Research Square

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

A Facial Reenactment Inconsistency- Aware Interpretable Deepfake Image Identification Using DVPWT and SETCR2sC2N

Akshatha(Corresponding Author) G

akshathag.cse@bmsce.ac.in

BMS College of Engineering, Visvesvaraya Technological University

Kempanna M

Article

Keywords: Deepfake Image Identification (DII), Model Explainability (ME), Facial Reenactment Inconsistency (FRI), High-frequency Artifact Index (HAI), Face Detection (FD), Frequency-Domain Transform (FDT), and Manipulated Region Localization (MRL)

Posted Date: June 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9433893/v1

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

Additional Declarations: No competing interests reported.

A FACIAL REENACTMENT INCONSISTENCY-AWARE INTERPRETABLE DEEPFAKE IMAGE IDENTIFICATION USING DVPWT AND SETCR2Sc2N

Akshatha G and Dr. Kempanna M

Computer Science and Engineering, BMS College of Engineering, Bangalore, Visvesvaraya Technological University, Belagavi-590018, India Email: akshathag.cse@bmsce.ac.in

Artificial Intelligence and Machine Learning, Bangalore Institute of Technology, Bangalore, Visvesvaraya Technological University, Belagavi-590018, India Email: kempsindia@gmail.com

Abstract: In order to reduce identity misuse and digital fraud, Deepfake Image Identification (DII) is fundamental in the digital era, thus ensuring increased public trust. Nevertheless, the Facial Reenactment Inconsistency (FRI) that left minimal spatial traces was rarely addressed in the existing Deepfake (DF) identification frameworks, thereby overlooking the subtle distortions in facial parts. Therefore, the DII's significance and reliability are reduced. Hence, in this paper, an FRI-aware interpretable DII model is proposed using Discrete Vertex of the Parabola Wavelet Transform (DVPWT) and Stacked Ensemble Transfer Convolutional Root2sigmoid Cubic Neural Network (SETCR2sC2N). Initially, from publicly available sources, the input images are collected, followed by image pre-processing. Further, the face region is detected from the pre-processed images. After that, to perform frequency-domain transformation, the proposed DVPWT is established, followed by High-Frequency Artifact Index (HFAI) calculation, fine-grained visual pattern learning, and feature extraction. Subsequently, to classify the image into real and fake, the proposed SETCR2sC2N is utilized. In the meantime, the manipulated regions in the fake images are localized. A novel SHapley Additive Diophantine exPlanations (SHADP)-based decision interpretation is carried out to ensure the model's trust. Therefore, in DF detection, the proposed approach has better reliability with an accuracy of 99.3263%.

Keywords: *Deepfake Image Identification (DII), Model Explainability (ME), Facial Reenactment Inconsistency (FRI), High-frequency Artifact Index (HAI), Face Detection (FD), Frequency-Domain Transform (FDT), and Manipulated Region Localization (MRL).*

1. INTRODUCTION

Social media allows users to access multimedia content in the modern information-centric society (Patel et al., 2023a). The emergence of DFs has also increased with the rapid growth of generative models and diffusion-centric architectures (Bird & Lotfi, 2024) (Arshed et al., 2024). By changing facial expressions and lip movements, DFs generate highly realistic synthetic images and video content, making it increasingly challenging for humans and traditional forensic tools for differentiating betwixt authentic and forged content (Preeti et al., 2023). Due to the advancement of DF technology, the face-swap problem has been a major concern over the past 2 years (Shahzad et al., 2022). DF detection or GAN-generated face detection has become an

active research area to date (Akshatha & Kempanna, 2025). For manipulating facial features, DFs utilize generative models, thus creating modified expressions or identities with impressive realism (Waseem et al., 2023). In DF image generation, an unsupervised training approach named Generative Adversarial Network (GAN) has gained immense popularity (Waqas et al., 2022) (Heidari et al., 2024).

DFs have been used for misinformation campaigns, political manipulation, financial scams, and identity fraud (Raza et al., 2022). The presence of undetected DFs degrades the public trust and poor security in the context of sensitive domains, such as digital forensics, social media, biometric authentication systems, and healthcare (Malik et al., 2022). Therefore, owing to severe societal, political, and security implications of DF image detection, it has become critically important (Wolter et al., 2022). Yet, the traditional image forensic strategies, including metadata analysis and reverse image searchers, struggled to identify AI-generated content (Akshathag et al., 2025). Hence, to detect DF images, the existing studies established algorithms, such as Vision Transformer (VT), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Bidirectional Long Short-Term Memory (Bi-LSTM), thereby improving the security level (Patel et al., 2023) (Mansoor & Iliev, 2025). The traditional framework (Akshatha et al., 2026) used Transfer Learning (TL)-assisted Deep Learning (DL) algorithms, comprising DenseNet201, InceptionV3, and ConvNeXt, for optimizing the model's generalizability. Nevertheless, because of the black-box architecture, some traditional frameworks had limited trust.

The advanced explainability techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM), SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Layer-wise Relevance Propagation (LRP), are developed to ensure the model's trust and transparency (Momin et al., 2025). Forensic validation and legal admissibility are enabled by the explainable AI techniques through revealing why an image is categorized as genuine or manipulated (Yu et al., 2023). Yet, none of the prevailing studies concentrated on addressing the facial reenactment inconsistencies, thereby overlooking the subtle distortions in facial parts. Hence, this research is motivated by the need to design an FRI-aware interpretable DII using DVPWT and SETCR2sC2N.

1.1 Research overview

This section outlines the challenges faced by the existing studies during DF detection. In addition, the fundamental targets of the proposed work are also outlined.

Research gap: The majority of the DF detection studies concentrated on obvious manipulations, such as face swaps, identity replacement, and synthetic faces. However, to convey misinformation, manipulators considered facial reenactment inconsistencies. Yet, facial reenactment inconsistencies, which were subtle distortions in facial parts that left minimal spatial traces, were rarely addressed in the existing DF detection frameworks, increasing misclassification and reducing the model's reliability.

Problem definition: The conventional frameworks' potential limitations are given as follows,

- The existing DF detection model (Mathews et al., 2023) didn't accurately detect and focus on the facial region, thereby causing the model to process irrelevant background

information and overlook critical facial cues. Hence, DF detection performance was degraded.

- The traditional (Islam & Belhaouari, 2023) only detected whether the image was genuine or fake; however, it failed to point out the exact manipulated location in the image.

- The prior study (Tran et al., 2023) depended on black-box architectures that made decisions without providing interpretable reasoning. The inadequate transparency limited forensic reliability and decreased user trust.

- Most DF detection techniques were trained on specific datasets, encompassing known manipulation methods. Nevertheless, the existing frameworks struggled to adapt to the unseen or innovative generation techniques. Therefore, the system's overall reliability was reduced under real-world scenarios.

Significance and scope of the study: In the proposed framework, an FRI-aware interpretable DII is effectively designed using DVPWT and SETCR2Sc2N. The proposed work reduces the misclassification rate by capturing subtle distortions. Moreover, the model's decisions are interpreted by the proposed approach, thus ensuring reliability and trustworthiness.

Research aim and objectives: Capturing the subtle visual artifacts is the proposed approach's key aim, thereby accurately detecting the DF images. The proposed work's major objectives are rendered as follows,

- A novel DVPWT-based frequency domain transformation and HFAI calculation are done to improve the model's accuracy.

- The proposed work employs the Quadratic-Violo-Jones (QVJ) to accurately detect the exact facial region.

- An effective Segment Inverse Anything Model (SIvAM) is introduced to emphasize the manipulated regions in the image.

- The proposed work employs the SHADP, which highlights the highly influential features, to interpret the model's decision.

- TL is integrated with the proposed DII model to adapt to the different types of manipulation techniques.

This paper is structured as: The related works are exemplified in Section 2; the mathematical modeling of the proposed framework is exhibited in Section 3; the performance validation is demonstrated in Section 4; the paper is concluded with future enhancement strategies in Section 5.

2. LITERATURE SURVEY

(Khalifa et al., 2022) employed a DF recognition framework using diverse Gabor filters and a CNN. To detect the face from the images, a multi-task cascade CNN was employed. Here, to recognize the DF images, a Dual-scale Large Receptive Field Network (DSLRFN) was utilized. Owing to the inclusion of large-scale datasets, this approach had high generalizability.

Nevertheless, since the Gabor filters amplified the high-frequency facial components, this framework had maximum false positives in real-world conditions.

(Li et al., 2022) developed a DL-centric one-class detection model for fake-faces utilizing a GAN network. Here, to enhance the input image, Gaussian blur, Gaussian noise, and homomorphic filter enhancement strategies were applied. Thereafter, to categorize the image into fake and real, a multi-channel CNN was used. Therefore, as per the experimental outcomes, the suggested model had better superiority. Yet, it had less generalizability, which degraded the performance of the DF identification under real-world environments.

