



ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.60, p.181-194, July/August, 2026.

DOI: <https://doi.org/10.5935/jetia.v12i60.3686>

RESEARCH ARTICLE

OPEN ACCESS

DETECTING IMAGE FORGERIES USING HYBRID RESIDUAL-DENSE SWIN TRANSFORMER FRAMEWORK WITH ATTENTION-BASED UNET++

Sruthi Anand*¹, Saranya V²¹Assistant Professor, Department of Information Technology, Vasavi College of Engineering, Hyderabad- 500031, Telangana, India.²Associate Professor, Department of Information technology, Sri Eshwar College of Engineering, Vadasithur, Coimbatore- 641202, India.¹<https://orcid.org/0000-0001-5344-4826>^{id}, ²<https://orcid.org/0000-0002-1883-4383>^{id}Email: *sruthianand@staff.vce.ac.in, saranya.v@sece.ac.in

ARTICLE INFO

Article History

Received: February 14, 2026

Reviewed: June 6, 2026

Accepted: July 6, 2026

Published: July 31, 2026

Keywords:Image forgery,
Image preprocessing,
Residual-Dense Swin transformer,
UNET++,
Channel-Dual attention.

ABSTRACT

Establishing the authenticity of digital images is widely acknowledged as a crucial aspect of multimedia security since contemporary alteration techniques like copy-move, splicing, and retouching create remarkably convincing forgeries that are difficult to spot. This research proposes a Deep Learning (DL) based image forgery detection methodology that combines attention-guided segmentation and hybrid feature extraction to deliver robust tamper detection and classification. The preprocessing procedures including resizing and normalizing images are used to ensure consistent model performance. The hybrid encoder, records local texture-level inconsistencies and long-range contextual dependencies, composed of Swin Transformer modules and Residual-Dense blocks. Multi-scale feature maps that are highly discriminative for forgery cues are the outcome of this. These features are decoded using a UNet++ segmentation network that has been altered using Channel-Dual Attention (CA-DA). This improves the spatial focus and boundary precision of the projected forgery mask. The created mask is analyzed in a detection stage, where threshold-based decision-making mechanism determines if the input is genuine or forged, non-empty masks that exceed the threshold indicate tampered images, whereas actual images are represented by empty masks. The proposed system is implemented in Python software, and the results demonstrate that the proposed experimental validation on benchmark datasets indicates improved detection accuracy, robustness to noise compression, and localization clarity when compared to conventional approaches.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Images and videos are attracting a lot of attention recently since they serve as the primary information source. In numerous contexts, images are utilized to convey information to various recipients. They are used in many different areas, including intelligence gathering, military operations, court testimony, social media disinformation, defamation of well-known individuals, and many more [1]. In addition to the unreliable methods of image dissemination, image manipulation compromise the legitimacy of an image. Authentication solutions were created in order to secure the image communication process [2]. There are numerous ways used for forging an image, however image manipulation techniques are usually categorized into two groups, content-changing and content-preserving. Semantic content is changed as a result of content-changing, which reshapes image content at random. The three most popular forms of image forgery detection that have been thoroughly researched are copy-move, splicing, and removal [3]. The most common type is copy-move forgery, in which a portion of the image is copied and pasted somewhere else in the same image. Contrarily, content-preserving forgeries change features like contrast, brightness, compression artifacts or noise patterns to deceive analysts rather than changing the image's semantic content. It is difficult to identify these subtle changes with conventional techniques [4]. Techniques for detecting image forgeries are broadly categorized into five groups. Geometry based techniques measure objects in the world and their positions in relation to the camera. Pixel based techniques detect statistical anomalies introduced at the pixel level. Format based techniques utilize the benefit of statistical correlations introduced by a specific lossy compression scheme [5].

Camera based techniques exploit artifacts introduced by the camera lens, sensor or on-chip post-processing and physical environment based techniques explicitly model and detect anomalies in the three dimensional interactions between physical objects, light and camera. One of the most prevalent types of tampering is copy or move forgery, which involves copying a portion of an image from one area and then passing it at more points in the same image or another image, ideally of the same scene [6]. This research illustrates the image forgery detection through various stages. The procedure commences with input image where raw data is provided to the system for forgery detection. This is followed by preprocessing where resize and normalization takes place to ensure consistent input for the model, which helps in training and inference stability [7]. In feature extraction, it extracts high level features from the input image using a hybrid model. The extracted features are then segmented to produce a pixel-wise segmentation mask indicating forged images. A predicted pixel mask highlighting suspected forged regions in the image. Then the mask is subjected to a threshold by the detection module and if exceeded, it marks the image as forged [8]. Convolutional Neural Networks (CNNs) have become increasingly popular in image detecting community throughout the past ten years. In order to identify the optimal features for categorizing image models, these algorithms have concentrated on training the CNN. CNN has the benefit of immediately extracting features from the image collection. The main benefit of CNN is that immediately extract classification features from image data. However, it is quite challenging to comprehend the generated image identification model and its reliance on a particular training dataset [9].

VGG operates by recursively partitioning the feature space into branches based on feature values, resulting in a tree structure with each leaf node representing a class label, authentic or fabricated. This makes VGG simple, interpretable, and easy to apply. However, it has a tendency to over fit training data, which is sensitive to noise in features, and has worse accuracy for complicated or nuanced forgeries [10]. To further enhance the performance of the detection model, segmentation methods are utilized. SegNet identifies image forgeries by automatically learning hierarchical features from raw images. Convolutional layers extract local patterns, edges, and textures first, followed by non-linear activation functions such as ReLU. Pooling layers lower spatial dimensions while keeping crucial properties, and fully connected layers combine these qualities to determine whether an image is legitimate or faked. Back propagation is used to train the network and minimize a loss function while also optimizing feature representations [11]. However, their effectiveness in automatic feature learning requires a considerable amount of labeled data, which are computationally costly and fail to catch long-range or global discrepancies, and might overlook very subtle or sophisticated forgeries [12]. To overcome these limitations, this research proposes a hybrid Residual-Dense Swin transformer for feature extraction encoder and UNet++ with CA-DA for segmentation encoder. This provides a predicted forgery mask that highlights potential altered places. The detecting stage assesses this mask to establish the image's validity. If the anticipated mask exceeds a predefined threshold, the image is classified as fabricated otherwise, an empty mask is classified as authentic.

1.1 RELATED WORKS

By [13] have proposed a feature extraction based LeNet-5 for detecting fake media content. LeNet-5 method processes and analyzes textual information by transforming text data into numerical representations like TF-IDF, Bag-of-Words or word embedding. This procedure enables LeNet-5 to detect bogus news on social media. In order to distinguish between fake and real postings using the attributes that were retrieved, it then learns a hyperplane in a high-dimensional space. However, it struggles with noisy and high-dimensional social media data, including slang, abbreviations, and grammatical errors which reduce the overall classification accuracy. By [14] have developed a fast and robust speckle pattern authentication with the Scale Invariant Feature Transform (SIFT) Algorithm in Physically Unclonable Functions (PUF). The use of SIFT with the Challenge Response Pairs (CRPs) protocol allows us to correctly authenticate a client while blocking any other fraudulent access. In this approach, the authentication procedure is highly trustworthy even in the face of answer rotation, zooming, and cropping, which might occur in consecutive PUF interrogations and to which other post processing algorithms are quite susceptible. Nonetheless, the analysis methodologies used are highly sensitive to minor alterations in the speckle pattern, which lead to failure in PUF authentication due to changes in position, illumination, or rotation. By [15] have presented an image forgery detection based on Histogram of Oriented Gradients (HOG) feature extraction. It is generated by combining edge orientations caused by image alteration into a one-dimensional histogram of specified image regions. The suggested method collects HOG characteristics from the YCbCr colour space and applies them to the discrete wavelet transform's high-frequency subbands.

