On Visual Saliency Maps for Identifying Fidelity of Deepfake Detection Datasets

Agniva Banerjee[1]
agniva24@iiserb.ac.in
Samiran Das[2]
samiran@iiserb.ac.in
Akshay Agarwal[2]
akshay@iiserb.ac.in

[1] Department of Electrical Engineering and Computer Science,
IISER Bhopal, India

[2] Department of Data Science and Engineering,
IISER Bhopal, India

Abstract

Forged and generated images, such as deepfakes and Photoshop forgeries, pose significant societal threats. Conventional fake image detection models cannot develop trust among users because they only classify the images without providing explanations. In this paper, we particularly address this concern and demonstrate that explainable AI (XAI) methods, including SHAP, LIME, and integrated gradients (IG), can highlight the specific regions in an image that influence a model's classification, and can identify the fidelity of the deepfake images. These attribution maps pinpoint manipulated regions in facial images and highlight geometric abnormalities. Furthermore, we analyzed the differences between different types of deepfake image datasets using cross-dataset experiments on Photoshop-generated, Celeb-DF, and FF++ datasets. Our findings show that models trained on deepfake images demonstrate superior robustness and generalization, especially in cross-dataset scenarios, compared to Photoshop-generated images. This research highlights how the complexity and structure of image manipulations directly affect a detection model's performance.

1 Introduction

The rapid advancement of generative AI has led to the generation of realistic, high-quality deepfake images [1, 2]. Detecting fake images is challenging as generative AI becomes sophisticated [3, 4]. However, the mere identification of deepfake images without in-depth insight does not provide an accurate framework for an extensive assessment of the models. Understanding why the model flags an image as fake is imperative to make the model trustworthy [5, 6]. Saliency maps highlight significant portions, revealing information about the bias and fairness of the models. Several works proposed diverse deep learning models to identify fake images, but few have analyzed feasible explanations [6, 7, 8]. Although preliminary approaches relied on hand-crafted features and machine learning models [9], recent works use advanced deep neural networks [10, 11, 12, 13]. Although these models perform well on images with noticeable irregularities, such as neighborhood inconsistency, geometric irregularity, and unnatural positioning of face parts [7, 14], they do not generalize well to images generated by more sophisticated generative models. Further, most of these approaches do not provide a proper explanation and do not build trust among end-users in such scenarios.

XAI models highlight salient image regions [15], providing valuable information on the behavior of the model [16]. Some studies have re-evaluated the adequacy of these visual feature maps for deepfake detection by comparing attribution outputs across models [17]. These concerns motivate us to develop an integrated feature attribution system to improve deepfake detection. We analyze various deep learning models, including shallow CNN, AlexNet [18], VGG19 [19], ResNet50 [20], ResNet101 [20], and EfficientNetV2 [21] to detect deepfakes of varying quality. We identify the attribution maps using SHAP [22], LIME [23], and IG [24], reevaluate model performance, and investigate the generalizability of the models by focusing on highlighted regions. The experimental results indicate that the attribution maps improve the models' detection accuracy. However, our primary goal is not to benchmark the latest architectures, but to systematically study how attribution methods influence models of varying depth and complexity, thereby isolating their role in improving trust and interpretability.

The key contributions of this work are as follows:

- We introduce a novel workflow that integrates XAI techniques into a deep learning pipeline to significantly enhance the accuracy and interpretability of forgery detection. Our workflow identifies visual saliency maps that reflect fiducial portions, geometric abnormalities, and manipulated facial regions, thereby providing a human-discernible understanding and building user trust.

- We utilize the attribution maps to evaluate the fidelity of the deepfake image datasets quantitatively. We guide the models in distinguishing different types of forged datasets using these attribution maps generated by the aforementioned XAI methods. This approach is validated through a comprehensive series of experiments conducted on deepfake datasets (FF++ and Celeb-DF) and Photoshop-generated (PG) datasets.

- We conducted extensive experiments demonstrating that models trained on deepfake datasets generalize better to PG forgeries than the reverse. We performed cross-dataset experiments to analyze these two types of fake images with distinct structural and feature-level differences, thereby addressing a critical gap in prior work.

This paper presents a systematic way to understand different deepfake types using XAI models, paving the way for developing advanced and reliable solutions. Our paper attempts to create a novel framework for combating the threat of deepfakes and making the digital environments more trustworthy.

2 Related Works

Several recent studies have proposed various ML and DL approaches for deepfake detection [25, 26, 27]. Safwat *et al.* [28] introduced a hybrid DL model that combines ResNet50 and GAN with channel-wise attention mechanisms to improve the detection accuracy of fake faces. Similarly, Ishrak *et al.* [29] combined CapsuleNet with Long-Short Term Memory (LSTM) to analyze deepfake video frames. Rafique *et al.* [8] proposed a hybrid approach that combined DL and traditional ML techniques for deepfake image detection and obtained 89.5% accuracy. Khalid *et al.* [30] proposed a graph neural network (GNN) framework that incorporates interpretability for deepfake detection. Similarly, attention-based models leveraging weight mechanisms and the LayerCAM technique were introduced in [31, 32]. Silva *et al.* [33] explored ensemble models combining standard CNNs with attention-based networks using Grad-CAM visualizations. Transformer-based solutions on Celeb-DF and FF++ datasets were presented in [34, 35]. Additionally, Ilyas *et al.* [36] introduced a prototype learning approach using ConvNext-PNet, which achieved notable generalization with 98.70% accuracy on FF++ and 97.09% on Celeb-DF. Similarly, Ahmad *et al.* [37] demonstrated the importance of privacy and security while mitigating fake content, a concern also highlighted by Huang *et al.* [38], listed specific challenges in explicit and implicit identity detection, such as face swapping, shifting, and face2face. Lin *et al.* [39] proposed domain-agnostic features to ensure the model's satisfactory performance in different domains. Malolan *et al.* [40] have attempted to bridge this gap by incorporating XAI techniques to highlight manipulated regions within images and offering visual cues about the model's decision-making process. Despite this progress, researchers lack a proper understanding of how attribution-based explanations affect model behavior for different types of forgeries. A notable gap remains in explaining the patterns observed across different deepfake detectors. However, identifying both deepfakes and PG images remains a significant challenge. To our knowledge, no study has explicitly addressed this nuanced distinction. In contrast, our focus is on the impact of feature attribution during classification.

