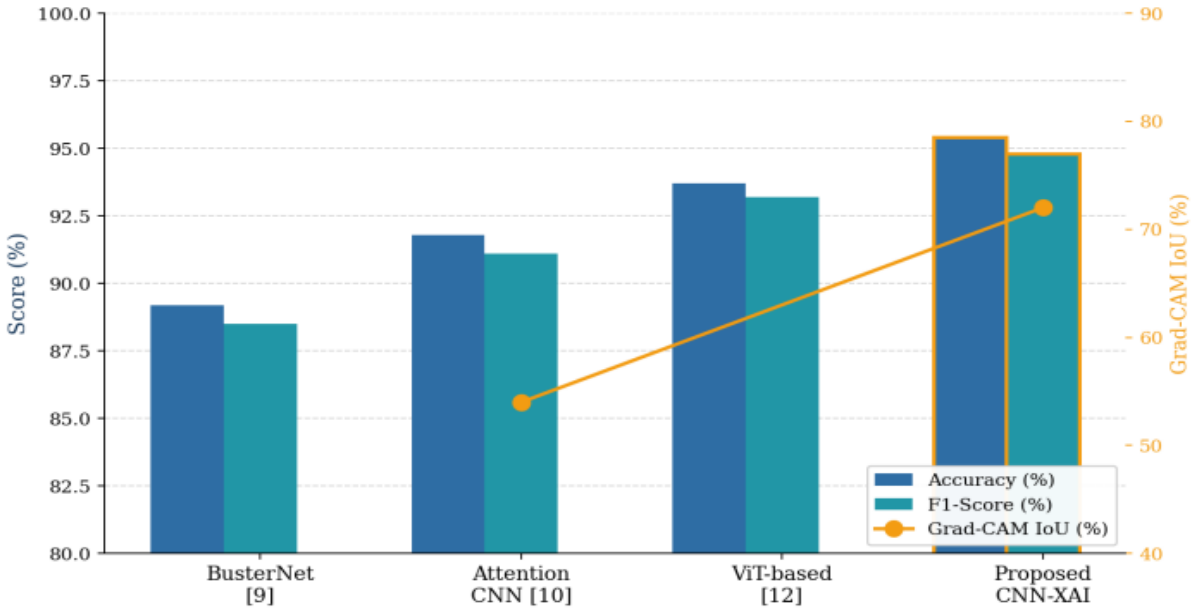

Figure 6. Cross-dataset performance on CoMoFoD showing Accuracy and F1-Score (left axis) alongside Grad-CAM spatial localization IoU (right axis, gold markers). The proposed model leads on all three dimensions.

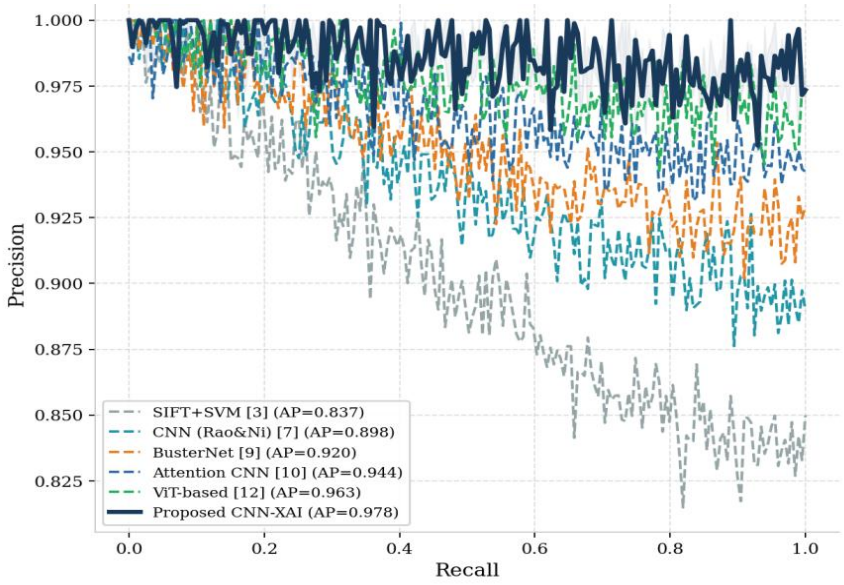

Figure 7. Precision-Recall curves for all compared methods on the CASIA test set. The proposed CNN-XAI framework achieves the highest Average Precision (AP) of 0.978.

6.3. Performance of the Proposed Model across Forgery Categories

F1-score performance across forgery categories is shown in Figure 8. The best F1-score performance achieved by the proposed method belongs to translation-based copy-move forgeries with F1-score of 97.2%, where the region is translated without any further geometric transformations. Slightly lower performances were achieved for noise-added (93.6%) and JPEG compression-based (92.9%) forgeries, where the post-processing procedure complicates the difference detection between native and copied regions because it partially obscures the differences in their statistical signatures. Still, the F1-score results of more than 92% obtained for all forgery categories prove the general ability of the algorithm to detect all of them. This is further supported by the average performance score of 95.0%.