(Li et al., 2024) employed a probing explainable fake image detection framework using larger multimodal models. Here, to detect deep fake images, the models, including Generative Pre-trained Transformer (GPT), CNN Spot (CNNSpot), and Gram-Net, were introduced. Furthermore, the explainability mechanism helped to enhance the model's robustness. Nevertheless, owing to error propagation, this model had a risk of flawed descriptions and erroneous judgments.

(Kolagati et al., 2022) propounded a DF prediction regarding a deep multilayer perceptron and CNN model. Primarily, facial landmark detection was performed. Thereafter, the facial landmark detected data was subjected to the multi-layer perceptron that learned differences in real and DF frames. In the meantime, the CNN was utilized for extracting features. The DF images were effectively detected centered on the hybrid network model. This model had a higher detection rate. Nevertheless, owing to the limited computational resources, this framework was not suitable for large-scale applications.

(Islam & Belhaouari, 2023) presented wavelet-packets-based DF image analysis and identification. At first, to capture the image artifacts and crucial feature characteristics, the wavelet transform was applied to the images. Afterward, to identify the DF images, a CNN was established. Owing to the wavelet-based features, this framework had high robustness in DII. The reliability of this approach was degraded as a result of inadequate exact-manipulated location identification.

(Mathews et al., 2023) implemented an interpretable DF identification model based on an unconstrained dataset. Here, the input image was subjected to the MesoInception-based network, where the frame-level features were extracted and the manipulated content was automatically identified. Furthermore, to visualize the most influential region in the manipulated images, the Grad-CAM was employed, ensuring trustworthiness. However, this model didn't identify the exact facial regions, degrading the DF detection performance.

(Khormali & Yuan, 2024) presented a self-supervised graph transformer for effective DF identification. Firstly, to extract the fundamental spatial features, the VT architecture was employed. Next, to understand the matheulated regions, a graph convolution network with a transformer discriminator and a graph transformer relevancy map was employed. In DF identification, this approach had better efficacy and generalizability. Nevertheless, owing to the numerous graph nodes, this framework had high computational overhead.

(Coccomini et al., 2024) recommended an effective framework named multi-identity size-invariant video DF identification. Here, a spatio-temporal transformer with a CNN was utilized

for the identification of the DF frames while analyzing the spatio-temporal inconsistencies. This model achieved fine-grained predictions with interpretability. However, as the number of identities increased, this approach had considerable training complexity.

(Tran et al., 2023) analyzed a generalization of image forgery detection with a meta DF detection framework. Initially, to optimize the model's generalizability, the block shuffling transformation was applied to the images. Here, to perform DF detection, the meta-learning-assisted EfficientNet was utilized. This framework significantly improved the generalizability of the model and reduced the overfitting issues. However, this approach had a black-box nature, thus displaying poor model trust.

(Al-Dulaimi & Kurnaz, 2024) combined a hybrid CNN and LSTM algorithm for robust DF image classification concerning TL. Primarily, from publicly available sources, the input images were collected, followed by pre-processing and DF detection. The TL's inclusion helped to increase the model's adaptability. Nevertheless, owing to the lack of fine-grained feature learning, this approach had suboptimal performance.

3. PROPOSED METHODOLOGY FOR ENHANCED DEEP FAKE IMAGE IDENTIFICATION

By utilizing DVPWT and SETCR2Sc2N, an FRI-aware interpretable DII framework is designed in this section. The subtle distortions in the images are captured by the proposed DVPWT, thus improving the accuracy of the DII. Hence, Figure 1 depicts the proposed framework's conceptual design.

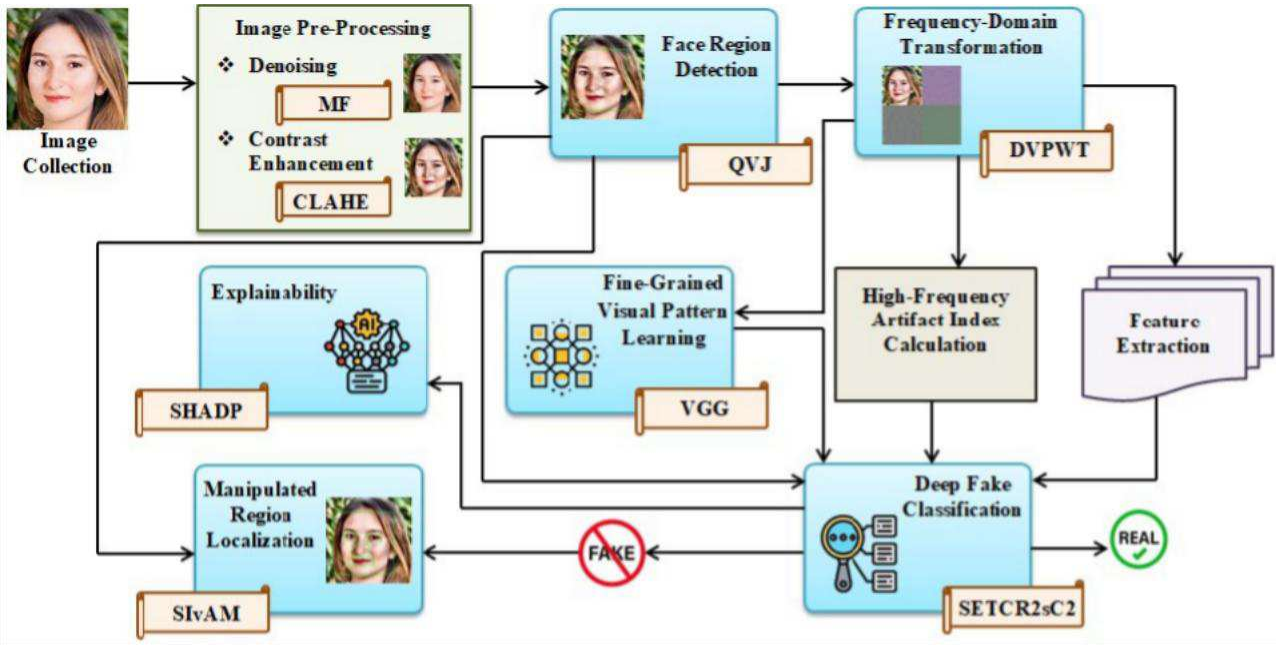


Figure 1: Architecture of the proposed methodology

In the proposed work, the key components, such as image pre-processing, face region detection, frequency-domain transformation, HFAI calculation, DF classification, manipulated region localization, and explainability, are encompassed.

3.1 Image collection

From publicly available sources, input images are collected in the image collection phase to train the proposed DII framework.

$$d_i = \{d_1, d_2, K\ K\ d_I\} \text{ Where, } i = 1 \text{ to } I \tag{1}$$

Here, the number of input images d_i is signified as I. Both authentic and manipulated samples are included in the collected input images.

3.2 Image pre-processing

After that, the raw input images d_i are subjected to image pre-processing, which enhances the image quality and characteristics.

Noise removal – The irrelevant noises in the d_i are eliminated in this step by applying the Median Filter (MF). By replacing every pixel with the median value of its neighboring pixels, the MF removes noise, thus suppressing isolated outliers.

$$h(a,b) = \max_{(q,p) \in v_{ab}} \{d_i(q,p)\} \tag{2}$$

Where, the input images with neighborhood pixels (q,p) are indicated as $d_i(q,p)$, the noise-removed images with neighborhood pixels (a,b) are represented as $h(a,b)$, and the window size is depicted as v.

Contrast enhancement – Afterward, to enhance the noise-removed images' resolution, the h is fed into the Contrast Limited Adaptive Histogram Equalization (CLAHE), thus leading to robust DF identification outcomes. To ensure local region enhancement, h is divided into small and non-overlapping regions called tiles. A histogram is computed for each tile to map the distribution of pixel intensity levels within that specific region. Next, the histogram is clipped at a predefined clip limit, thereby preventing excessive contrast enhancement. After that, histogram equalization is applied to each clipped tile. Thus, the input pixel intensities are mapped to an equalized range.

$$Hs_u = \sum_{k=1}^{u} \left(N_{his}{}^{u}\right)_k \tag{3}$$

Here, the cumulative histogram is depicted as Hs, the normalized histogram is indicated as N_{his}, the number of gray levels in the h is given as $u = 1 \text{ to } U$ $u = 1 \text{ to } U$, and the total number of pixel values in the h is denoted as $k = 1,2,K\ K$. Eventually, the bilinear interpolation is applied to integrate the region-enhanced tiles, thereby creating smooth transitions and eliminating visible boundaries between tiles. Here, the pre-processed images (contrast enhancement images) are demonstrated as (c).