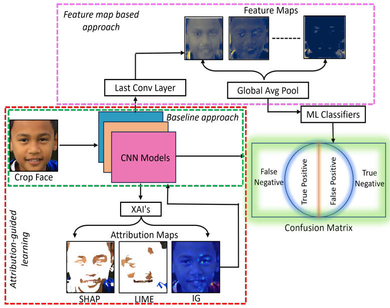


Figure 1: An end-to-end three-stage pipeline of the proposed approach, considering: baseline CNNs on raw images (green), attribution maps (red), and feature maps from the final convolutional layer (magenta).

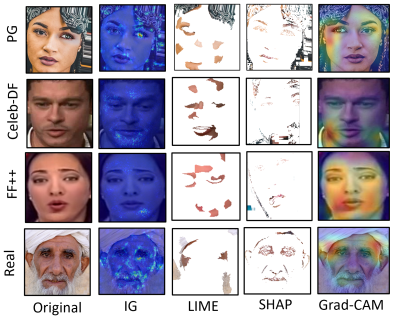


Figure 2: Feature attribution maps of PG, Celeb-DF, FF++, and real images. Each row shows the original image, IG, LIME, SHAP, and Grad-CAM visualization.

3 Methodology

3.1 Feature Attribution Methods

We incorporate feature attribution techniques into our workflow (shown in Figure 1) to enhance the model's interpretability by highlighting its most relevant regions. The XAI methods identify which parts of that image most influence the classifier's decision, enabling us to focus on semantically meaningful areas rather than the entire image. This allows us to comprehend whether the model looks at spurious features and correctly identifies the manipulated region. This section outlines the attribution tools employed and details the procedure for generating and interpreting these maps within our pipeline.

3.1.1 SHapley Additive exPlanations (SHAP)

We quantify the influence of individual pixels on model decisions using the SHAP explainer. SHAP is a widely used XAI method based on information theory that considers image features $\mathcal{S} = \{1, \ldots, d\}$ as the "players" of a cooperative game, assigning each feature $j \in \mathcal{S}$ a Shapley value ϕ, that indicates its marginal contribution to the model output [22]. For a given input image $\mathbf{X} \in R^{H \times W \times 3}$, SHAP yields $\phi \in R^{d}$ (where $d = H \cdot W \cdot 3$). The final saliency map for the three channels is represented by $\Phi \in R^{H \times W}$ and the positive entries mark forged regions (such as mouth corners, eye rims, and blending seams). We define a binary mask to retain the positively attributed pixels as, $M_{u,v} \;=\; 1\big[\text{where } \Phi_{u,v} > 0\big]$, where $u = 1, \ldots, H,\ v = 1, \ldots, W$. Then apply this mask channel-wise to the original image according to the formula $\mathbf{X}^{+} = \mathbf{X} \odot M_{u,v}$. The visual inspection of the resulting masked attribution map $\mathbf{X}^{+}$ in Figure 2 (Fourth column) confirms that the model focuses on plausible forgery cues. Fine-tuning with these masked attribution maps steers learning toward genuine manipulation artefacts. However, the extra computation is negligible compared to a standard forward pass and enables spontaneous attribution during training and evaluation.

3.1.2 Local Interpretable Model-agnostic Explanation (LIME)

We capture important local visual features using the LIME explainer. LIME is a popular feature attribution method [23] that explains a model's predictions by fitting a local, sparse surrogate around the input. For an image $\mathbf{X}$, we first obtain N super-pixels $\{\mathcal{S}_j\}_{j=1}^{N}$ using Simple Linear Iterative Clustering (SLIC) [41]. Each perturbation is a binary vector $\mathbf{z} \in \{0,1\}^{N}$ that masks super-pixels, sampling n such vectors yields a neighbourhood $\mathcal{Z} = \{\mathbf{z}^{(i)}\}_{i=1}^{n}$. The surrogate model is a sparse, linear model $g(\mathbf{z}) = \beta_0 + \sum_{j=1}^{N} \beta_j z_j$, that is fitted by weighted least squares with a kernel $\pi_{\mathbf{X}}$ favouring perturbations close to $\mathbf{X}$. The learned weights $\beta = \{\beta_j\}_{j=1}^{N}$ represent an attribution map $\Psi_{u,v} = \beta_{s(u,v)}$, where $s(u,v)$ represents the super-pixel index containing (u,v). Similar to our approach with SHAP, we define a binary mask $\mathbf{M}_{u,v} = 1\big[\text{where } \Psi_{u,v} > 0\big]$ to obtain the masked attribution map $\mathbf{X}^{+}$. The third column of Figure 2 shows that LIME highlights fine-grained regions rather than broad facial structures. Per prediction, its computational cost is higher because every sample $\mathcal{Z}$ requires a forward pass and the surrogate fit, which makes LIME slower than SHAP for deep networks.