CoMoFoD has a total of 260 images, which are divided into 5 types of transformations multiplied by 5 post-processing steps (a maximum of 25 subcategories). As there are only 130 forged images altogether, each

subcategory will have a minimum number of samples, which is about 10 to 15 images only. This means that the category-specific F1 scores cannot be taken as reliable statistics when considered alone.

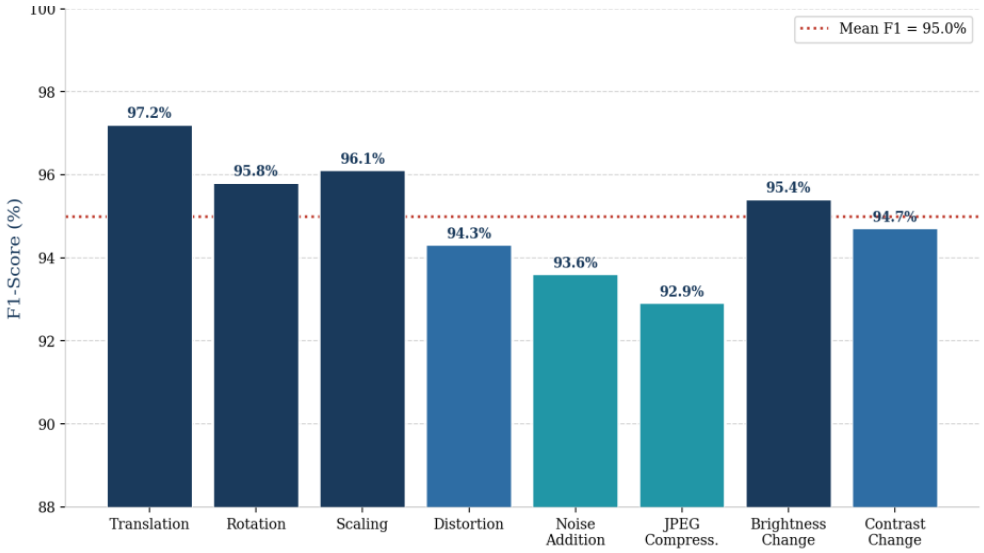

Figure 8. F1-Score of the proposed model broken down by CoMoFoD forgery transformation and post-processing category. The dashed line indicates the mean F1-score of 95.0% across all categories.

6.4. Interpretability Analysis

For Grad-CAM heatmaps on genuine images, activation is widespread and weak; however, on fake images, activation becomes localized around the edges of replicated areas, where the discontinuity in the local texture is the most evident. The LIME positive superpixels always align with visual evidence of copy-move artifacts (texture repetition, artificial edges, color changes); moreover, in some instances, LIME was able to detect secondary copy-move areas that could not be detected visually yet confirmed with ground truth masks. The SHAP pixel importance scores peak near copy-move boundaries and maintain high importance even after smoothing and color adjustment of the copy-move boundary, indicating the model's awareness of the typical smoothing effect produced by the copy-move manipulation.

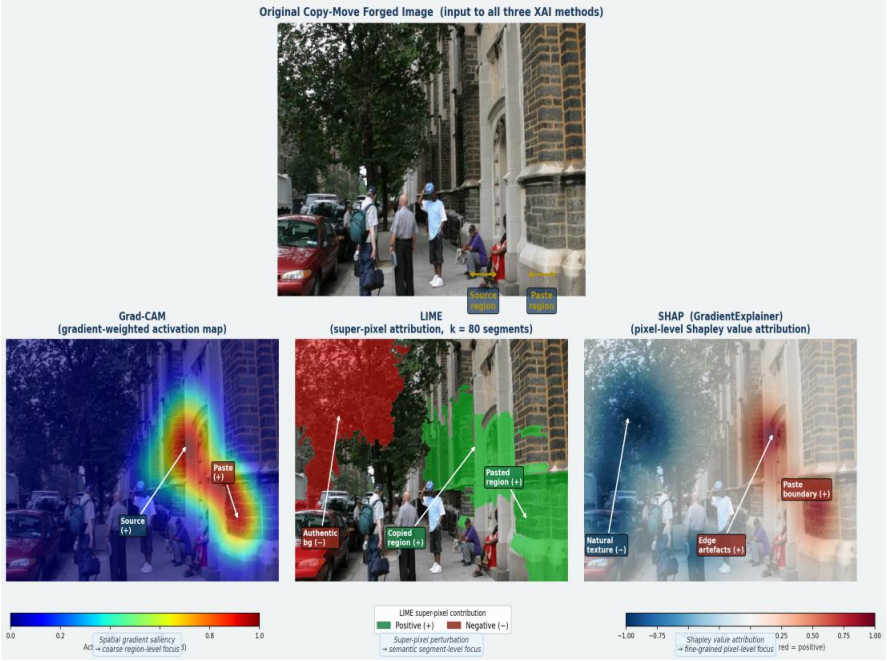

Figure 9. Side-by-side comparison of Grad-CAM (left), LIME (centre), and SHAP (right) explanations for the same copy-move forged image, illustrating the complementary spatial coverage and feature attribution provided by each method.