3.3 Face region detection

The face region is detected from C using the proposed QVJ algorithm, thus resulting in accurate DII. The existing Viola-Jones (VJ) algorithm is extremely fast and also has a high detection rate. However, it is sensitive to the choice of window size, thereby struggling to detect very small faces or very large faces. Therefore, in the proposed approach, the Quadratic Equation (QE) is employed to initialize the proper window size by offering an optimal initial estimate concerning image dimensions.

At first, the input image C is transformed into a grayscale image $(g\Phi)$, reducing computational complexity.

$$g\Phi(p,q)=0.299r(p,q)+0.587g(p,q)+0.114b(p,q) \tag{4}$$

Where, the red, green, and blue channels are illustrated as (r,g,b), respectively, and the pixel coordinates are notated as (p,q). Afterward, to represent the cumulative sum of all pixels $(I(p,q))$, an integral image is computed, thereby processing the image at multiple scales.

$$I(p,q)=\sum_{h=0}^{p}\sum_{l=0}^{q} g\Phi(p,q) \tag{5}$$

Here, the random indices are specified as $(h,l)(h,l)$. Then, using the QE, a sliding window size $(\tilde{W})$ is computed. Next, to detect the face region of different sizes, the window slides across the image at multiple scales.

$$\tilde{W}(I)=\&^{2}I+cs \tag{6}$$

Where, the quadratic coefficient is depicted as $\&$, the linear coefficient is illustrated as I, and the constant term is demonstrated as cs. Haar-like features are extracted for each window by calculating the intensity difference between adjacent rectangular regions. In capturing the characteristics of facial patterns, such as eye-cheek contrast and nose bridge structure, the Haar-like features (H_D) play an essential role.

$$H_D(\tilde{W})=I(\sum A-\sum A) \tag{7}$$

Here, the pixels in the white area and black area are shown as A and A, respectively. Thereafter, to evaluate the extracted features, a cascade classifier is applied for each sub-window. Early stage classifier and later stage classifier are involved in the cascade classifier. To eliminate the non-face regions, early stages consider fewer and highly effective features. Later stages consider more complex and numerous features to accurately identify the face region. Hence, the windows that pass all cascade stages are labeled as face regions. Also, bounding boxes are generated for localizing the detected faces in the image. Next, the face-detected images $(I^{\cdot})$ are fed to the frequency-domain transformation.

3.4 Frequency-domain transformation

The proposed DVPWT effectively decomposes the $F^{'}$ into localized frequency sub-bands to capture the subtle distortions in the $F^{'}$, where subtle reenactment artifacts are determined as abnormal high-frequency energy distributions. The Discrete Wavelet Transform (DWT) captures fine details and offers effective spatial localization. Subsequently, the DWT is robust to noise and outliers, which ensures better performance. Nevertheless, considerable detail loss in the images is caused by improper wavelet coefficient selection. Hence, to select the appropriate wavelet coefficient, the proposed work employs the Vertex of the Parabola (VP). The most informative wavelet coefficients are automatically selected by VP while avoiding excessive detail loss.

The proposed DVPWT decomposes $F^{'}$ into sub-bands that represent different frequency components at multiple scales. The low- and high-pass filters are applied to $F^{''}$ to extract the sub-bands, including Low-Low (LL), Low-High (LH), High-Low (HL), and High-High (HH). The low-frequency components (smooth regions) are included in LL (approximation), whereas the texture and fine artifacts are included in the HH (diagonal details). LH and HL signify the vertical edges and horizontal edges, respectively.

Primarily, the low- and high-pass filters are applied row-wise, followed by downsampling. Therefore, the intermediate results (Row-low and Row-high) are acquired. The proposed work utilizes the VP $(\tilde{Z}$ to weight the wavelet coefficient.

$$\mathrm{Lw}(t)=\sum_{e} n[e]\times F^{''}[2t-e]\times \tilde{Z}\left(\frac{\tilde{a}}{2b},\frac{-\P}{4b}\right) \tag{8}$$

$$\mathrm{Hw}(t)=\sum_{e} o[e]\times F^{''}[2t-e]\times \tilde{Z}\left(\frac{\tilde{a}}{2b},\frac{-\P}{4b}\right) \tag{9}$$

Where, the row-low and row-high are denoted as Lw and Hw, respectively, the sample index is represented as e, the parabola coefficients are given as $(\tilde{a},\tilde{b}$, the discriminant of the parabola is specified as $\P$, the low- and high-pass filter coefficients are depicted as $n[e]$ and $o[e]$, respectively, and the index of the filter coefficients is depicted as t. Likewise, the low- and high-pass filters are applied to the intermediate two images column-wise. The downsampling is carried out after column-wise filtering. Thus, 4 sub-bands, encompassing LL (u_{LL}), HL (V_{LH}), HL (B_{HL}), and HH (C_{HH}), are acquired.

$$\mathcal{P}\{u_{\mathrm{LL}},V_{\mathrm{LH}},B_{\mathrm{HL}},C_{\mathrm{HH}}\} \tag{10}$$

Here, the decomposed images are indicated as $\mathcal{P}$. The pseudo-code of the proposed DVPWT is depicted below,

---

Input: Face detected images $F^{'}$
Output: Decomposed images $\mathcal{P}$

Begin

Initialize $F^{'}$, $\mathrm{Lw}(t)$, $\mathrm{Hw}(t)$ and $\tilde{Z}$

For 1 to each $F^{''}$ do,

#row-wise

Apply low-pass filter with downsampling

Integrate VP

$$\mathrm{Lw}(t)=\sum_{e} \mathrm{n}[e]\times F^{''}[2t-e]\times Z\left(\frac{\bar{q}}{2b},\frac{-\P}{4b}\right)$$

Perform high-pass filter with downsampling

$$\mathrm{Hw}(t)=\sum_{e} \mathrm{o}[e]\times F^{''}[2t-e]\times Z\left(\frac{\bar{q}}{2b},\frac{-\P}{4b}\right)$$

#column-wise

Apply low-pass filter and high-pass filter

Perform downsampling

Extract sub-bands

End For

Return $\left|\mu_{\mathrm{LL}}, V_{\mathrm{LH}}, B_{\mathrm{HL}}, C_{\mathrm{HH}}\right|$

End

3.5 High-frequency artifact index calculation

The HFAI is computed from to identify the most influential artifacts in the manipulated region, thus leading to accurate outcomes. The proportion of abnormal high-frequency components in the image is measured by HFAI $(\mathrm{k}_{\mathrm{index}})$.

$$\mathrm{k}_{\mathrm{index}}=\frac{\sum_{(\hat{x},\hat{y})\,\mathrm{Hf}}\left|(\hat{x},\hat{y})\right|}{\sum_{(\hat{x},\hat{y})\,\mathrm{Fq}}\left|(\hat{x},\hat{y})\right|} \tag{11}$$

Where, the frequency location is exhibited as $(\hat{x},\hat{y})$, and the set of high-frequency components and the set of all frequency components are notated as Hf and Fq, respectively.

3.6 Fine-grained visual pattern learning

In the meantime, using the Visual Geometry Group (VGG) algorithm, the fine-grained visual patterns are automatically learned from . To increase network depth, the VGG utilizes many stacked 3*3 convolutional layers. Here, early layers capture simple features (edges and textures), while deeper layers help to capture complicated shapes and objects. Primarily, the input layer transmits the into stacked 3*3 convolutional layers, thus generating a feature map.

$$F^{\mathrm{map}}=\mathrm{Conv}\times$$ (12)

Here, the convolution operation is denoted as Conv. Thereafter, using the max pooling operation $(\mathrm{Max}_{\mathrm{pool}})$, the feature map's spatial dimensions (Dim) are reduced.

$$\mathrm{Dim}\neg \text{reduce} F^{\mathrm{map}}\times\mathrm{Max}_{\mathrm{pool}}$$ (13)

Moreover, the flatten operation (Flt) is applied for projecting the feature map into a 1-dimensional space $(\cap \mathbb{C})$.

$$\mathrm{Flt} \frac{3}{4} \frac{3}{4} \frac{3}{4} \circledR \cap \mathbb{C} \tag{14}$$

Therefore, the fine-grained learned patterns are exemplified as (L_+).

3.7 Feature extraction

Additionally, from $\mathbb{C}$, essential features, such as texture, shape, edge, spatial, frequency, keypoint (Scale-Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF)), histogram (Histogram of Gradient (HOG) and color histogram), and structural features, are extracted to optimize the proposed DII framework's learning efficiency.

$$M_y = (M_1, M_2, \mathrm{K}\ \mathrm{K}\ M_Y) \text{ Here, } y = 1 \text{ to } Y \tag{15}$$

Where, the number of extracted features M_y is depicted as Y.