3.1.3 Integrated Gradient (IG)

Integrated gradient [24] is a widely used XAI method that computes feature importance by performing a line integration of the model's gradient along a straight path from a baseline $\mathbf{X}_0$ to the input $\mathbf{X}$. Let $\mathbf{X}_{\alpha} \;=\; \mathbf{X}_0 + \alpha\ (\mathbf{X} - \mathbf{X}_0)$, where $\alpha \in [0,1]$. The attribution for pixel (u,v,c) approximates the integral with 50 Riemann steps as,

$$\mathrm{IG}_{u,v,c} \;=\; \big(X_{u,v,c} - X_{0,u,v,c}\big) \int_0^1 \frac{\partial f(\mathbf{X}_{\alpha})}{\partial X_{u,v,c}}\, d\alpha. \tag{1}$$

IG yields time complexity $O(K)$, much lower than deep SHAP and LIME. The IG method produces a heat map $\Gamma \in R^{H \times W}$ summing all channels, highlighting distinct and prominent facial regions such as eyes and nose, as depicted in the second column of Figure 2. These attribution maps improve transparency and can steer fine-tuning toward authentic manipulation artefacts. Similarly, Grad-CAM reveals that the shallow CNN focuses on manipulated areas.

3.1.4 Feature Map From Last Convolution Layer

The model processes the input image through multiple convolutional layers and progressively extracts different-level features. The feature maps from the final convolutional layer of each CNN model are passed through a Global Average Pooling (GAP) layer [42] to obtain compact feature vectors. These feature vectors are then used as input to various ML models (see Figure 1).

3.2 Experimental Setup

We considered three popular, distinct datasets: a PG image dataset [43], Celeb-DF [44], and FaceForensics++ (FF++) [45]. The PG dataset contains 2,041 face images, comprising 1,081 real and 960 fake images. The fake images in this dataset are categorized into easy, medium, and hard classes based on their ease of detection. The easy, medium, and hard classes contain noticeable geometric aberrations, subtle distortion, and minimal visual anomalies. Celeb-DF [44] consists of high-quality deepfake videos of celebrities, generated with advanced synthesis techniques that produce artefacts such as lip-sync mismatch and flickering. FF++ is a challenging benchmark dataset [45] containing 1,000 videos for manipulation methods, including DeepFakes, Face2Face, FaceSwap, and NeuralTextures. This study focuses on the DeepFakes subset of the FF++ dataset to evaluate model performance. These datasets offer a diverse and challenging benchmark for training and testing forgery detection models.

To evaluate the impact of attribution maps, we trained models with two types of inputs: (i) original images and (ii) attribution-based masked images. We performed both within-dataset and cross-dataset evaluations to assess the generalizability. For instance, we trained models on the PG dataset and tested them on Celeb-DF and FF++, and vice versa. We initially resized and normalized the images for preprocessing and then split the PG dataset into training and testing sets using a 70:30 ratio. From the Celeb-DF and FF++ datasets, we extracted 10 random frames per video, ensuring variations in expression, pose, and lighting. Subsequently, we split these video-based datasets into a train-test subset, according to the official protocols [44, 45].

We explored six convolution-based models: a custom CNN model with three 3×3 convolutional layers (32 / 64 / 128 filters); AlexNet as a shallow baseline; VGG-19 for deeper hierarchical features; ResNet-50 and ResNet-101 with residual connections; and EfficientNet-V2, a relatively lightweight model. We first trained the models on raw images for binary classification (real vs. fake) using 20 epochs, a batch size 32, and a learning rate of 0.01. After convergence, feature attributions are generated using SHAP, LIME, and IG. We then used these attribution maps to fine-tune the models and re-evaluate performance, systematically analyzing the efficacy of attribution-guided learning across multiple datasets and settings.

Training on PG dataset						
	Celeb-DF		FF++		PG	
Modality	Accuracy	ROC-AUC	Accuracy	ROC-AUC	Accuracy	ROC-AUC
Baseline						
CNN	50.58	0.69	50.09	0.5	71.97	0.72
Alexnet	52.58	0.51	50.98	0.5	70.26	0.71
VGG19	54.31	0.52	51.91	0.51	75.59	0.75
ResNet50	54.58	0.53	52.08	0.51	76.30	0.75
ResNet101	55.07	0.54	52.71	0.52	76.78	0.75
EfficientNetV2	**55.91**	**0.55**	**53.07**	**0.53**	**78.78**	**0.77**
Attribution Map From SHAP						
CNN	56.78	0.54	51.58	0.51	86.19	0.85
Alexnet	58.73	0.57	52.11	0.52	85.31	0.84
VGG19	59.66	0.56	53.46	0.53	85.93	0.84
ResNet50	61.03	0.59	56.49	0.55	88.00	0.87
ResNet101	61.74	0.58	57.41	0.56	89.07	0.87
EfficientNetV2	**62.77**	**0.62**	**60.15**	**0.6**	**90.78**	**0.86**
Attribution Map From LIME						
CNN	56.35	0.54	54.01	0.53	84.33	0.81
Alexnet	58.60	0.56	54.12	0.54	84.03	0.82
VGG19	59.64	0.58	56.40	0.56	85.44	0.84
ResNet50	60.87	0.59	58.93	0.58	86.95	0.83
ResNet101	61.83	0.60	59.68	0.59	85.59	0.84
EfficientNetV2	**62.11**	**0.61**	**62.95**	**0.61**	**86.00**	**0.85**
Attribution Map from Integrated Gradient						
CNN	60.18	0.57	56.28	0.56	**96.00**	**0.91**
Alexnet	61.73	0.59	60.91	0.6	88.32	0.87
VGG19	62.60	0.61	61.66	0.61	89.65	0.88
ResNet50	63.43	0.62	62.19	0.61	91.47	0.82
ResNet101	63.84	0.63	63.11	0.63	91.47	0.84
EfficientNetV2	**64.87**	**0.64**	**64.05**	**0.64**	91.81	0.89
Feature Map From Last Conv Layer						
CNN	61.02	0.6	60.21	0.6	82.03	0.80
Alexnet	61.06	0.6	60.39	0.6	78.86	0.77
VGG19	61.50	0.61	60.55	0.59	80.51	0.77
ResNet50	62.02	0.6	61.16	0.59	82.03	0.79
ResNet101	62.74	0.61	61.52	0.60	82.49	0.82
EfficientNetV2	**63.02**	**0.60**	**62.20**	**0.61**	**83.03**	**0.82**