These collected features are then fed into a classifier, which determines whether the image is legitimate. Nevertheless, HOG features primarily collect edge and gradient information, which is insufficient to detect forgeries that do not drastically change these characteristics. By [16] have established a forged image detection based on clustering algorithms and transform domain feature analysis. The proposed method is highly effective in terms of detection accuracy, offering a viable solution to the issues of modern multimedia manipulation. Texture features are retrieved using the Discrete Wavelet Transform (DWT) technique, which improves the resilience and reliability of the forgery detection process. Though, it requires large amounts of labeled video data and significant computational resources, making training and real-time deployment challenging. In turn [17] have proposed a semantic segmentation-based approach for identifying and locating manipulated regions in images. In this research, the semantic segmentation is used to improve object awareness and increase object localization accuracy. Instead of considering semantic segmentation as a separate submodule, we incorporate it as a third branch of an end-to-end framework. This method balances visual, compression, and segmentation artifacts while minimizing overemphasis on any single branch. However, it struggles with noisy, unclear or extremely context-dependent content particularly in low-resource languages or images with subtle alterations, thereby lowering overall detection accuracy.

1.2 CONTRIBUTIONS

The main contributions of this hybrid residual-dense swin transformer framework with attention-based unet++ research includes,

- To effectively capture both local and global image features for forgery detection a hybrid residual-dense swin transformer framework with attention-based Unet++ is integrated.

- To develop a global context with local texture details for effective feature extraction, a hybrid residual-dense Swin transformer encoder is utilized.
- To enable precise localization of manipulated regions, spatial channel attention with nested skip connections, an Attention-based U-Net++ decoder is implemented.

I.3 PAPER ORGANIZATION

This paper is organized as follows: Section II provides an explanation of the proposed system description. The study methodology is thoroughly described in Section III, which also includes examples of the process, methods, and mathematical representations. The experimental results of this work are described in Section IV, which also include the simulation setup, comparison analysis, and research summary. The conclusion of the proposed investigation is explained in Section V.

II. PROPOSED SYSTEM DESCRIPTION

This paper aims to enhance the accuracy, robustness and localization clarity in detecting image forgeries by integrating image preprocessing and hybrid residual-dense swin transformer framework with attention-based Unet++ for feature extraction and segmentation. The proposed block diagram is displayed in Figure 1.

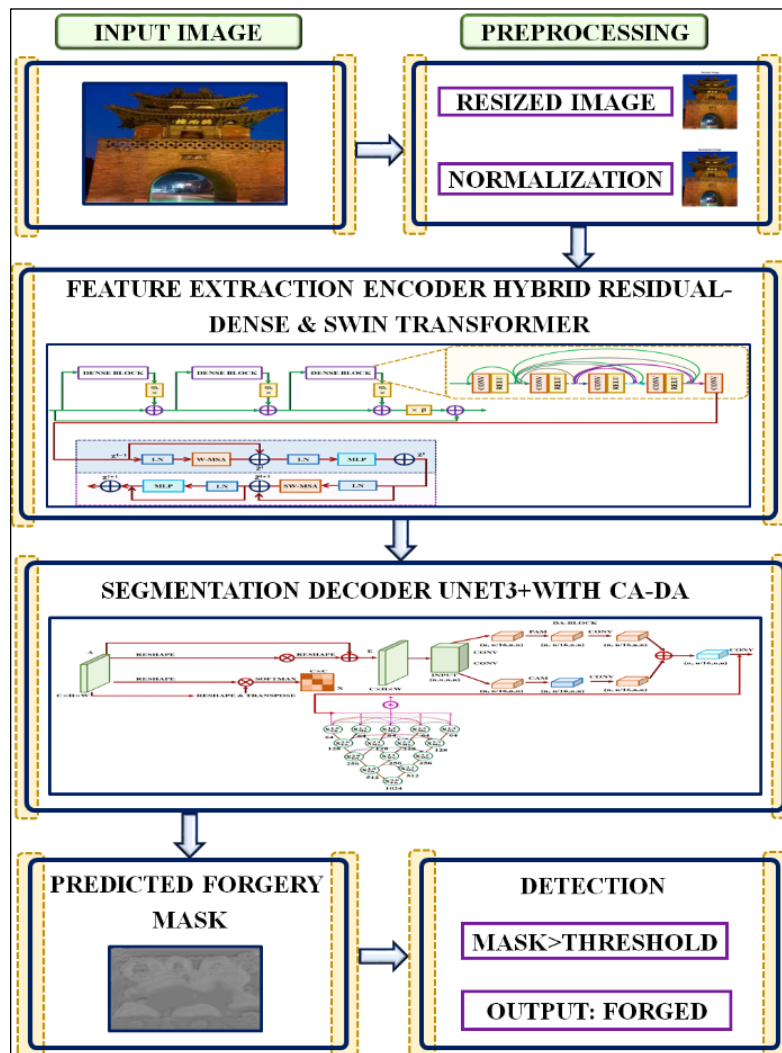


Figure 1: Proposed block diagram.

Source: Authors, (2026).

The proposed approach integrates transformer based and convolutional techniques to present a strong DL framework for image forgery detection. To ensure consistency, the image that arrives is initially processed by resized and normalizing. Following processing, the image is fed into a hybrid feature extraction encoder that successfully captures both local and global contextual data by combining residual-dense blocks with a Swin transformer. A UNet++ architecture supplemented with CA-DA is then used to decode these complex feature representations, allowing for the exact localization of modified regions. A predicted forgery mask is produced by the decoder and examined during the detection stage. The image is considered fake if the mask's confidence level exceeds a predetermined threshold. High accuracy and locating image forgeries is ensured by this end-to-end process.

III. PROPOSED SYSTEM MODELLING

III.1 IMAGE PREPROCESSING

In image preprocessing, resizing and normalizing is a crucial step in image analysis pipeline. To ensure consistency and compatibility with the DL model, an image must be standardized before being examined for potential alteration.

III.1.1 Resizing

The input image is resized to a fixed resolution, ensuring that all images have uniform dimensions regardless of their original size. This step makes the computational process efficient and allows the neural network to handle images in batches. In addition, resizing helps maintain consistency across the dataset, preventing variations in image size from affecting model learning. It also reduces memory requirements and speeds up both training and inference. By standardizing resolution, the network focus on learning important patterns instead of adjusting to irregular input sizes.

III.1.2 Normalizing

Pixel values are normalized and typically scaled to a certain range. Normalization stabilizes the training process, accelerates convergence, and assures that changes in image brightness or contrast do not influence forged detection. It also prevents numerical instability by maintaining input values within a reasonable range, hence avoiding bursting or vanishing gradients during back propagation. Normalization ensures that all features contribute proportionally, which enhances the model's capacity to detect small irregularities. It improves generalization by making the model less susceptible to differences in lighting, exposure, or camera settings. In some circumstances, mean subtraction and standard deviation scaling are employed to align pixel distributions with retrained models, which improves feature extraction performance. The diagram for image preprocessing is shown in Figure 2.