3.8 Deepfake classification

Here, the inputs (Ip), such as $F^{''}, L_+, \mathrm{k}_{\mathrm{index}}$, and $M_y \boxed{M_y}$, are fed into the proposed SETCR2sC2N, which classifies the image into real and fake. The proposed work uses the stacked ensemble techniques, including XceptionNet, EfficientNet, and CNN, to perform DII. Spatial feature extraction is performed by the existing XceptionNet and EfficientNet. Subtle visual artifacts and fine-grained textures in manipulated facial regions are captured by the XceptionNet. Owing to EfficientNet's optimized scaling methods, it effectively captures facial contours and structural anomalies. Here, the XceptionNet and EfficientNet are taken into account as base learners. The CNN is regarded as a meta-model. Relevant visual features are automatically learned and extracted by the CNN, thereby identifying subtle inconsistencies introduced during the DF generation process. Nevertheless, owing to the improper Activation Function (AF), the existing CNN has the risk of dead neurons. Therefore, in the proposed approach, the Root2sigmoid (R2s) AF is introduced, thus avoiding dead neurons by eliminating zero-gradient regions and preserving neuron responsiveness across the entire input domain. Similarly, owing to the improper weight initialization, the CNN has over-fitting or under-fitting issues. Thus, to initialize the weight value, the proposed work employs the Cubic Function (CF). The CF-centric weight initialization distributes initial weights with controlled variance, thus preventing excessively large or near-zero weights. The proposed work integrates TL to adapt to various types of manipulation techniques. Figure 2 presents the proposed SETCR2sC2N's architecture.

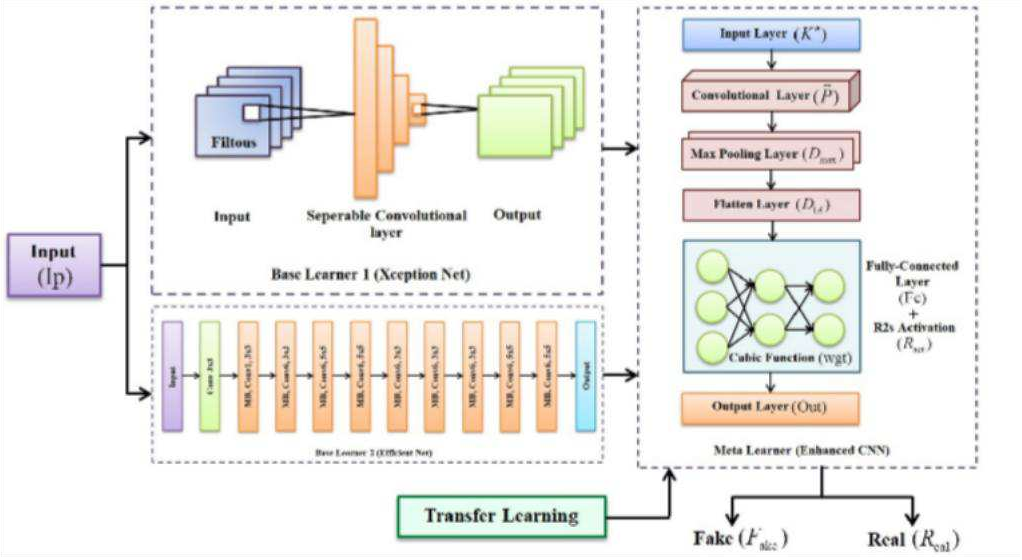


Figure 2: Design of the proposed SETCR2sC2N

Base learner 1: Firstly, Ip is fed to the base learner 1 (XceptionNet), which includes three main flows, namely entry flow (initial feature extraction), middle flow (deeper feature extraction with repeated modules), and exit flow (final feature aggregation and classification).

Entry flow – To extract the local features (L_{fea}), two 3*3 convolutions with Rectified Linear Unit (ReLU) activation as well as batch normalization are applied to Ip in the first step. Here, the ReLU activation is applied for introducing non-linearity. Thereafter, the depthwise separable convolution is applied to learn spatial patterns (edges, textures, and local artifacts) within each channel separately. Afterward, a 1*1 convolution layer is utilized for fusing information across channels, thus learning inter-channel dependencies.

$$O^{map} = (L_{fea} * D_{ker}) \tilde{A} \circ \quad O^{map} = (L_{fea} * D_{ker}) \tilde{A} \circ \tag{16}$$

Here, the feature map is exhibited as O^{map}, the pointwise (1*1) convolution kernel is indicated as $\tilde{A}$, and the depthwise convolution kernel is denoted as D_{ker}. After that, for reducing the feature map size, the max pooling layer is utilized. Next, the dimensionality-reduced feature map O^{map} is transmitted to the middle flow.

Middle flow – High-level features with repeated depthwise separable convolution blocks are extracted by the middle flow. ReLU activation, Depthwise convolution, and pointwise convolution are encompassed in each block. Furthermore, for maintaining gradient flow and preventing vanishing gradients, skip connections are enabled.

Exit flow – Exit flow aggregates features and further reduces them into a single vector. Here, to extract the global features, depthwise separable convolution blocks with residual connections are utilized. Thereafter, the global average pooling is applied to reduce each channel to a single value. Lastly, to obtain the output class probabilities, the Fully-Connected Layer (FCL) with softmax is utilized.

Hence, the prediction 1 results are represented as $(\hat{W}$.

Base learner 2: Here, Ip is subjected to the EfficientNet that conducts DII. At first, to automatically generate the feature map $(Y\text{¢})$, the convolutional layer is utilized. Next, a set of Mobile inverted Bottleneck Convolutional (MBConv) layers with different configurations is used. To capture fundamental spatial details, the MBConv utilizes depth-wise separable convolution $(\mathrm{f}_{\mathrm{depth}}$ and 1*1 point-wise convolution $(\tilde{O}$. Next, the 1*1 convolution is utilized for increasing the number of channels, thus attaining a relevant feature set (H_{fea}).

$$H_{\mathrm{fea}} = \mathrm{f}_{\mathrm{depth}} * \tilde{O}(Y\text{¢}) \tag{17}$$

Thereafter, the global average pooling operation (gb) is applied for diminishing the convoluted features' spatial dimensions.

$$\mathrm{Pl} = \mathrm{gb}(H_{\mathrm{fea}}) \quad \mathrm{Pl} = \mathrm{gb}(H_{\mathrm{fea}}) \tag{18}$$

Here, the spatial dimension reduced feature map is displayed as Pl. Then, the FCL $(\tilde{Q}$ significantly captures the probability value for each class.

$$\tilde{Q} = a_{\mathrm{fun}} \times (\mathrm{Pl}' \ \mathrm{w}_{\mathrm{gt}}) + \mathrm{B}_{\mathrm{is}} \tag{19}$$

Where, the sigmoid AF is notated as a_{fun}, the weight value of the FCL is given as w_{gt}, and the bias value of the FCL is represented as B_{is}.

$$U_{\mathrm{out}} \ \text{¾¾¾®} \ \tilde{M} \tag{20}$$

At last, the prediction 2 results $(\tilde{M}$ are produced by the output layer (U_{out}).

Meta learner: In this phase, the inputs $(K\text{¢})$ like $\hat{W}$ and $\tilde{M}$ are subjected to the enhanced CNN algorithm.

Convolution layer – At first, the inputs $(E\text{¢})$, such as $\hat{W}$ and $\tilde{M}$, are transmitted to the convolutional layer that learns the input's spatial patterns.

$$\overset{\mathrm{t}}{P} = \frac{K\text{¢} - \mathrm{Fl} + 2\,\mathrm{Pd}}{\mathrm{St}} + 1 \tag{21}$$

Here, the generated feature set is given as $\tilde{P}$, the convolutional kernel is indicated as Fl, the padding is demonstrated as Pd, and the stride is given as St.

Max pooling layer - Afterward, by utilizing the max pooling layer, which slides a window over the input feature map, the spatial dimensions of the $\tilde{P}$ is diminished. From each window, the max pooling layer selects the maximum value to reduce the spatial dimensions.

$$D_{\max} = \mathrm{Nu} \times \tilde{P} \tag{22}$$

Here, the dimensionality-reduced feature map is depicted as $D_{\max}$, and the max pooling operation is specified as Nu.