Table 1: Comparative performance assessment of the models across different modalities. Models are trained on the PG dataset.

Training on Celeb-DF dataset						
	Celeb-DF		FF++		PG	
Modality	Accuracy	ROC-AUC	Accuracy	ROC-AUC	Accuracy	ROC-AUC
Baseline						
CNN	87.69	0.91	57.45	0.54	71.09	0.70
Alexnet	92.8	0.95	58.82	0.55	72	0.71
VGG19	93.1	0.93	59.14	0.57	72.28	0.71
ResNet50	88.34	0.87	60.34	0.6	73.91	0.72
ResNet101	94.02	0.94	62.46	0.62	74.37	0.73
EfficientNetV2	**97.63**	**0.96**	**64.83**	**0.64**	**77.48**	**0.76**
Attribution Map From SHAP						
CNN	88.13	0.88	61.62	0.61	71.61	0.70
Alexnet	91.29	0.90	61.69	0.6	72.23	0.71
VGG19	92.81	0.91	62.71	0.61	72.91	0.72
ResNet50	93.79	0.91	63.39	0.63	73.73	0.72
ResNet101	94.51	0.92	64.31	0.63	75.52	0.73
EfficientNetV2	**94.63**	**0.93**	**66.23**	**0.65**	**76.65**	**0.75**
Attribution Map From LIME						
CNN	87.82	0.85	60.27	0.59	65.46	0.63
Alexnet	91.06	0.89	60.65	0.6	66.81	0.63
VGG19	91.71	0.91	61.22	0.61	66.80	0.64
ResNet50	89.57	0.87	62.79	0.62	68.15	0.66
ResNet101	89.09	0.85	62.01	0.62	67.33	0.66
EfficientNetV2	**97.99**	**0.94**	**64.14**	**0.64**	**71.56**	**0.69**
Attribution Map from Integrated Gradient						
CNN	89.91	0.90	63.01	0.61	75.29	0.71
Alexnet	93.75	0.89	63.12	0.62	76.15	0.72
VGG19	94.17	0.92	63.54	0.63	76.42	0.72
ResNet50	92.57	0.89	64.86	0.64	77.2	0.74
ResNet101	95.07	0.95	65.22	0.65	78.28	0.74
EfficientNetV2	**98.06**	**0.98**	**69.38**	**0.69**	**78.83**	**0.77**
Feature Map From Last Conv Layer						
CNN	86.02	0.84	61.34	0.61	64.02	0.64
Alexnet	91.75	0.90	61.55	0.61	69.75	0.68
VGG19	93.05	0.91	61.71	0.61	70.51	0.69
ResNet50	87.83	0.85	62.78	0.62	71.81	0.71
ResNet101	93.82	0.90	63.93	0.63	72.87	0.70
EfficientNetV2	**97.34**	**0.95**	**64.56**	**0.64**	**75.51**	**0.75**

Table 2: Comparative performance assessment of models across different modalities. Models are trained on the Celeb-DF dataset.

4 Experimental Results and Discussions

4.1 Baseline Performance

We trained the six aforementioned CNN models on the original datasets to establish a benchmark for within and cross-dataset settings. The comparative results tabulated in Tables 1-3 confirm that all models obtain high accuracy and ROC-AUC. These Tables suggest that the EfficientNetV2 achieves comparatively higher accuracy and ROC-AUC across all settings. A detailed analysis reveals that the accuracy and ROC-AUC of the models increase steadily with an increase in the model depth. Table 1 reflects that the models struggle to accurately classify images from the PG dataset, as it contains subtle artefacts. The best performing model, EfficientNetV2, achieved 78.78% accuracy and 0.77 ROC-AUC. However, the performance of the EfficientNetV2 displayed noticeable degradation in the cross-dataset setting, as expected. Table 2 suggests that EfficientNetv2 achieved excellent performance on the Celeb-DF dataset, with an accuracy of 97.63 and an ROC-AUC of 0.96. Models trained on the Celeb-DF dataset obtained better recognition performance in cross-dataset evaluation than those trained on the PG dataset. The diverse features in these two datasets influence recognition performance and impede generalization in cross-dataset experiments. EfficientNetV2 achieved the highest performance on FF++ with 99.58% accuracy and 0.97 ROC-AUC, though cross-dataset scores dropped significantly for most models. The loss plots in Fig. 3 demonstrate that shallow CNN models converge slowly, especially for the FF+ dataset. In contrast, models with complex network architectures, such as ResNet101 and

EfficientNetV2, achieve lower loss and higher accuracy in fewer epochs. Models trained on other deepfake datasets attain relatively high accuracy on the PG dataset. Other cross-dataset experiments produce comparatively lower accuracy, indicating that the PG dataset contains realistically blended manipulations or duplicated facial regions. These plots underline that more profound and complex models inadvertently learns useful features.