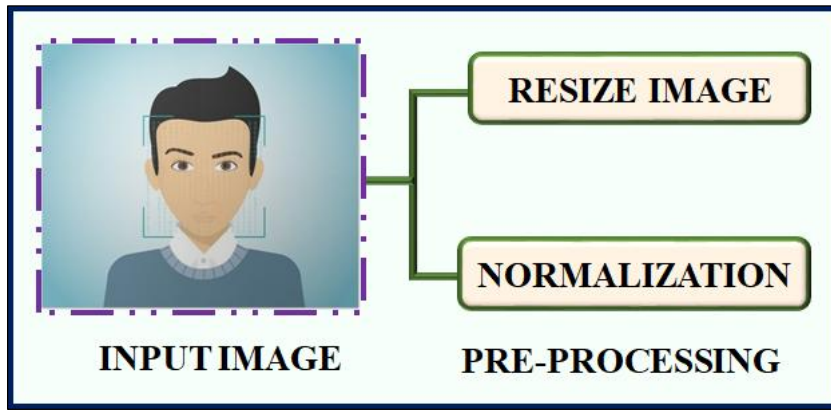


Figure 2: Image preprocessing.
Source: Authors, (2026).

By combining both, Resizing and normalization prepare the raw image for subsequent processing phases, such as feature extraction and segmentation, allowing the detection model to concentrate on forgery-specific patterns rather than changes caused by size or illumination discrepancies. Resizing maintains constant input dimensions, allowing for batch processing and reducing computing effort, whereas normalization scales pixel intensity values to a uniform range, enhancing model convergence and stability during training. Furthermore, these processes serve to reduce the impact of noise, contrast discrepancies, and illumination variations making the model more adaptable to a variety of image situations.

III.2 FEATURE EXTRACTION- RESIDUAL-DENSE SWIN TRANSFORMER

Residual Dense Network (RDN) is an architecture that identifies image forgery due to its outstanding feature extraction capabilities. In order to detect fabricated regions in images, fine-grained details, subtle discrepancies, and manipulation effects need to be precisely captured. The RDN combines residual and dense connectivity, allowing for effective extraction of both local and global characteristics. The dense block structure allows the network to transfer information within each block, efficiently merging low-level and high-level elements including texture irregularities, edges, and noise patterns, which frequently signal tampering. By combining the input and output of each dense block, the network maintains a strong information flow, which is crucial for detecting small changes that would otherwise go undetected. Mathematically, the residual network adds the input x_{i-1} to its mapped output $F(x_{i-1})$, enhancing gradient flow and preventing overfitting as network depth increases,

$$F_{Res}(x_{i-1}) = F(x_{i-1}) + x_{i-1} \quad (1)$$

However, fundamental characteristics within a block are unable to fused by residual connections alone. By combining information from all preceding layers, the dense block overcomes this restriction and allows the network to detect minute local imperfections including texture anomalies, noise disturbances, and duplicated regions. Dense blocks increase the network's sensitivity to fine-grained modifications that conventional convolutional layers could miss by combining data from several layers.

Dense blocks offer an additional technique that combines local detail capture with global information flow when paired with residual connections. This improves the overall performance of image fraud detection across a variety of manipulation types, including splicing, copy-move, and inpainting.

$$F_{DB}^i = \text{ReLU}(F_{DB}^{i-1} + \sum_{j=0}^{i-1} F_j(x_j)) \quad (2)$$

The detection of forgeries is further enhanced by the introduction of residual dense blocks (RDBs), which create global residual connections between blocks because dense blocks only concentrate on internal feature fusion. By combining the benefits of dense and residual connections, RDBs make sure that data moves throughout the network as well as within individual blocks. RDBs improve the network's capacity to identify minute manipulations like splicing edges, copy-move duplications, or anomalies in inpainting by preserving inter-block dependencies. All things considered, RDBs offer a strong foundation for merging local and global data, which makes the network extremely efficient at precisely locating and detecting tampered areas in a variety of forging scenarios.

$$F_{RDB} = \beta \cdot F_{DB}(x_i) + x_{i-1}, \beta = 0.2 \quad (3)$$

In addition to improving the network's capacity to extract local and global features, RDN's residual and dense connections provide effective gradient flow, which permits the training of deeper architectures without overfitting. Dense blocks capture tiny tampering artifacts like duplicated regions, noise discrepancies, or splicing edges by allowing the fusion of all previous layer characteristics within a block. The network preserves inter-block connections by incorporating these dense blocks into Residual Dense Blocks (RDBs), which highlights fine-grained local characteristics while keeping global contextual information. RDN is extremely resistant to a variety of forgery techniques, including as splicing, copy-move, and inpainting manipulations, thanks to its hierarchical feature extraction. The Swin Transformer is a hierarchical vision model that is well suited for detecting image forgeries because it capture both local defects and global contextual anomalies. The Swin Transformer uses hierarchical feature maps and a shifting window attention mechanism to analyze data at many scales, which is necessary for detecting manipulations like splicing, copy-move, or inpainting. Subsequently, Swin Transformer Blocks (STBs) are used to convert the image into non-overlapping patches. Layer Normalization (LN), a two-layer Multilayer Perceptron (MLP) with Gaussian Error Linear Unit (GELU) activation, and Window-Based Multi-Head Self-Attention (W-MSA) and Shifting Window MSA (SW-MSA) modules make up each block. The computational complexity of the conventional multi-head self-attention is as follows,

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \quad (4)$$

To address this, the Swin Transformer uses window-based multi-head self-attention (W-MSA), which restricts attention computation to local windows of size $M \times M$. This reduces the complexity to,

$$\Omega(W-MSA) = 4hwC^2 + 2M^2hwC \quad (5)$$

This reduction in complexity allows the network to efficiently focus on local regions of the image, which is particularly important for detecting small-scale forgeries such as copy-move or subtle splicing artifacts. By limiting attention computation to each window, the model analyse fine-grained textures, edges, and noise inconsistencies without being overwhelmed by the total number of patches in high-resolution images. This local attention mechanism forms the foundation for identifying micro-level tampering cues, which are often missed by traditional convolutional or global attention methods. The structure of residual-dense swin transformer is shown in Figure 3.

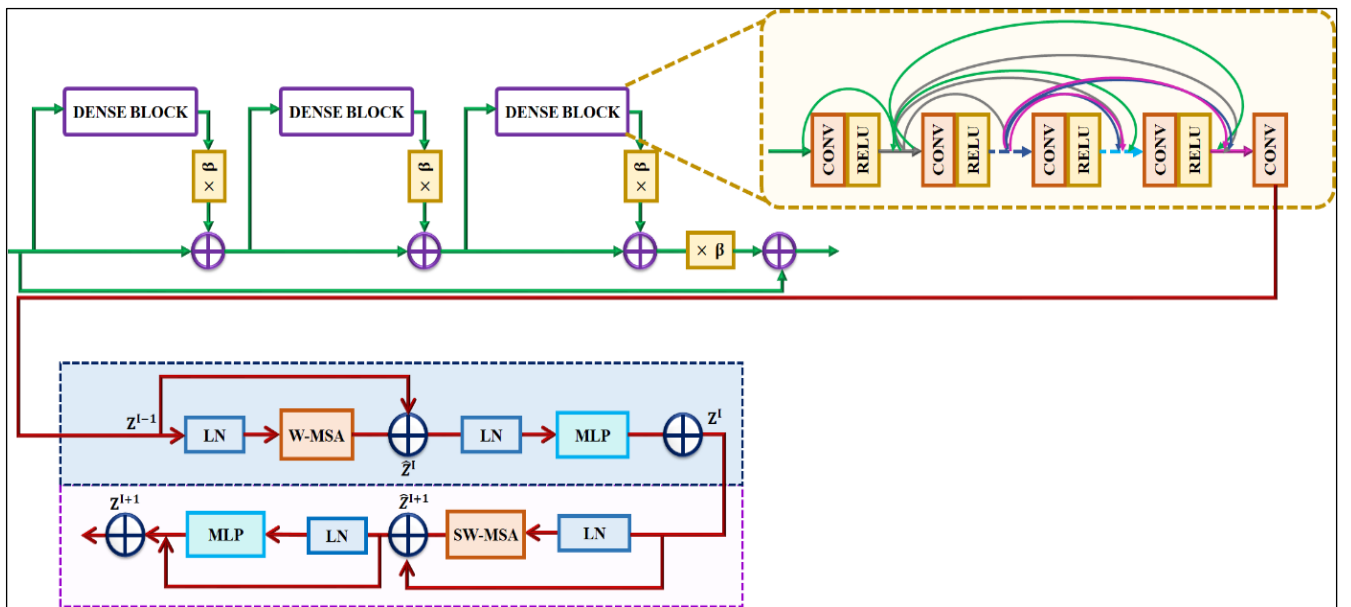


Figure 3: Structure of residual-dense swin transformer.
Source: Authors, (2026).