Flatten layer – Here, for better computational efficiency, the $D_{\max}$ is flattened into a single-vector value (D_{1d}).

$$D_{1d} = D_{\max} \tag{23}$$

Fully-connected layer – In this step, D_{1d} is processed by the FCL (Fc) through sharing its weight and bias values. The R2s AF (R_{act}) is utilized to improve the neuron's learning efficiency by introducing non-linearity.

$$\mathrm{Fc} = R_{\mathrm{act}}\left(D_{1d} \cdot \mathrm{wgt}\right) + b_{\mathrm{ias}} \tag{24}$$

$$R_{\mathrm{act}}\left(D_{1d}\right) = \frac{\sqrt{2^{R_{\mathrm{act}}}} - \sqrt{2}^{-R_{\mathrm{act}}}}{2\sqrt{2}\sqrt{2\left(\sqrt{2}^{2R_{\mathrm{act}}} + \sqrt{2}^{-2R_{\mathrm{act}}}\right)}} \tag{25}$$

Where, the weight and bias of the FCL are exhibited as wgt and b_{ias}, respectively. By using the CF, the wgt is regularized in the proposed approach.

$$\mathrm{wgt} = v_1 D_{1d}^{\ 3} + v_2 D_{1d}^{\ 2} + v_3 D_{1d} + v_4 \tag{26}$$

Here, the cubic parameters are depicted as (v_1, v_2, v_3, v_4).

Output layer – The output layer (Out) classifies the images into real (R_{cal}) and fake (F_{ake}), centered on the maximum class probability value.

$$\mathrm{Out} = \left(R_{\mathrm{cal}}, F_{\mathrm{ake}}\right) \tag{27}$$

Furthermore, in the proposed work, the concept of TL, which reuses the learned knowledge from another source model for a related task, is adapted. The pseudo-code of the proposed SETCR2sC2N is depicted as,

Input: $F^{\cdot}$, L_{+}, $\mathsf{k}_{\text{index}}$ and M_{v}

Output: Real or fake

Begin

Initialize lp, L_{fea}, O^{map}, H_{fea}, $E\text{¢}$ and Out

For 1 to each input do,

#Base learner 1 (XceptionNet)

Implement entry flow

$O^{\text{map}} = \left(L_{\text{fea}} * D_{\text{ker}}\right)\text{Ã}$

Execute middle flow and exit flow

Produce prediction 1 results $\hat{W}$

#Base learner 2 (EfficientNet)

Generate feature map $Y\text{¢}$

Apply MBConv layers

Extract feature map $H_{\text{fea}} = \mathsf{f}_{\text{depth}} * \tilde{O}\left(Y\text{¢}\right)$

Implement global pooling layer

$\mathrm{Pl} = \mathrm{gb}\left(H_{\text{fea}}\right)$

Evaluate fully-connected layer $\hat{Q}$

Obtain prediction 2 results $\hat{M}$

#Meta learner (Enhanced CNN)

Generate convolutional layer

$$\overset{\mathsf{t}}{P} = \frac{K\text{¢} - \mathrm{Fl} + 2\,\mathrm{Pd}}{\mathrm{St}} + 1$$

Apply max pooling operation

$D_{\max}$ ¾ ¾ ® Nu × $\hat{P}$

Obtain single vector D_{1d}

Perform fully-connected layer

$$\mathrm{Fc} = R_{\text{act}}\left(D_{1d}{}' \ \mathrm{wgt}\right) + b_{\text{ias}}$$

Activate R2s function R_{act}

Regularize wgt using CF

Evaluate output layer Out

End For

Return $\left(R_{\text{cal}}, F_{\text{ake}}\right)$

End

Overall, the DF faces are significantly predicted by the proposed SETCR2sC2N, thereby reducing the identity threat.

3.9 Manipulated region localization

The manipulated regions are localized from the fake images F_{ake} using the proposed SIvAM. By learning complex features and patterns from the image, the Segment Anything Model (SAM) precisely localizes manipulated regions within images. SAM is well appropriate for capturing the

altered area with subtle manipulation. It renders accuracy and zero-shot region localization without task-specific training, thus facilitating efficient focus on the manipulated area. Nevertheless, owing to the fixed weight values, the existing SAM fails to detect fine-grained changes, resulting in suboptimal performance. Thus, the Inverse Function (IvF) is utilized for selecting the suitable weight value. By computing a transformed value of the original fixed weight, the IvF selects suitable weight values. In addition, IvF dynamically adjusts the weights, improving segmentation efficiency.

- The F_{ake} is subjected to image encoding in the first step, where the image embedding (I_{emb}) is obtained, and the embedding signifies the visual features. The image is divided into fixed-size patches during embedding. Every single patch is embedded into a higher-dimensional feature vector. Thereafter, the prompt encoder is utilized for converting the prompt (P_{rompt}) into a prompt embedding (P_{emb}). Here, prompts are utilized to guide the model for region identification.

$$\left(F_{ake}, P_{rompt}\right) ¾¾® \left(I_{emb}, P_{emb}\right) \tag{28}$$

- Thereafter, the image embedding and prompt embedding are fused and further subjected to a lightweight decoder, thereby predicting the segmentation mask matching the prompt.

$$F_{use} = I_{emb} + P_{emb} \tag{29}$$

Where, the fused embedding is signified as F_{use}.

- After that, the decoder generates a segmentation mask (M_{ask}) with confidence scores for the prompted objects. The mask with the highest confidence score is selected. Using IvF, weight values (L^{weight}) are adjusted in the proposed work.

$$M_{ask} = sg\left(L^{weight} \times att\left(F_{use} \times W_{matrix}\right)\right) \tag{30}$$

$$L^{weight} = \sin^{-1}\left(F_{use}\right) \tag{31}$$

$$sg = \frac{1}{1+\exp\left(-F_{use}\right)} \tag{32}$$

Where, the sigmoid AF is specified as sg, the learnable mask projection weight is given as L^{weight}, the attention mechanism is depicted as att, and the learnable attention weight matrix is demonstrated as W_{matrix}. Therefore, to obtain an accurate localization, the refinement is done iteratively.

3.10 Explainability

Furthermore, a novel SHADP is utilized for interpreting the prediction results Out regarding Ip, enhancing the model's trust. In the proposed work, the SHAP is introduced to provide detailed explanations regarding the model's predictions. By emphasizing the facial regions or features

that most influence DF predictions, the SHAP offers clear and model-agnostic explanations. Trust and forensic reliability are improved by SHAP. Nonetheless, the influential features are underestimated by the improper marginal contribution calculations, thus resulting in incorrect interpretation. Hence, to compute the marginal contribution, the Diophantine Equation (DE) is employed in the proposed approach. The DE enables integer-based contribution balancing. Therefore, it is ensured that each feature's contribution fulfills the global additivity constraint without approximation errors.

Here, each feature is regarded as a "player" in a game that has the common goal (DF image prediction). The marginal contribution $(M_{\bar{A}})$ is calculated for each feature to measure how much the model's prediction changes regarding the feature.

$$M_{\bar{A}} = \mathrm{ab}(\mathrm{Ip}) + \mathrm{bc}(\mathrm{Out}) \tag{33}$$

Where, the Diophantine coefficients are denoted as ab and bc. Thereafter, to identify the feature's fair share of the prediction change, the average value of the contributions (Shapley value) is computed.

$$S_{\mathrm{ley}}(M_{\bar{A}}) = \sum_{\mathrm{xtf}\ U_{\bar{Y}} \setminus \{L\}} \frac{|\mathrm{Xt}|!(|U_{\bar{Y}}| - |\mathrm{Xt}| - 1)}{|U_{\bar{Y}}|!} \left[P_{\nu}(\mathrm{Xt} \dot{\mathrm{E}} \{\blacksquare\}) - P_{\nu}(\mathrm{Xt}) \right] \tag{34}$$

Here, the Shapley value is represented as S_{ley}, the subset of features without feature $\blacksquare$ is given as Xt, the prediction function is depicted as P_{ν}, and the total number of features is signified as $U_{\bar{Y}}$. Based on the Shapley values, the local model (L^{model}) effectively visualizes the feature's significance.

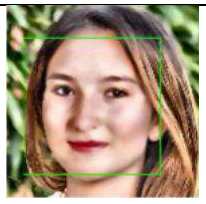
$$L^{\mathrm{model}}(S_{\mathrm{ley}}) = \blacksquare \blacksquare \blacksquare

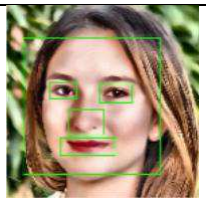
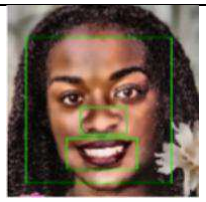
Table 1: Dataset characteristics

Real	Fake	Total
1081	960	2041

The sample distribution of the RFFD dataset is demonstrated in Table 1. Totally, the RFFD dataset contains 2041 images. From the whole images, 80% and 20% of the images are utilized to perform training and testing, respectively.