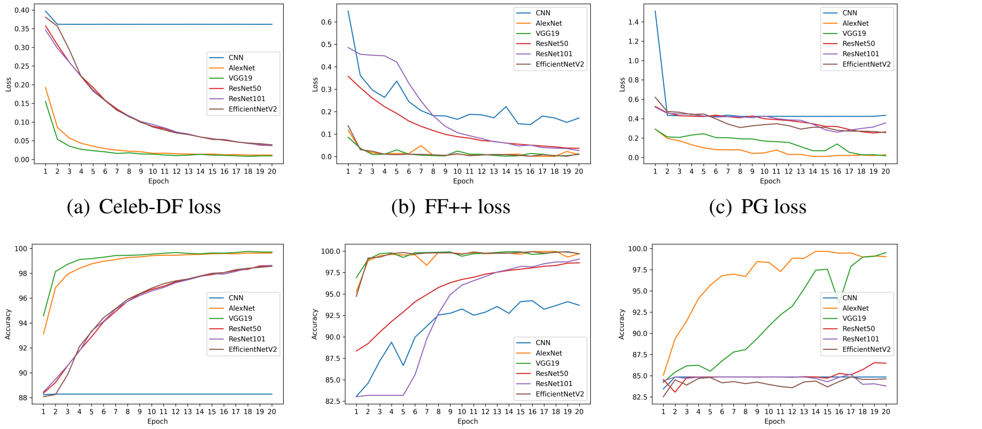
Training on FF++						
	Celeb-DF		FF++		PG	
Modality	Accuracy	ROC-AUC	Accuracy	ROC-AUC	Accuracy	ROC-AUC
Original Images						
CNN	58.46	0.52	90.03	0.81	71.38	0.70
Alexnet	60.65	0.59	89.22	0.81	71.92	0.71
VGG19	62.74	0.61	94.90	0.84	74.14	0.74
ResNet50	68.34	0.61	92.22	0.84	74.19	0.74
ResNet101	68.76	0.64	96.33	0.85	75.07	0.75
EfficientNetV2	**69.98**	**0.69**	**99.58**	**0.97**	**76.67**	**0.77**
Attribution Map From SHAP						
CNN	64.32	0.62	89.87	0.89	73.88	0.71
Alexnet	65.69	0.63	89.98	0.90	75.38	0.74
VGG19	65.71	0.63	91.95	0.91	78.68	0.77
ResNet50	70.31	0.65	94.56	0.94	81.99	0.81
ResNet101	71.32	0.7	95.52	0.93	82.13	0.82
EfficientNetV2	**72.83**	**0.71**	**97.09**	**0.95**	**84.58**	**0.84**
Attribution Map From LIME						
CNN	61.27	0.60	91.61	0.90	72.79	0.71
Alexnet	62.65	0.62	92.12	0.90	73.01	0.73
VGG19	63.22	0.63	98.40	0.94	75.77	0.74
ResNet50	67.79	0.65	97.93	0.94	79.99	0.79
ResNet101	67.01	0.65	95.68	0.92	78.34	0.77
EfficientNetV2	**70.14**	**0.69**	**99.95**	**0.96**	**81.33**	**0.81**
Attribution Map from Integrated Gradient						
CNN	67.15	0.66	94.86	0.95	75.35	0.75
Alexnet	70.62	0.69	89.22	0.95	79.19	0.79
VGG19	71.84	0.7	94.90	0.95	81.71	0.81
ResNet50	72.48	0.7	92.22	0.96	82.62	0.82
ResNet101	73.12	0.71	93.67	0.97	83.35	0.82
EfficientNetV2	**73.88**	**0.71**	**99.58**	**0.98**	**86.15**	**0.83**
Feature Map From Last Conv Layer						
CNN	64.34	0.6	89.02	0.87	73.02	0.72
Alexnet	65.55	0.63	91.73	0.90	75.73	0.74
VGG19	66.71	0.64	96.82	0.94	77.82	0.75
ResNet50	68.78	0.67	97.30	0.95	77.30	0.76
ResNet101	68.93	0.67	94.81	0.92	78.81	0.77
EfficientNetV2	**71.56**	**0.70**	**98.42**	**0.97**	**80.42**	**0.80**

Table 3: Comparative performance assessment of the models across different modalities. The models are trained on the FF++ dataset.

4.2 Model Performance with Insights from Feature Attribution

We masked the salient regions and fed these images to the aforementioned deep learning models to analyze the effectiveness of the saliency maps obtained from the XAI methods. According to the results shown in Table 1-3, using saliency maps improved the detection performance in most cases. SHAP-based saliency maps result in relatively higher detection accuracy for most datasets. Specifically, SHAP-based saliency maps led to relatively higher detection accuracy for most datasets. On average, SHAP and LIME improved accuracy by 5-6% and 3-4%, respectively. For the PG dataset, the accuracy and ROC-AUC score of EfficientNetV2 increased from 78.78% to 90.78% and from 0.77 to 0.86, respectively. However, cross-dataset evaluation generally reduced the classification accuracy for models trained on attribution maps. Models trained on the Celeb-DF dataset that utilized SHAP feature maps produced the best performance in a cross-dataset setting. For evaluations within the dataset, training in Celeb-DF and FF++ using SHAP maps yielded a precision of 94.63% and 97.09% with ROC-AUC scores of 0.93 and 0.95, respectively.