By focusing on local regions, W-MSA efficiently captures fine-grained artifacts, such as subtle texture inconsistencies or duplicated areas, which are common in tampered images. Within each Swin Transformer Block (STB), features are processed sequentially. Initially, the W-MSA module computes local dependencies for each patch, incorporating residual connections to stabilize training and preserve low-level details,

$$\hat{Z}_l = W - MSA(LN(Z_{l-1})) + Z_{l-1} \quad (6)$$

Here, Z_{l-1} is the input feature from the previous layer, LN is Layer Normalization, and the addition of Z_{l-1} ensures that the original patch information is retained. After this, the features are passed through a two-layer MLP with GELU activation to enhance non-linear representations,

$$Z_l = MLP(LN(\hat{Z}_l)) + \hat{Z}_l \quad (7)$$

Whereas W-MSA focuses on local interactions, it cannot capture relationships across neighboring windows, which is important for detecting manipulations spanning multiple regions, such as a spliced object. The shifted window MSA (SW-MSA) addresses this by cyclically shifting the windows and computing attention over overlapping patches,

$$\hat{Z}_{l+1} = SW - MSA(LN(Z_l)) + Z_l \quad (8)$$

$$Z_{l+1} = MLP(LN(\hat{Z}_{l+1})) + \hat{Z}_{l+1} \quad (9)$$

Cross-window connections are essential for modeling long-range dependencies and this shifted window method makes computation efficient. The model successfully collects both local and global contextual information by switching between normal window based self-attention and shifting window based self-attention across consecutive levels. Finally, the attention mechanism is shown below,

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (10)$$

Here, Q, K , and V are the query, key, and value matrices for a given window, d is the dimension of feature vectors, and B encodes relative positions between patches. By using this mechanism, the Swin Transformer accurately identify irregularities in spatial structure, texture, and edges that indicate image forgery. Overall, this combination of residual-dense swin transformer is used to detect forgeries at multiple scales, generating pixel-level manipulation maps and ensuring both local artifact sensitivity and global structural consistency, which are essential for robust forensic analysis.

III.3 SEGMENTATION- UNet++ WITH CHANNEL-DUAL ATTENTION

To detect image forgery, it is necessary to determine the exact location of the tampered areas in addition to determining if the image was actually altered. Although a conventional UNet is capable of recording some semantic and spatial data, it eliminate important details like boundaries and forged textures since its straight skip connections only convey information at the same size. To get around this, UNet++ employs layered skip connections, which combine features from several layers while maintaining high-level semantic information and low-level geometric details. This fusion process is formalized by equations,

III.3.1 First Merge – Preserving Boundary Details

The first merging operation combines the shallowest encoder feature map $X1:0_{En}$ (low-level details) with the next encoder output $X2:0_{En}$

$$X1:1_{Me} = X1:0_{En} + X2:0_{En} \quad (11)$$

By employing layered skip connections to combine encoder features at several layers, UNet++ improves image detection. Low level features $X1:0_{En}$ and $X2:0_{En}$ are first combined to create $X1:1_{Me}$. This method enhances semantic comprehension and spatial detail, which increase object localization and border accuracy.

III.3.2 Second Merge – Adding Contextual Support

This equation expands the merge by reusing $X1:1_{Me}$, combining it with the shallow encoder feature $X1:0_{En}$ again and an intermediate feature from the second column $X2:1_{Me}$

$$X1:2_{Me} = X1:1_{Me} + X1:0_{En} + X2:1_{Me} \quad (12)$$

The frequent application of the shallow feature map $X1:0_{En}$ in this phase of the UNet++ model for image fraud detection helps to maintain fine-grained image characteristics including surface textures and small-scale imperfections by reinforcing local texture patterns. The intermediate feature map $X2:1_{Me}$, introduces spatial consistency, enabling the model to compare data from various image scales. By highlighting differences that might not be apparent at a single resolution, this cross-scale fusion aids in differentiating real textures from ones that have been altered. Because of this, the merged feature map $X1:2_{Me}$ improves its ability to detect minute differences between forged and real areas, like crooked or misaligned edges that are frequently observed in spliced forgeries.

III.3.3 Third Merge – Balancing Detail and Semantics

By integrating many previous nodes to create $1:3_{Me}$, the third merging stage becomes more thorough. In addition to deeper contextual features $2:2_{Me}$, this fusion incorporates data from both shallow and intermediate layers. This strikes a compromise between semantic-level abstraction and the preservation of fine-grained details.

$$X1:3_{Me} = X1:2_{Me} + X1:1_{Me} + X1:0_{En} + X2:2_{Me} \quad (13)$$

In the third merging stage of the UNet++ model, the fusion process integrates progressively higher-level semantic features from $X2:2_{Me}$ while still preserving low-level detail through $1:0_{En}$.

III.3.4 Final Merge – Forgery Segmentation Map

The output $X1:4_{De}$ is produced by the final decoder node aggregating all of the prior intermediate merges and deep decoder data.

$$X1:4_{De} = X1:3_{Me} + X1:2_{Me} + X1:1_{Me} + X1:0_{En} + X2:3_{De} \quad (14)$$

According to the equation, a decode image from system $X1$ and an enable image from system $X1$ add up to the decode image. On the contrary, include semantic information that enables the model to determine whether a suspicious area fits into the image's overall context, ensuring that additional significant irregularities or irrational alterations are detected. High-level semantics and granular detail are combined in the final result, which shows a hierarchical integration of these multi-scale elements. The network consistently discover forged areas even when tampering is subtle, well-blended, or purposefully done to avoid detection thanks to this fusion, which strengthens and improves the detection process. The Channel Attention Module (CAM) improves the network's capacity to focus on the most informative channels in feature maps, enabling it to discern between authentic and tampered regions. CAM reshapes a convolutional feature map $A \in R^{C \times H \times W}$ into $\in R^{C \times N}$, where $N = H \times W$, allowing for channel connections to be determined. The module then computes a channel attention map $X \in R^{C \times C}$ using the following softmax-based operation,

$$X_{ji} = \frac{\exp(A_i A_j)}{\sum_{i=1}^C \exp(A_i A_j)} \quad (15)$$

In this case, A_i and A_j are the flattened feature vectors for the i and j channels, as well as X_{ji} are shown. This metric assesses the impact of the i channel on the j channel. This step collects inter-channel relationships, allowing the network to determine which channels are most useful in detecting forgeries. Channels with high influence weights, for example, will be more sensitive to edge irregularities or odd texture patterns. The enhanced feature map $E \in R^{C \times H \times W}$ is then computed as follows,

$$E_j = \beta \sum_{i=1}^C X_{ji} A_i + A_j \quad (16)$$

Where β is a learnable scaling parameter. This process effectively generates a weighted total of all channels, along with the original feature, emphasizing informative channels and suppressing less relevant channels. By incorporating CAM into the UNet++ framework, the network gains both multi-scale feature fusion and channel-focused attention. Shallow features in UNet++ capture tiny details such as edges and textures, whereas deep features offer semantic context. The Dual Attention (DA) Block combines the strengths of the Position Attention Module (PAM) and the CAM to improve feature extraction for image fraud detection. PAM detects spatial correlations in feature maps, thereby learning the location where tampering happens by focusing on the positional links between pixels. CAM, on the other hand, captures inter-channel dependencies without highlighting what features are relevant for detecting manipulations, such as texture inconsistencies, odd edges, or cloned sections. By merging these two modules, the DA-Block extract spatial and channel-specific clues, which is critical for identifying subtle or complicated forgeries. The DA-Block has two parallel components. The first component uses convolution to scale the number of channels by a sixteenth, resulting in α_1 .