Table 2: Sample image results

Sample / Steps	Real	Fake (hard)	Fake (mid)	Fake (easy)
Input				
Noise removal				
Contrast enhancement				
Facial region detection				
Frequency-domain transformation				

Manipulated region localization				

In Table 2, the proposed work's sample image results regarding input, noise removal, contrast enhancement, facial region detection, frequency-domain transformation, and manipulated region localization are showcased.

4.2 Empirical analysis for frequency-domain transformation

To exhibit the proposed model's superiority in frequency-domain transformation, the proposed DVPWT's performance is analyzed by comparing it with several traditional methods, such as DWT, Continuous Wavelet Transform (CWT), Stationary Wavelet Transform (SWT), and Multi-Resolution Analysis (MRA).

Table 3: Performance analysis of the proposed DVPWT and existing methods

Methods	MSE	MAE	PSNR (dB)	SSIM
Proposed DVPWT	0.13	1.32	55.78	0.9467
DWT	0.23	1.89	42.45	0.9056
CWT	0.47	2.22	36.98	0.8677
SWT	0.56	4.23	28.89	0.7856
MRA	0.78	6.34	22.67	0.6789

In Table 3, the proposed DVPWT's performance is compared with conventional algorithms regarding frequency-domain transformation. VP-based wavelet coefficient selection was done to improve the proposed DVPWT's efficiency. Here, the proposed DVPWT attained Mean Squared Error (MSE) of 0.13, Mean Absolute Error (MAE) of 1.32, Peak Signal-to-Noise Ratio (PSNR) of 55.78dB, and Structural Similarity Index Measure (SSIM) of 0.9467. The existing SWT attained MSE of 0.56, MAE of 4.23, PSNR of 28.89dB, and SSIM of 0.7856 due to the improper coefficient selection. The proposed DVPWT attained impressive outcomes in frequency-domain transformation by appropriately selecting the wavelet coefficients.

4.3 Numerical investigation for deepfake identification

Similarly, to illustrate the proposed framework's significance in DII, the proposed SETCR2sC2N's performance assessment is carried out. Here, the quantitative metrics like accuracy, precision, recall, f-measure, and training time are taken into account.

Table 4: Efficacy analysis of the proposed SETCR2sC2N

Methods	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)	Training time (ms)
Proposed SETCR2sC2N	99.3273	99.3023	98.3923	99.3053	54363
MobileNet	98.5933	97.3433	96.3343	97.9676	76789
ShuffleNet	95.2423	93.4892	94.5893	94.5647	89234
PVT	92.5893	90.4892	92.1922	92.6757	98688
DBN	90.6352	88.2322	90.7856	90.6444	113532

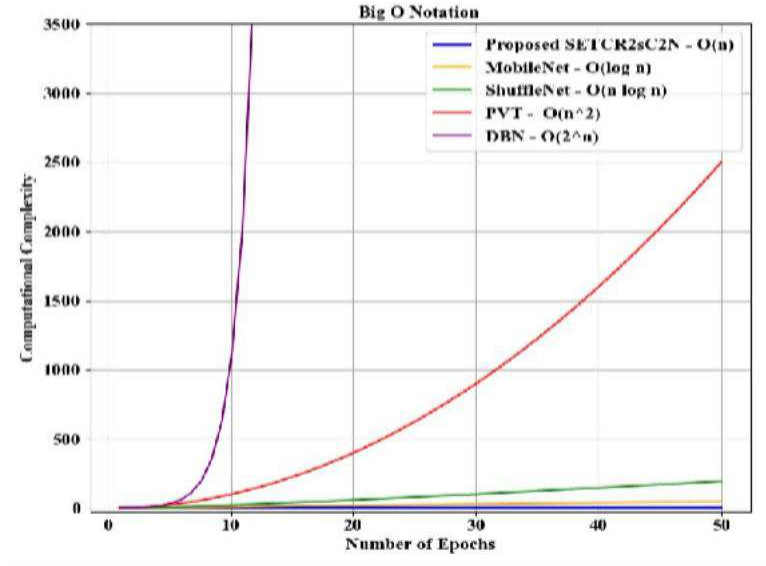


Figure 3: Complexity evaluation

Table 4 and Figure 3 evaluate the proposed SETCR2sC2N's performance by comparing it with several conventional classifiers, such as MobileNet, ShuffleNet, Pyramid VT (PVT), and Deep Belief neural Network (DBN). The proposed SETCR2sC2N obtained an accuracy of 99.3273%, precision of 99.3023%, recall of 98.3923%, f-measure of 99.3053%, and training time of 54363ms. The traditional techniques attained an average accuracy of 94.2650%, precision of 92.3884%, recall of 93.4753%, f-measure of 93.9631%, and training time of 94560ms owing to the lack of effective neuron activation. Due to the inclusion of the R2sAF, the proposed SETCR2sC2N's performance was enhanced. For varying epochs, the proposed SETCR2sC2N obtained stable complexity, while the conventional classifiers had a sudden spike in complexity. The proposed approach minimized the computational complexity due to the presence of CF-based weight initialization. Hence, according to the experimental results, the proposed framework had higher superiority in DII.

4.4 Experimental evaluation for explainability

Likewise, to showcase the proposed method's trustworthiness, the performance analysis of the proposed SHADP and traditional techniques like SHAP, Partial Dependence Plot (PDP), LIME, and Global Surrogate (GS) is done.

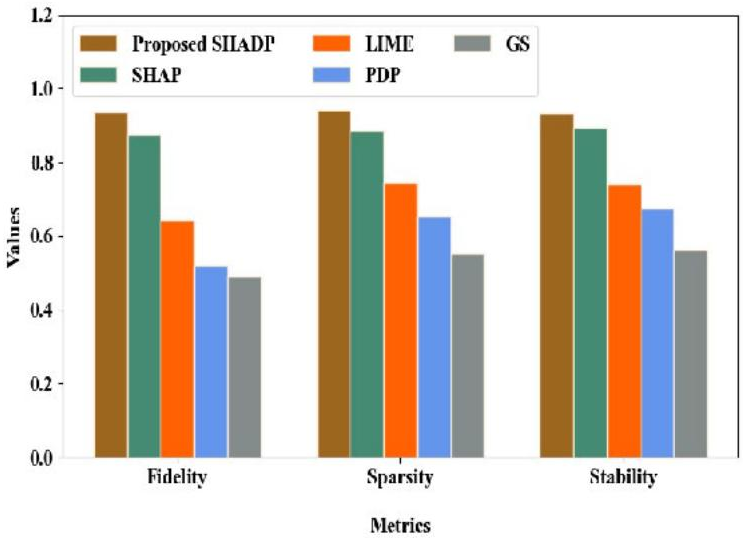


Figure 4: Empirical analysis of the proposed SHADP and existing methods

Figure 4 analyzes the performance of the proposed SHADP and conventional approaches concerning fidelity, sparsity, and stability. The proposed SHADP obtained 0.935 fidelity, 0.942 sparsity, and 0.932 stability. Likewise, the conventional SHAP had fidelity, sparsity, and stability of 0.875, 0.884, and 0.894, respectively. The presence of DE helped to enhance the model's interpretation efficiency by computing the suitable marginal contribution. Nevertheless, the performance of the existing algorithms was degraded due to the lack of proper marginal contribution calculation, evidencing the robustness of the proposed approach.

4.5 Performance validation for face region detection

The proposed QVJ's performance is contrasted with the existing techniques, such as VJ, You Only Look Once (YOLO), Multi-task Cascaded CNN (MCCNN), and Dual Shot Face Detector (DSFD), to illustrate the proposed system's efficiency in face region detection.