The results indicate that LIME's reliance on local features limits its effectiveness, while the global features obtained by SHAP help attain better recognition. Consequently, LIME-based training was less effective than training with SHAP maps. The models trained on the PG dataset showed weak generalization compared to deepfake-trained models, which adapted more easily to PG images. IG generally improved accuracy by more than 10%



(a) Celeb-DF loss (b) FF++ loss (c) PG loss

(d) Celeb-DF acc. (e) FF++ acc. (f) PG acc.

Figure 3: Training loss and accuracy curves of CNN models under baseline conditions across Celeb-DF, FF++, and PG datasets.

compared to the baseline. IG combines global and local features, identifying regions with aberrations and unnaturally shaped areas. While the custom CNN model achieved better performance on the PG dataset with 96% accuracy and a 0.91 ROC-AUC score, it performed poorly in cross-dataset evaluations. In contrast, models trained with the IG method achieved more than 70% accuracy on unseen datasets. The IG-based saliency map outperformed other methods because it considers the entire input path and captures more comprehensive feature interactions.

The real-world viability of attribution-enhanced systems hinges not only on their detection capabilities but also on their computational demands. Since attribution methods operate as post-hoc explanations, they do not alter the backbone network size but require additional forward and backward passes to generate saliency maps. Table 4 reports the resulting overhead in terms of floating-point operations per second (FLOPs) and inference time, measured on a 20GB NVIDIA ADA Generation 4000 GPU. For reference, baseline EfficientNetV2 inference is ∼12 ms per image. The results shown in Table 4 highlight a trade-off between interpretability and efficiency. SHAP and LIME provide helpful but slower explanations. In contrast, IG offers more detailed attributions with substantially lower latency, making it the most practical choice for real-time and large-scale deployment.

XAI Method	Extra FLOPs (× baseline)	Time / Image
SHAP	6-8×	80 ms (vs. 12 ms)
LIME	8-10×	100 ms (vs. 12 ms)
IG	∼5×	60 ms (vs. 12 ms)

Table 4: Computational overhead of attribution methods relative to baseline inference.

We also evaluated feature-based transfer learning using representations from the last convolutional layer of the CNNs. The output of the average-pooling layers was fed into 26 machine learning classifiers. The performance gains of these approaches were smaller than those of the deep models but remained above the baseline. ExtraTreesClassifier [46] achieved 83.03% accuracy and a 0.82 ROC-AUC score on the PG dataset. At the same time, Gaussian Naive Bayes (GaussianNB) and Nu-Support Vector Classification (NuSVC) obtained the best performance on Celeb-DF and FF++, with 63.02% and 62.02% accuracy, respectively. Cross-dataset generalization remained limited, although features from Celeb-DF showed better transferability than others. ExtraTreesClassifier achieved 97.34% accuracy and a 0.95 ROC-AUC on Celeb-DF, while Light Gradient Boosting Machine (LightGBM) achieved 64.56% on FF++ and 75.51% on PG. Training the models on FF++ features yielded the strongest results. Under this setting, ExtraTreesClassifier achieved 98.42% accuracy and a 0.97 ROC-AUC, and it maintained 71.56% accuracy on Celeb-DF and 80.42% on PG.

Saliency maps consistently improved detection, with IG providing the most significant gains. While SHAP offers strong global interpretability, LIME adds localized insights. However, PG manipulations remain the most difficult to detect and generalize compared to deepfake forgeries.

Model	FF++	Celeb-DF
DFGNN [30]	98.97	93.90
MRT-Net [31]	96.70	–
AW-MSA [32]	98.05	96.12
Ensemble [33]	–	93.64
ViXNet [34]	89.10	94.40
CviT [35]	93.00	–
ConvNext-PNet [36]	98.70	97.09
M2TR [47]	99.50	**99.76**
FakeFormer (AUC) [48]	97.76	95.21
GenConViT [49]	**99.60**	90.94
Ours (EfficientNetV2 + IG)	99.58	98.06

Table 5: Comparison of accuracy/AUC (%) for SOTA deepfake detection models on FF++ and Celeb-DF datasets.

Method	FF++	Celeb-DF	PG
Trained on FF++			
DFGNN [30]	–	73.40	–
ResNet-Swish-Dense54 [50]	–	70.04	–
ViXNet [34]	–	69.30	–
ConvNext-PNet [36]	–	68.45	–
M2TR [47]	–	68.2	–
Ours (EfficientNetV2 + IG)	–	**73.88**	**86.15**
Trained on Celeb-DF			
DFGNN [30]	69.60	–	–
ViXNet [34]	68.00	–	–
ConvNext-PNet [36]	41.28	–	–
Ours (EfficientNetV2 + IG)	**69.38**	–	**78.83**
Trained on PG			
Ours (EfficientNetV2 + IG)	**63.02**	**62.02**	–

Table 6: Cross-dataset accuracy (%) of SOTA fake face detection methods when trained on one dataset and tested on others.

4.3 Comparative Analysis with SOTA Deepfake Detectors

We benchmark our work against several SOTA methods [30, 31, 32, 33, 34, 35, 36, 47, 48, 49]. Most models consistently attained an accuracy of over 90% in within-dataset settings for all saliency maps. The EfficientNetv2 model with IG feature maps attained the best accuracy, 99.58% and 98.06% on FF++ and Celeb-DF datasets, respectively. Table 5 highlights that recent models [47, 49] achieve better performance in within-dataset settings. GenConViT [49] achieves the highest accuracy on the FF++ dataset, while M2TR [47] performs best on Celeb-DF. Although our approach does not surpass these methods in terms of accuracy, it addresses the critical gap. Our work analyzes cross-domain generalization and explores saliency maps to highlight the differences in the inherent quality of the images in the deepfake datasets. The work is significant because this area has received little attention despite its societal relevance. In-depth understanding of the attribution maps and generalization aspects will play a key role in developing explainable, trustworthy deepfake detection approaches.