$$\alpha_1 = \text{Conv}(\text{input}) \quad (17)$$

This stage reduces the complexity of the feature map in preparation for PAM processing. The position attention module processes α_1 to capture spatial correlations, followed by another convolution to obtain the refined feature map $\tilde{\alpha}_1$,

$$\tilde{\alpha}_1 = \text{Conv}(\text{PAM}(\alpha_1)) \quad (18)$$

The second component of the Dual Attention (DA) Block for image forgery detection uses the CAM to stress channel-wise dependencies and highlight what aspects are most relevant for identifying alterations. The input features are initially run through a convolution to reduce dimensionality and create the feature map for the attention mechanism. CAM then computes inter-channel correlations and refines the features, followed by another convolution to generate the enhanced feature map

$$\alpha_2 = \text{Conv}(\text{input}) \quad (19)$$

$$\tilde{\alpha}_2 = \text{Conv}(\text{CAM}(\alpha_2)) \quad (20)$$

This procedure prioritizes channels that include crucial tampering cues, such as cloned textures, odd edges or inconsistent colours, while suppressing less informative channels. The first component of the DA-Block, led by the PAM, generates $\tilde{\alpha}_1$. Schematic diagram of UNet++ with Channel-Dual Attention is shown in Figure 4.

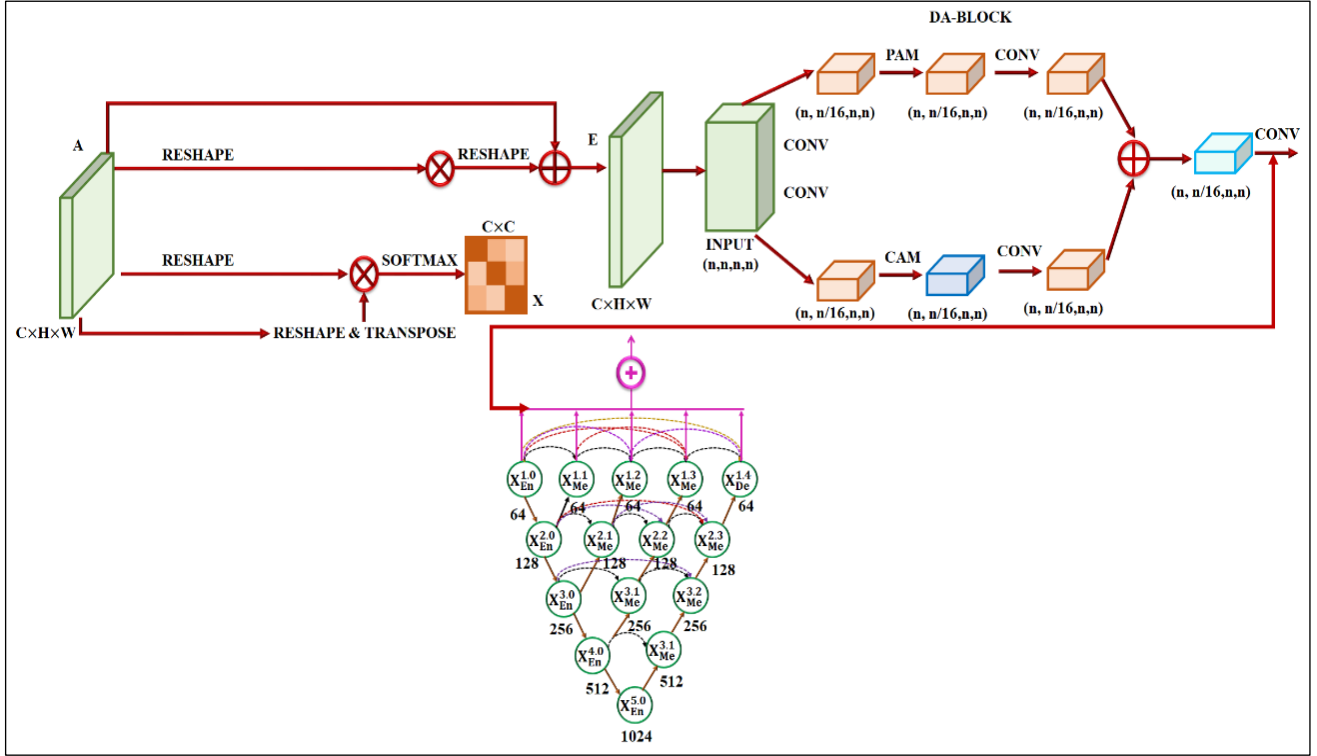


Figure 4: schematic diagram of UNet++ with Channel-Dual Attention.
Source: Authors, (2026).

By focusing on spatial dependencies or wherever modifications take place. The outputs from both routes are then aggregated and summed, and a final convolution restores the original number of channels, providing the final DA-Block output.

$$output = Conv(\tilde{\alpha}_1 + \tilde{\alpha}_2) \quad (21)$$

Combining $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$, the DA-Block blends spatial and channel-wise attention, allowing the network to detect subtle, well-blended, or sophisticated forgeries more efficiently by combining fine-grained details with global contextual information. In image forgery detection, this DA-Block allows the network to focus on both spatial inconsistencies and channel-specific information. The combination of PAM and CAM ensures that the model captures both fine-grained information and global context, allowing it to detect subtle, well-blended, or large-scale alterations. By modifying intermediate channel dimensions and convolutional procedures, the DA-Block modified to prioritize the most relevant forgery features, boosting detection accuracy and robustness across a wide range of image formats. Thus the integration of UNet++ with CA-DA focuses on accurately identifying tampered regions. CA-DA improves spatial and contextual awareness, whereas dense skip connections are used for detailed feature extraction. This makes the model useful for digital image detection by assisting it in identifying subtle modifications.

IV. RESULT AND DISCUSSION

This research provides a notable explanation of the manner in which a hybrid residual-dense Swin transformer framework coupled with an attention-based UNet++ is implemented to detect image forgeries efficiently.

Table 1: Train vs Test Distribution Table.

Class	Train samples	Test samples
Forged	5200	1300
Original	5200	1300

Source: Authors, (2026).

Train vs Test distribution of forged and original class is shown in Table 1. The train samples for both forged and original class is 5200. Similarly, for both forged and original class the test samples are 1300. The balanced split reduces bias towards any class. Overall, this distribution facilitates the development of a dependable and effective classification model.