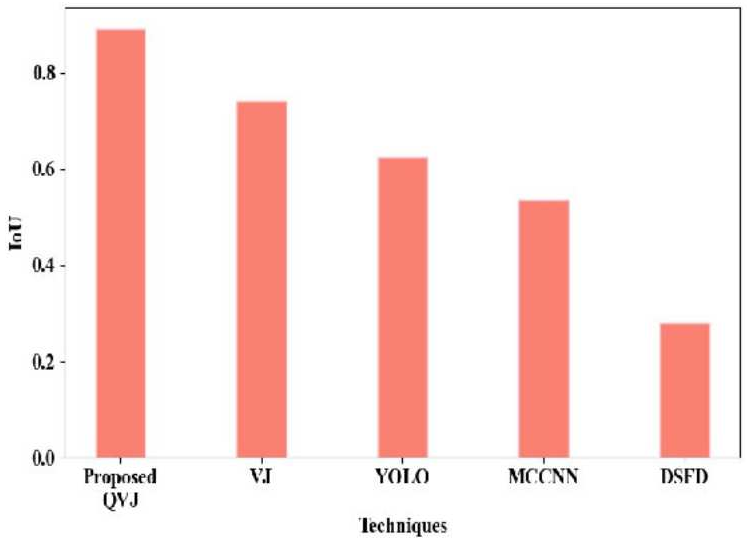


Figure 5: Intersection over Union validation

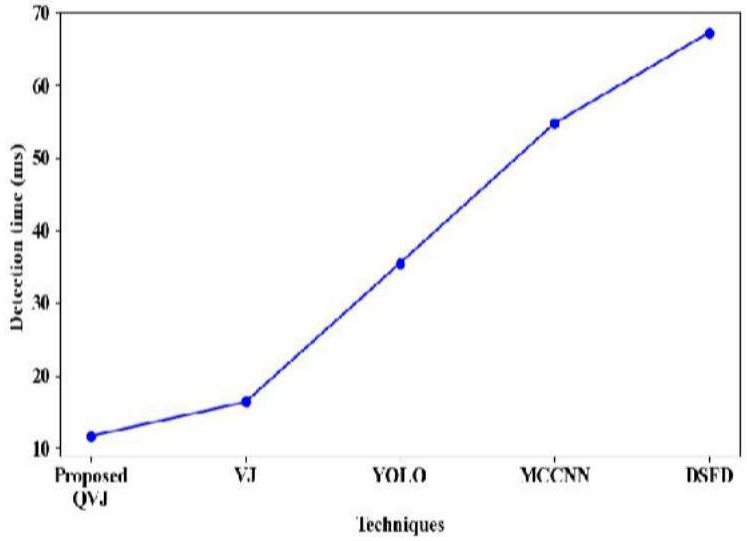


Figure 6: Performance assessment of the proposed QVJ

Figures 5 and 6 assess the Intersection over Union (IoU) and detection time of the proposed QVJ and conventional algorithms, respectively. The proposed QVJ obtained an IoU of 0.8921 and a detection time of 11.66ms. However, IoU of 0.7422, 0.6233, 0.5343, and 0.2784 were attained by the conventional VJ, YOLO, MCCNN, and DSFD, respectively. Furthermore, an average detection time of 43.45ms was taken by the traditional approaches. Here, owing to the inclusion of QE-based proper window size selection, the proposed approach's performance was improved. In contrast, the existing algorithms' performance was mainly affected due to improper window size selection. Therefore, as per the analysis results, the proposed framework had higher robustness when contrasted with the conventional techniques.

4.6 Efficiency assessment for manipulated region localization

Based on the manipulated region localization, the performance validation of the proposed SIvAM and traditional algorithms is carried out concerning the metrics, namely jaccard index and dice coefficient.

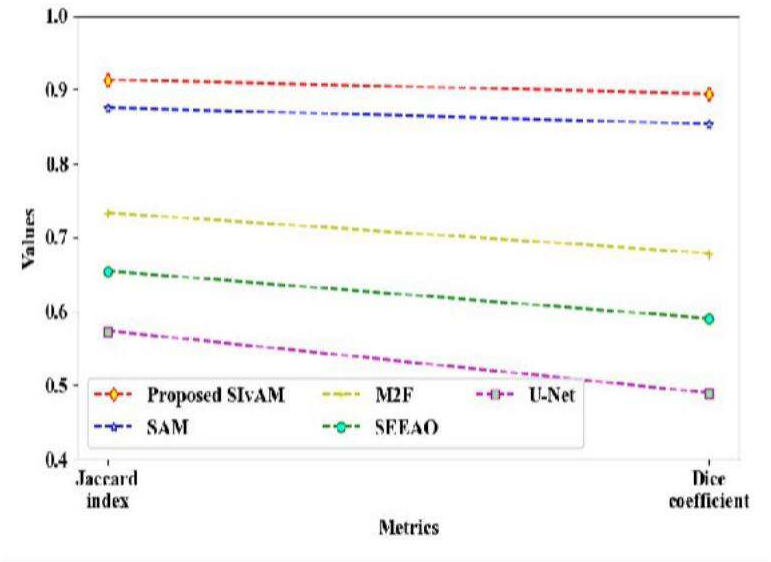


Figure 7: Performance analysis of the proposed SIvAM and conventional algorithms

To showcase the proposed model's efficiency in manipulated region localization, the proposed SIvAM's performance is compared with conventional techniques, such as SAM, Mask2Former (M2F), Segment Everything Everywhere All at Once (SEEAO), and U-Net, in Figure 7. The proposed SIvAM attained jaccard index of 0.9133 and a Dice coefficient of 0.8943. Nevertheless, the conventional SAM achieved jaccard index of 0.8753 and a Dice coefficient of 0.8534. Here, due to improper weight initialization, the traditional methods had limited performance. Owing to the inclusion of IvF-based weight initialization, the proposed approach's performance was improved. Hence, it was absolutely clear that the proposed work had better performance in manipulated region localization.

4.7 Comparative analysis of the proposed methodology

Here, to prove the proposed framework's dominance in DII, the proposed work's performance is contrasted with several existing studies.

Table 5: Comparative validation of the proposed framework

Author's name	Algorithm	Accuracy (%)	Precision (%)	F-measure (%)
Proposed work	SETCR2Sc2N	99.3273	99.3023	99.3053
(Soudy et al., 2024)	CNN	97	-	-

(Kumar et al., 2025)	Enhanced MobileNetV2	99.10	98.23	99.10
(Javiya et al., 2025)	Adaptive deep convolutional GAN (Adaptive-DCGAN) with a lightweight CNN	95	95	95
(Xia et al., 2023)	Multi-Collaboration and Multi-Supervision Network (MMNet)	73.93	-	-
(Anwar et al., 2022)	Support vector machine	96.24	99.6	98

In Table 5, the proposed work's performance is compared with related studies regarding DF identification. Owing to the inclusion of subtle facial reenactment inconsistencies, the proposed approach obtained superior performance in DII. Here, the proposed SETCR2Sc2N acquired an accuracy of 99.3273%, precision of 99.3023%, and f-measure of 99.3053%. Similarly, to perform DII, the existing studies utilized methods, namely support vector machine, MMNet, enhanced MobileNetV2, and CNN. The existing enhanced MobileNetV2 (Kumar et al., 2025) attained an accuracy of 99.10%, precision of 98.23%, and f-measure of 99.10%. Nevertheless, owing to the lack of fine-grained feature pattern learning and subtle feature analysis, the traditional frameworks had limited performance. Overall, the inclusion of R2s activation helped to enhance the proposed neural network's performance, demonstrating the proposed work's superiority.

5. CONCLUSION

This paper presented an FRI-aware explainable DII framework using DVPWT and SETCR2Sc2N. The images were significantly decomposed into various sub-bands by the proposed DVPWT, thereby capturing the subtle distortions. The proposed DVPWT obtained an MAE of 1.32 and an SSIM of 0.9467 regarding frequency-domain transformation. Here, the images were effectively classified into real and fake by the proposed SETCR2Sc2N with an accuracy of 99.3273%. Furthermore, a novel SHADP helped to enhance the model's trust through decision interpretation. Therefore, as per the experimental outcomes, the proposed SHADP had a fidelity of 0.935. To upgrade the proposed system's efficiency, the critical processes, such as accurate face region detection, fine-grained visual pattern learning, and manipulated region localization, were carried out. Collectively, the identity threat was significantly mitigated by the proposed approach, which also improved public trust through the robust DII.

Practical implications: A variety of real-time applications, including digital forensics, social media content moderation, biometric authentication systems, and cybercrime investigation platforms, are effectively supported by the proposed FRI-aware DII framework. Trust and accountability are enhanced by the interpretable decision mechanism of the proposed work, making it appropriate for real-world applications.

Future scope: To enhance the DF identification's accuracy and significance, the future work will concentrate on temporal modeling and spatio-temporal consistency analysis. Furthermore, in the future, data balancing/data augmentation will be done to upgrade the robustness against different DF manipulation techniques.

Funding:

I have not received any kind of funds for this research.

Conflict of Interest:

There is no conflict of interest between the authors.