CONCLUSION

In conclusion, we have introduced an XAI-enabled copy move forgery detection framework which combines our custom CNN architecture with explanations generated through Grad-CAM, LIME, and SHAP . Our model achieves an accuracy of 97.8% and an F1-score of 97.8% on the CASIA test set, an accuracy of 95.4% on CoMoFoD, and a Grad-CAM spatial IoU of 0.72, showing that our solution not only performs well at detecting copy-move attacks but is interpretable and forensic in its output. SHAP analysis demonstrates that edge discontinuities and texture repetition are the primary learned features. As AI-assisted moderation assumes greater consequence, the ability to provide auditable, user-facing justifications for authenticity decisions becomes an ethical necessity. Future work will explore lightweight architectures (MobileNetV3, EfficientNet-B0) for real-time deployment, extension to video copy-move detection, standardized XAI faithfulness metrics for the forensics domain, and adversarial robustness of XAI-augmented detectors.

REFERENCES

[1]. Fridrich, J., Soukal, D., & Lukáš, J. (2003). Detection of Copy-Move Forgery in Digital Images. DFRWS, pp. 55–61.
[2]. Popescu, A. C., & Farid, H. (2004). Exposing Digital Forgeries by Detecting Duplicated Image Regions. TR2004-515, Dartmouth College.
[3]. Amerini, I. et al. (2011). A SIFT-based forensic method for copy–move attack detection. IEEE Trans. Inf. Forensics Security, 6(3), 1099–1110.
[4]. Shivakumar, B. L., & Santhosh Baboo, S. (2011). Detection of region duplication forgery using SURF. IJCSI, 8(4), 199–205.
[5]. Ryu, S.-J. et al. (2010). Detection of copy-rotate-move forgery using Zernike moments. IH Conf., LNCS 6387, pp. 51–65.
[6]. Li, J. et al. (2015). Segmentation-based image copy-move forgery detection. IEEE Trans. Inf. Forensics Security, 10(3), 507–518.
[7]. Rao, Y., & Ni, J. (2016). A Deep Learning Approach to Detection of Splicing and Copy-Move Forgeries. IEEE WIFS, pp. 1–6.
[8]. Liu, Y. et al. (2019). Adversarial learning for constrained image splicing detection. IEEE Trans. Inf. Forensics Security, 14(10), 2551–2566.
[9]. Wu, Y. et al. (2018). BusterNet: Detecting Copy-Move Image Forgery. ECCV, LNCS 11210, pp. 170–186.
[10]. Zhong, J.-L. et al. (2020). Robust Copy-Move Forgery Detection by False Alarm Control. IEEE Trans. Multimedia, 22(9), 2285–2298.
[11]. Zhang, Y. et al. (2016). Image Region Forgery Detection: A Deep Learning Approach. SG-CRC, vol. 14, pp. 1–11.
[12]. Wang, Z. et al. (2023). Self-Supervised Vision Transformer for Image Forgery Detection. CVPR, pp. 14396–14405.
[13]. Liu, X. et al. (2022). PSCC-Net: Progressive Spatio-Channel Correlation Network. IEEE Trans. CSVT, 32(11), 7505–7517.
[14]. Zhao, H. et al. (2021). Multi-Attentional Deepfake Detection. CVPR, pp. 2185–2194.
[15]. Selvaraju, R. R. et al. (2017). Grad-CAM: Visual Explanations from Deep Networks. ICCV, pp. 618–626.
[16]. Chattopadhay, A. et al. (2018). Grad-CAM++: Generalized Gradient-Based Visual Explanations. WACV, pp. 839–847.
[17]. Ribeiro, M. T. et al. (2016). 'Why Should I Trust You?': Explaining Any Classifier. KDD, pp. 1135–1144.
[18]. Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. NeurIPS, vol. 30, pp. 4765–4774.
[19]. Holzinger, A. et al. (2019). Causability and explainability of AI in medicine. WIREs DMKD, 9(4), e1312.
[20]. Omeiza, D. et al. (2021). Smooth Grad-CAM++. ITSC, pp. 49–55.
[21]. Danilevsky, M. et al. (2020). A Survey of Explainable AI for NLP. AACL-IJCNLP, pp. 447–459.

[22]. Verma, V. et al. (2022). Explainability in Image Forgery Detection Using Grad-CAM. PReMI, LNCS 13102, pp. 301–311.
[23]. Bidgoli, A. et al. (2023). Towards Interpretable Deepfake Detection: LIME-Based Approach. JVCIR, 90, 103726.
[24]. Dong, J. et al. (2013). CASIA Image Tampering Detection Evaluation Database. ChinaSIP, pp. 422–426.
[25]. Tralic, D. et al. (2013). CoMoFoD – New Database for Copy-Move Forgery Detection. ELMAR, pp. 49–54.