We also evaluate our attribution-based methods on PG images for comparison. A hybrid model [28] was designed to detect fake faces on the PG dataset by leveraging the generative strength of GANs and the discriminative capabilities of ResNet, achieving an accuracy of 82.98%. As shown in Figure 4, the baseline classifiers in our standard training setup initially perform compared to this hybrid model presented by Safwat *et al.* [28]. However, when using attribution maps as input, the performance of these classifiers significantly improves and surpasses the hybrid model's accuracy. Notably, the IG-based attribution maps demonstrate superior generalization ability, leading to the highest improvement in detection accuracy

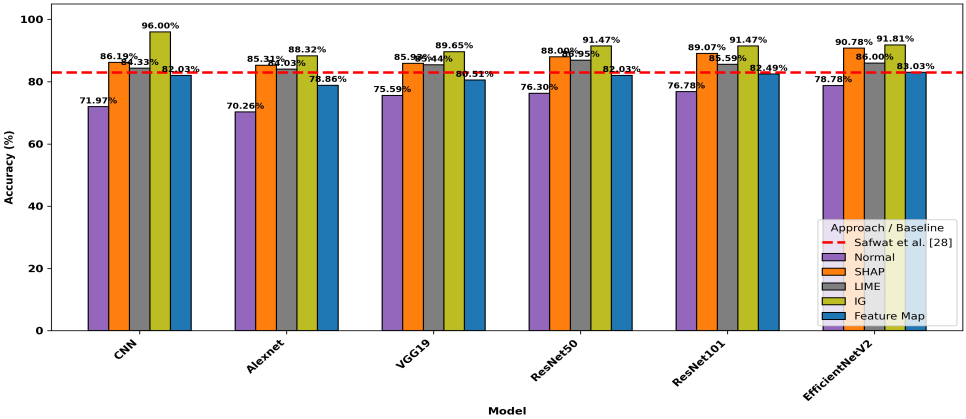


Figure 4: Performance comparison of different CNN models using different saliency mapping methods on the PG dataset. A red dashed line marks the benchmark accuracy (82.98%) [28].

across datasets.

Moreover, to assess the generalizability of our proposed approach, a cross-dataset evaluation is conducted, comparing it with existing methods, where models are trained on one dataset and tested on another. As the models listed in Table 6 are not evaluated on PG images, their results are not directly comparable. However, since these models are trained on the Celeb-DF dataset and tested on FF++, and vice versa, we include those settings for a fair comparison. As shown in Table 6, our proposed approach consistently achieves higher accuracy in both cross-dataset scenarios than the existing deepfake detection models.

Conclusion

In conclusion, this study underlines the importance of visual feature attribution in enhancing fake face detection. Incorporating attribution maps for training improves model interpretability and enables the models to learn informative and discriminative visual cues. Our cross-dataset evaluation further underscores the importance of generalization, suggesting that models trained with deepfake images, predominantly when guided by attribution maps, exhibit superior performance across diverse data distributions. The analysis highlights the practical advantage of using feature attribution as an explanation tool as an integral workflow component. The improvements in robustness and explainability pave the way for real-world deployment of the approach, particularly in high-stakes scenarios. Future work could expand this framework to detect deepfakes in multi-modal data, jointly utilizing visual and audio attribution. Our work emphasizes the necessity of explainable AI in building transparent, reliable, and high-performing deepfake detection systems.

Acknowledgement

A. Agarwal is partially supported through the ANRF PMECRG grant, Gov. of India.

References

[1] Rosa Gil, Jordi Virgili-Gomà, Juan-Miguel López-Gil, and Roberto García. Deepfakes: evolution and trends: R. gil et al. *Soft Computing*, 27(16):11295–11318, 2023.

[2] Aman Mehra, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Motion magnified 3-d residual-in-dense network for deepfake detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 5(1):39–52, 2023.

[3] Zhongjie Ba, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, and Kui Ren. Exposing the deception: Uncovering more forgery clues for deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 719–728, 2024.

[4] Akshay Agarwal and Nalini Ratha. Manipulating faces for identity theft via morphing and deepfake: Digital privacy. In *Handbook of statistics*, volume 48, pages 223–241. Elsevier, 2023.

[5] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE access*, 10:25494–25513, 2022.

[6] Akshay Agarwal and Nalini Ratha. Deepfake catcher: Can a simple fusion be effective and outperform complex DNNs? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3791–3801, 2024.

[7] Fakhar Abbas and Araz Taeihagh. Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, 252:124260, 2024.

[8] Rimsha Rafique, Rahma Gantassi, Rashid Amin, Jaroslav Frnda, Aida Mustapha, and Asma Hassan Alshehri. Deep fake detection and classification using error-level analysis and deep learning. *Scientific reports*, 13(1):7422, 2023.

[9] Md Shohel Rana, Beddhu Murali, and Andrew H Sung. Deepfake detection using machine learning algorithms. In *2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 458–463. IEEE, 2021.

[10] Vedant Jolly, Mayur Telrandhe, Aditya Kasat, Atharva Shitole, and Kiran Gawande. Cnn based deep learning model for deepfake detection. In *2022 2nd Asian conference on innovation in technology (ASIANCON)*, pages 1–5. IEEE, 2022.

[11] Manoj Kumar, Hitesh Kumar Sharma, et al. A gan-based model of deepfake detection in social media. *Procedia Computer Science*, 218:2153–2162, 2023.