REFERENCES:

Dataset:

https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection

Akshatha, G., & Kempanna, M. (2025). Review of Deep Learning Strategies in Modern Fake Face Identification Systems. *Grenze International Journal of Engineering and Technology*, 1–11. https://thegrenze.com/pages/servej.php?fn=464_21.pdf&name=Review of Deep Learning Strategies in Modern FakeFace Identification Systems&id=5038&association=GRENZE&journal=GIJET&year=2025&volume=11&issue=2

Akshatha, G., Kempanna, M., Ashoka, S. B., Prasanth, J., & Chinta, K. (2026). Deep Learning for Facial Forgery Detection Performance Evaluation of DenseNet201, InceptionV3 and ConvNeXt. *Journal of Machine and Computing*, *6*(1), 48–57. https://doi.org/10.53759/7669/jmc202606005

Akshathag, Madhurakavi, S., Shetty, S., Suhas, C. M., & Kumar, R. (2025). Advancing AI-Generated Image Detection. *Grenze International Journal of Engineering and Technology*, 1–9. https://thegrenze.com/pages/servej.php?fn=650_21.pdf&name=Advancing AI-Generated Image Detection&id=5066&association=GRENZE&journal=GIJET&year=2025&volume=11&issue=2

Al-Dulaimi, O. A. H. H., & Kurnaz, S. (2024). A Hybrid CNN-LSTM Approach for Precision Deepfake Image Detection Based on Transfer Learning. *Electronics (Switzerland)*, *13*(9), 1–22. https://doi.org/10.3390/electronics13091662

Anwar, M. A., Tahir, S. F., Fahad, L. G., & Kifayat, K. (2022). Image forgery detection by transforming local descriptors into deep-derived features. *Applied Soft Computing*, 1–33.

https://doi.org/10.1016/j.asoc.2023.110730

Arshed, M. A., Mumtaz, S., Ibrahim, M., Dewi, C., Tanveer, M., & Ahmed, S. (2024). Multiclass AI-Generated Deepfake Face Detection Using Patch-Wise Deep Learning Model. *Computers*, *13*(1), 1–18. https://doi.org/10.3390/computers13010031

Bird, J. J., & Lotfi, A. (2024). CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *IEEE Access*, *12*, 15642–15650. https://doi.org/10.1109/ACCESS.2024.3356122

Coccomini, D. A., Zilos, G. K., Amato, G., Caldelli, R., Falchi, F., Papadopoulos, S., & Gennaro, C. (2024). MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection. *IEEE Transactions on Information Forensics and Security*, *19*, 6084–6096. https://doi.org/10.1109/TIFS.2024.3409054

Heidari, A., Navimipour, N. J., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *14*(2), 1–45. https://doi.org/10.1002/widm.1520

Islam, A., & Belhaouari, S. B. (2023). Fast and Efficient Image Generation Using Variational Autoencoders and K-Nearest Neighbor OveRsampling Approach. *IEEE Access*, *11*, 28416–28426. https://doi.org/10.1109/ACCESS.2023.3259236

Javiya, D., Jethava, G., Parmar, H., & Jethava, S. (2025). Enhancing deepfake detection with Adaptive-DCGAN and Lite-CNN: a novel approach to image classification. *Discover Applied Sciences*, *7*(11), 1–21. https://doi.org/10.1007/s42452-025-07690-y

Khalifa, A. H., Zaher, N. A., Abdallah, A. S., & Fakhr, M. W. (2022). Convolutional Neural Network Based on Diverse Gabor Filters for Deepfake Recognition. *IEEE Access*, *10*, 22678–22686. https://doi.org/10.1109/ACCESS.2022.3152029

Khormali, A., & Yuan, J. S. (2024). Self-Supervised Graph Transformer for Deepfake Detection. *IEEE Access*, *12*, 58114–58127. https://doi.org/10.1109/ACCESS.2024.3392512

Kolagati, S., Priyadharshini, T., & Rajam, V. M. A. (2022). Exposing deepfakes using a deep multilayer perceptron – convolutional neural network model. *International Journal of Information Management Data Insights*, *2*(1), 1–8. https://doi.org/10.1016/j.jjimei.2021.100054

Kumar, B. A., Misra, N. K., Pathak, N., Ahmadpour, S. S., Krishnamoorthy, M., Shukla, D. K., Patidar, M., & Hakimi, M. (2025). Hybrid CMNV2: DeepFake faces classification and recognition using deep learning methods. *Results in Engineering*, *28*, 1–14. https://doi.org/10.1016/j.rineng.2025.107513

Li, S., Dutta, V., He, X., & Matsumaru, T. (2022). Deep Learning Based One-Class Detection System for Fake Faces Generated by GAN Network. *Sensors*, *22*(20), 1–23. https://doi.org/10.3390/s22207767

Li, Y., Liu, X., Wang, X., Lee, B. S., Wang, S., Rocha, A., & Lin, W. (2024). FakeBench:

Probing Explainable Fake Image Detection via Large Multimodal Models. *IEEE Transactions on Information Forensics and Security*, 1–13. https://doi.org/10.1109/TIFS.2025.3597211

Malik, A., Kuribayashi, M., Abdullahi, S. M., & Khan, A. N. (2022). DeepFake Detection for Human Face Images and Videos: A Survey. *IEEE Access*, *10*, 18757–18775. https://doi.org/10.1109/ACCESS.2022.3151186

Mansoor, N., & Iliev, A. I. (2025). Explainable AI for DeepFake Detection. *Applied Sciences (Switzerland)*, *15*(2), 1–21. https://doi.org/10.3390/app15020725

Mathews, S., Trivedi, S., House, A., Povolny, S., & Fralick, C. (2023). An explainable deepfake detection framework on a novel unconstrained dataset. *Complex and Intelligent Systems*, *9*(4), 4425–4437. https://doi.org/10.1007/s40747-022-00956-7

Momin, M. S., Sufian, A., Barman, D., Leo, M., Distante, C., & Damer, N. (2025). Explainable deepfake detection across different modalities: An overview of methods and challenges. *Image and Vision Computing*, *163*, 1–16. https://doi.org/10.1016/j.imavis.2025.105738

Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I. E., & Mazibuko, T. F. (2023). An Improved Dense CNN Architecture for Deepfake Image Detection. *IEEE Access*, *11*, 22081–22095. https://doi.org/10.1109/ACCESS.2023.3251417

Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., Aluvala, S., & Vimal, V. (2023a). Deepfake Generation and Detection: Case Study and Challenges. *IEEE Access*, *11*, 143296–143323. https://doi.org/10.1109/ACCESS.2023.3342107

Preeti, Kumar, M., & Sharma, H. K. (2023). A GAN-Based Model of Deepfake Detection in Social Media. *Procedia Computer Science*, *218*, 2153–2162. https://doi.org/10.1016/j.procs.2023.01.191

Raza, A., Munir, K., & Almutairi, M. (2022). A Novel Deep Learning Approach for Deepfake Image Detection. *Applied Sciences (Switzerland)*, *12*(19), 1–15. https://doi.org/10.3390/app12199820

Shahzad, H. F., Rustam, F., Flores, E. S., Mazón, J. L. V., Diez, I. de la T., & Ashraf, I. (2022). A Review of Image Processing Techniques for Deepfakes. *Sensors*, *22*(12), 1–28. https://doi.org/10.3390/s22124556

Soudy, A. H., Sayed, O., Tag-Elser, H., Ragab, R., Mohsen, S., Mostafa, T., Abohany, A. A., & Slim, S. O. (2024). Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Computing and Applications*, *36*(31), 19759–19775. https://doi.org/10.1007/s00521-024-10181-7

Tran, V. N., Kwon, S. G., Lee, S. H., Le, H. S., & Kwon, K. R. (2023). Generalization of Forgery Detection With Meta Deepfake Detection Model. *IEEE Access*, *11*, 535–546. https://doi.org/10.1109/ACCESS.2022.3232290

Waqas, N., Safie, S. I., Kadir, K. A., Khan, S., & KHEL, M. H. K. (2022). DEEPFAKE Image Synthesis for Data Augmentation. *IEEE Access*, *10*, 80847–80857.

https://doi.org/10.1109/ACCESS.2022.3193668

Waseem, S., Bakar, S. A. R. S. A., Ahmed, B. A., Omar, Z., Eisa, T. A. E., & Dalam, M. E. E. (2023). DeepFake on Face and Expression Swap: A Review. *IEEE Access*, *11*, 117865–117906. https://doi.org/10.1109/ACCESS.2023.3324403

Wolter, M., Blanke, F., Heese, R., & Garcke, J. (2022). Wavelet-packets for deepfake image analysis and detection. In *Machine Learning* (Vol. 111). Springer US. https://doi.org/10.1007/s10994-022-06225-5

Xia, R., Liu, D., Li, J., Yuan, L., Wang, N., & Gao, X. (2023). MMNet: Multi-Collaboration and Multi-Supervision Network for Sequential Deepfake Detection. *IEEE Transactions on Information Forensics and Security*, 1–13. https://doi.org/10.1109/TIFS.2024.3361151

Yu, C., Zhang, X., Duan, Y., Yan, S., Wang, Z., Xiang, Y., Ji, S., & Chen, W. (2023). Diff-ID: An Explainable Identity Difference Quantification Framework for DeepFake Detection. *IEEE Transactions on Dependable and Secure Computing*, *14*(8), 1–15. https://doi.org/10.1109/TDSC.2024.3364679