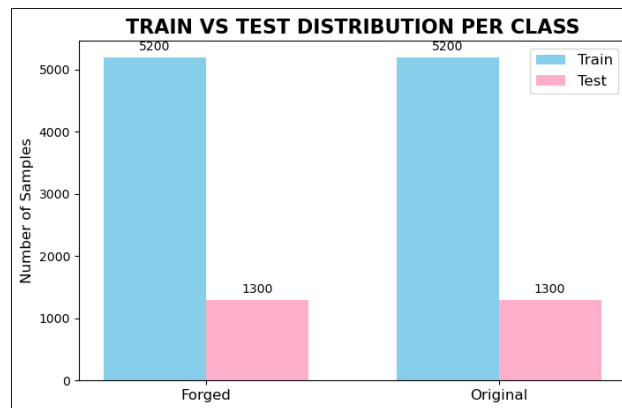


Figure 5: Train vs Test distribution per class.
Source: Authors, (2026).

The distribution of train and test samples for the two classes Forged and Original is displayed in Figure 5. There are 1,300 test samples and 5,200 training examples for every class. The distribution of the Forged class is comparable, with 1,300 samples used for testing and 5,200 samples used for training. With 5,200 training samples and 1,300 test samples, the Original class exhibits the same distribution pattern. This guarantees that the quantity of training and test samples for each class is equal. Comparing the quantity of samples for every class in the training and testing sets is made simple by the graph's visual representation of this division.

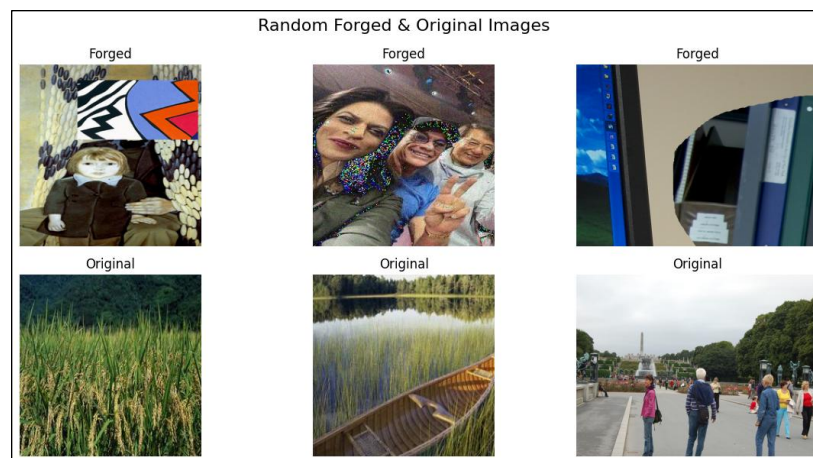


Figure 6: Random forged and original images.
Source: Authors, (2026).

A comparison of the original and forged images is shown in the Figure 6. The top row features three fabricated images, while the bottom row shows their real copies. Forgeries include distorted group selfies, cut-out regions that highlight irregularities, and a child's shot with an unnatural background. The originals feature natural sceneries such as a field, a boat on calm sea, and people walking in public. This side-by-side arrangement emphasizes visual cues and abnormalities. It helps to highlight the differences between authentic and digitally changed images. This comparison helps to identify digital forgeries more efficiently.

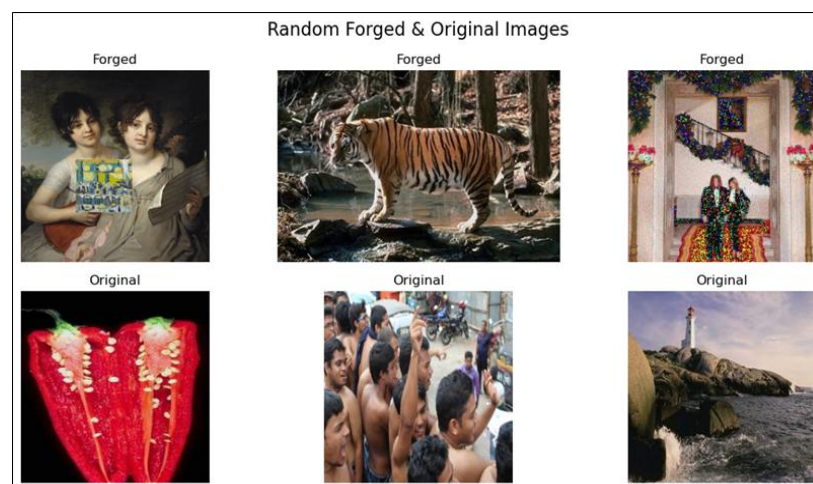


Figure 7: Random forged and original images.
Source: Authors, (2026).

A combination of forged and genuine images are displayed in Figure 7. Three Forged images a painting, a tiger in a forest, and a vibrant group portrait are displayed in the top row. A close-up of a pepper, a crowd of people outside, and a lighthouse near the sea are the three Original images that are shown in the bottom row. The arrangement provides insight into the kind of content contained in a classification job by emphasizing the difference between authentic and altered images.

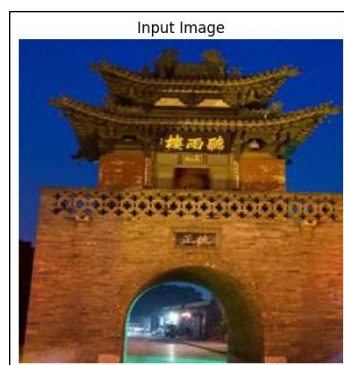


Figure 8: Input image.
Source: Authors, (2026).

An input image which serves as the initial sample in the image forgery detection procedure is depicted in Figure 8. The image depicts a conventional architectural structure with a brick base, archway, and a decorative roof. It depicts the raw input data for a forgery detection system prior to processing. This indicates the start of the analytical pipeline. The system uses such images to determine authenticity or modification. These inputs are critical for developing and testing detection systems.

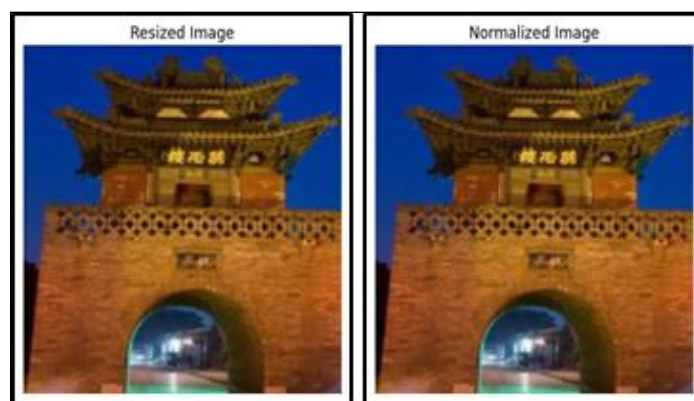


Figure 9: Resized and Normalized image.
Source: Authors, (2026).

Two image processing methods resizing and normalization are contrasted side by side in this figure 9. The Resized Image displays the outcome of adjusting the image's dimensions, which frequently results in distorted or blurry images if the aspect ratio is changed. The Normalized Image shows an image having its pixel values changed, perhaps for improved clarity and contrast, without changing the image's original size. In order to improve the image's visual balance and prepare its features for additional analysis normalization usually entails scaling the pixel values to a predetermined range.