[12] Akshay Agarwal and Nalini Ratha. Detection of identity swapping attacks in low-resolution image settings. *Journal of Information Security and Applications*, 89: 103911, 2025.

[13] Aayushi Agarwal, Akshay Agarwal, Sayan Sinha, Mayank Vatsa, and Richa Singh. Md-csdnetwork: Multi-domain cross stitched network for deepfake detection. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–8, 2021.

[14] Simranjeet Singh, Rajneesh Sharma, and Alan F Smeaton. Using gans to synthesise minimum training data for deepfake generation. *arXiv preprint arXiv:2011.05421*, 2020.

[15] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems*, 34:26422–26436, 2021.

[16] Balachandar Gowrisankar and Vrizlynn LL Thing. An adversarial attack approach for explainable ai evaluation on deepfake detection models. *Computers & Security*, 139: 103684, 2024.

[17] Konstantinos Tsigos, Evlampios Apostolidis, Spyridon Baxevanakis, Symeon Papadopoulos, and Vasileios Mezaris. Towards quantitative evaluation of explainable ai methods for deepfake detection. In *Proceedings of the 3rd ACM international workshop on multimedia AI against disinformation*, pages 37–45, 2024.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[22] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[24] Gary SW Goh, Sebastian Lapuschkin, Leander Weber, Wojciech Samek, and Alexander Binder. Understanding integrated gradients with smoothtaylor for deep neural network attribution. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4949–4956. IEEE, 2021.

[25] Lanzino Romeo, Fontana Federico, Diko Anxhelo, Marini Marco Raoul, and Cinque Luigi. Faster than lies: Real-time deepfake detection using binary neural networks. *arXiv preprint arXiv:2406.04932*, 2024.

[26] Richa Singh, Akshay Agarwal, Maneet Singh, Shruti Nagpal, and Mayank Vatsa. On the robustness of face recognition algorithms against attacks and bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13583–13589, 2020.

[27] Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Parameter agnostic stacked wavelet transformer for detecting singularities. *Information Fusion*, 95:415–425, 2023.

[28] Soha Safwat, Ayat Mahmoud, Ibrahim Eldesouky Fattoh, and Farid Ali. Hybrid deep learning model based on gan and resnet for detecting fake faces. *IEEE Access*, 12: 86391–86402, 2024.

[29] Gazi Hasin Ishrak, Zalish Mahmud, MD Farabe, Tahera Khanom Tinni, Tanzim Reza, and Mohammad Zavid Parvez. Explainable deepfake video detection using convolutional neural network and capsulenet. *arXiv preprint arXiv:2404.12841*, 2024.

[30] Fatima Khalid, Ali Javed, Hafsa Ilyas, Aun Irtaza, et al. Dfgnn: An interpretable and generalized graph neural network for deepfakes detection. *Expert Systems with Applications*, 222:119843, 2023.

[31] Ankit Yadav and Dinesh Kumar Vishwakarma. Aw-msa: Adaptively weighted multi-scale attentional features for deepfake detection. *Engineering Applications of Artificial Intelligence*, 127:107443, 2024.

[32] Ankit Yadav and Dinesh Kumar Vishwakarma. Mrt-net: Auto-adaptive weighting of manipulation residuals and texture clues for face manipulation detection. *Expert Systems with Applications*, 232:120898, 2023.

[33] Samuel Henrique Silva, Mazal Bethany, Alexis Megan Votto, Ian Henry Scarff, Nicole Beebe, and Peyman Najafirad. Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Science International: Synergy*, 4:100217, 2022.

[34] Shreyan Ganguly, Aditya Ganguly, Sk Mohiuddin, Samir Malakar, and Ram Sarkar. Vixnet: Vision transformer with xception network for deepfakes based video and image forgery detection. *Expert Systems with Applications*, 210:118423, 2022.

[35] Deressa Wodajo and Solomon Atnafu. Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126*, 2021.

[36] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. Convnext-pnet: An interpretable and explainable deep-learning model for deepfakes detection. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2024.

[37] Saadaldeen Rashid Ahmed, Emrullah Sonuç, Mohammed Rashid Ahmed, and Adil Deniz Duru. Analysis survey on deepfake detection and recognition with convolutional neural networks. In *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pages 1–7. IEEE, 2022.

[38] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4490–4499, June 2023.

[39] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16815–16825, June 2024.

[40] Badhrinarayan Malolan, Ankit Parekh, and Faruk Kazi. Explainable deep-fake detection using visual interpretability methods. In *2020 3rd International conference on Information and Computer Technologies (ICICT)*, pages 289–293. IEEE, 2020.

[41] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[42] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[43] https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection, accessed 5th Jan, 2025.

[44] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020.

[45] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[46] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[47] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 international conference on multimedia retrieval*, pages 615–623, 2022.

[48] Dat Nguyen, Marcella Astrid, Enjie Ghorbel, and Djamila Aouada. Fakeformer: Efficient vulnerability-driven transformers for generalisable deepfake detection. *arXiv preprint arXiv:2410.21964*, 2024.

[49] Deressa Wodajo Deressa, Hannes Mareen, Peter Lambert, Solomon Atnafu, Zahid Akhtar, and Glenn Van Wallendael. Genconvit: Deepfake video detection using generative convolutional vision transformer. *Applied Sciences*, 15(12):6622, 2025.

[50] Marriam Nawaz, Ali Javed, and Aun Irtaza. Resnet-swish-dense54: a deep learning approach for deepfakes detection. *The Visual Computer*, 39(12):6323–6344, 2023.