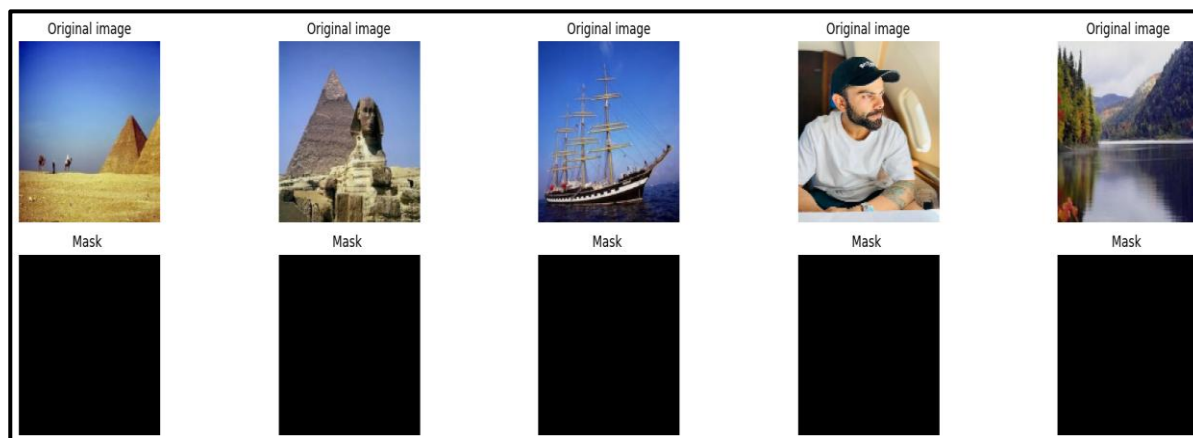


Figure 10: Image masking process.
Source: Authors, (2026).

Figure 10 displays a collection of original images along with matching black masks. The images depict locales such as the Pyramids of Giza, the Sphinx, a sailing ship, a man in a plane, and a lakeside setting. A black mask appears beneath each image, indicating that it was most likely utilized in image processing. These masks target specific areas for activities like as segmentation and object detection. The entirely black masks indicate that parts are masked for processing purposes. They utilized later to emphasize or preserve specific areas of the images during analysis.

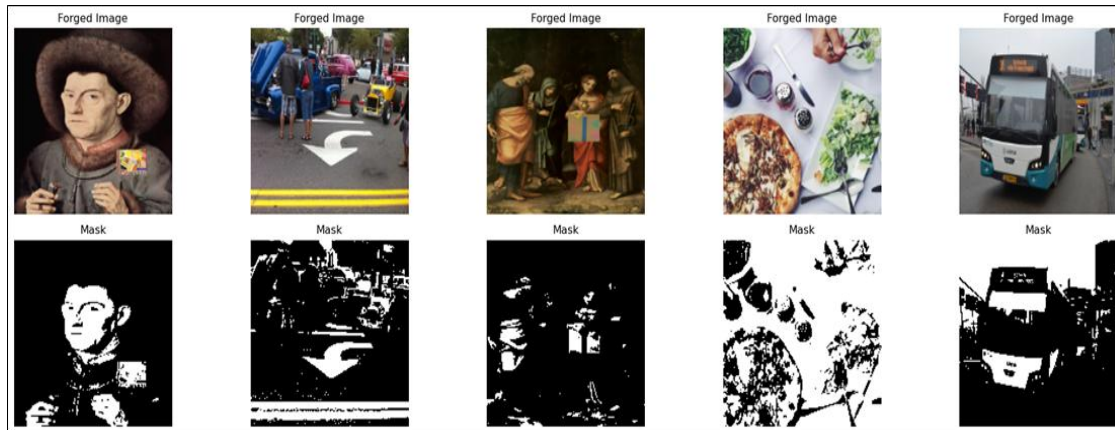


Figure 11: Forgery Detection with Masking Technique.
Source: Authors, (2026).

Figure 11 displays a series of manipulated images together with the black-and-white masks that go with them. This technique uses masks to identify altered or forged areas in images. Each mask focuses on manipulated regions, such as a digitally altered portrait, modifications in a street scene, a modified painting, a fabricated food arrangement, and edits on a bus. The masks draw attention to these altered areas. This is useful in forensic analysis, image validation, and forgery detection. The graphic shows how mask detection identifies modified parts in forged images.

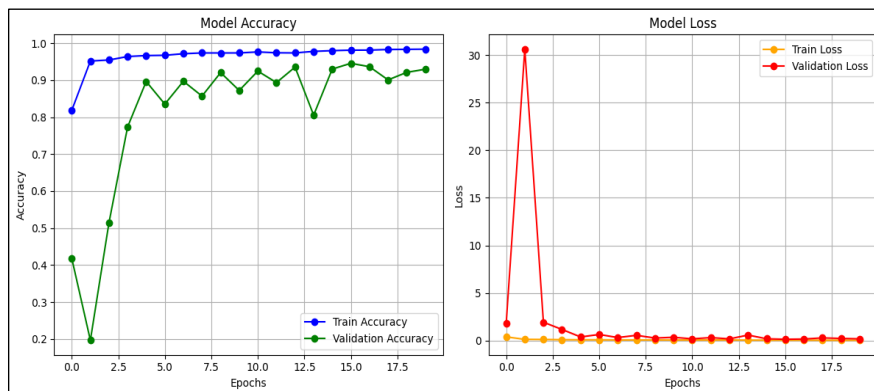


Figure 12: Model Accuracy and Loss over Epochs.
Source: Authors, (2026).

Figure 12 display a model's performance measures over 20 training epochs. The left figure depicts model accuracy, with training accuracy (blue) rapidly increasing and stabilizing near 1.0, whereas validation accuracy (green) improves slowly and varies, indicating some overfitting. The right plot depicts model loss, with both training (yellow) and validation loss (red) starting high but decreasing dramatically. Validation loss varies considerably, indicating potential instability or overfitting. Despite this, both losses stabilize near zero, indicating that the model has successfully trained.

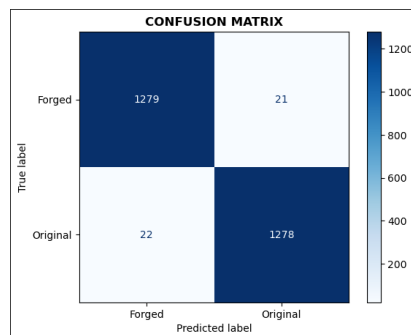


Figure 13: Confusion matrix.
Source: Authors, (2026).

The confusion matrix presented in Figure 13 shows a well classification model predicts two classes Forged and Original. The confusion matrix has 1,279 true positives for forged images and 1,278 true positives for original images. There are 21 false positives (original misclassified as forgery) and 22 false negatives (forged misclassified as original). The model is highly accurate, with minimal misclassifications. This demonstrates a great capacity to differentiate between the two classes successfully.

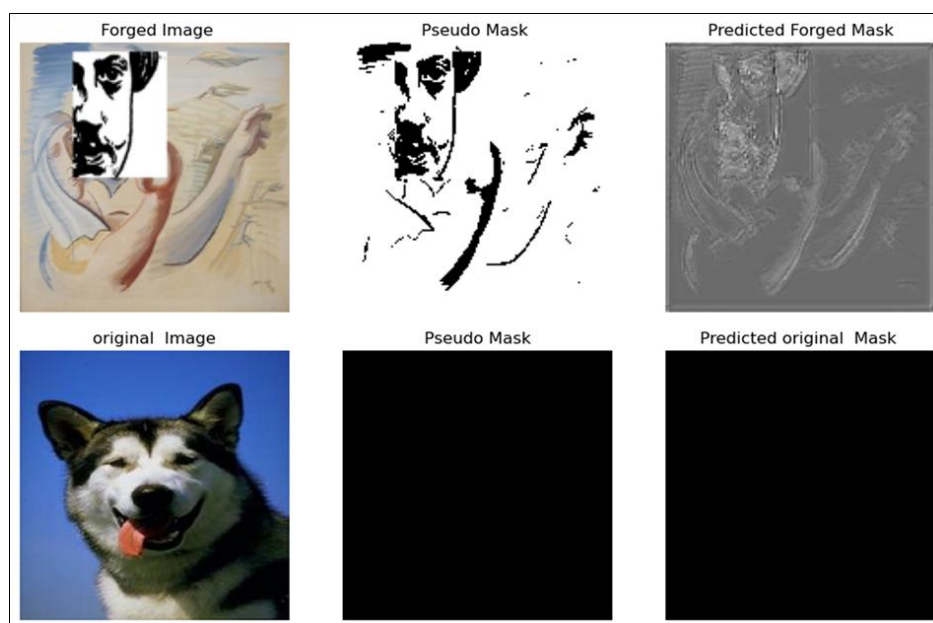


Figure 14: Image Forgery Detection and Mask Prediction Outcomes.
Source: Authors, (2026).

The process of identifying image forgery with mask prediction algorithms is depicted in Figure 14. The six sections demonstrate the forgery detection procedure. The top-left corner displays a faked image, followed by a faux mask that highlights probable modifications. The model predicts a forged mask in order to identify certain bogus regions. The original image and pseudo mask are shown here, with no modification. The accuracy of the detection model be evaluated by comparing anticipated masks to originals.

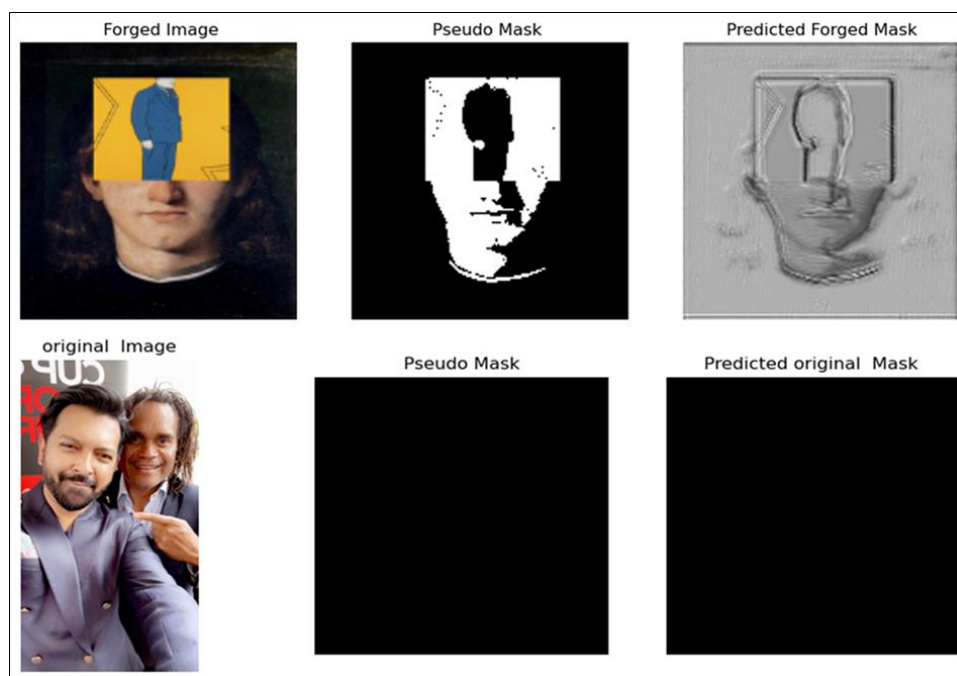


Figure 15: Forgery Detection in Images and Mask Prediction Results.
Source: Authors, (2026).

Figure 15 illustrates the process of image fraud detection, emphasizing the comparison of authentic and manipulated images using anticipated and faux masks. The top row displays the original image, its blank pseudo mask, and the forecasted mask, indicating no tampering. The bottom row shows a forged image, a faux mask that highlights the altered areas, and the model's anticipated forged mask. The anticipated mask attempts to match the pseudo mask's highlighted regions. This comparison measures the model's ability to reliably detect forgeries.

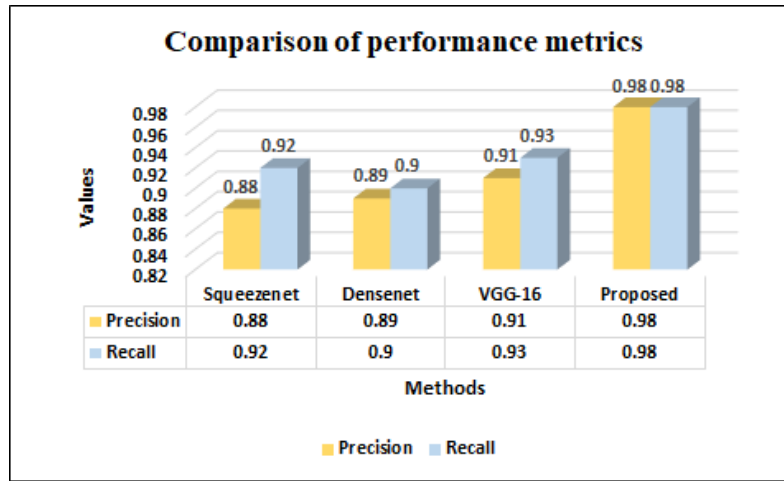


Figure 16: Comparison of Performance metrics.

Source: Authors, (2026).

Figure 16 illustrates the performance of four different methods such as SqueezeNet [18], DenseNet [19], VGG-16 [20], and a proposed method based on two metrics, Precision and Recall. SqueezeNet has the lowest performance, with a Precision of 0.88 and Recall of 0.92, DenseNet follows closely, showing slight improvements with a Precision of 0.89 and Recall of 0.92. VGG-16 performs better, with a Precision and Recall of 0.91. The proposed method significantly outperforms the others, achieving the highest scores Precision and Recall of 0.98 each.

Table 2: Comparative analysis of accuracy.

Methods	Accuracy
SqueezeNet [18]	0.89
Dense net [19]	0.92
VGG-16 [20]	0.93
Proposed	0.98

Source: Authors, (2026).

Table 2 shows the comparative analysis of accuracy for various methods such as SqueezeNet, Dense net, VGG-16 and a proposed method. In this comparison, SqueezeNet attains the lowest accuracy of 0.89, Dense net reaches the accuracy of 0.92, VGG-16 achieves the accuracy of 0.93 and the proposed method succeeds the accuracy of 0.98. Hence the proposed method outperforms the highest accuracy of overall methods.

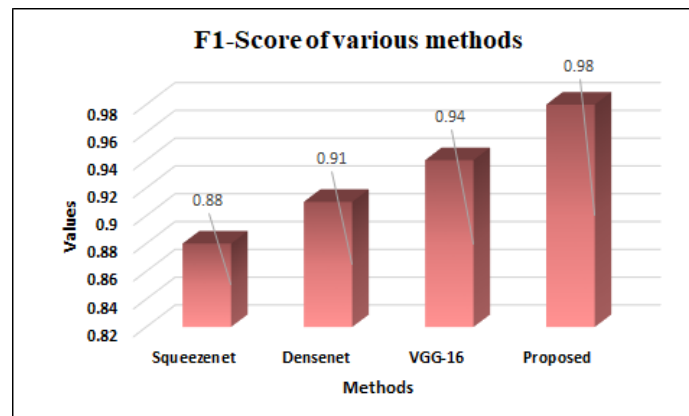


Figure 17: FI-score comparison of various methods.

Source: Authors, (2026).

Figure 17 depicts the FI- score of various methods such as Squeeze net, Dense net, VGG-16 and proposed method. Squeeze net achieves the lowest FI-score of 0.88, followed by Dense net and VGG-16 of 0.91 and 0.94. The proposed system achieves the highest FI-score of 0.98.

V. CONCLUSION

This study presents the contributions and most significant results for detecting image forgeries. The primary contribution lies in the introduction of an innovative hybrid DL model that seamlessly integrates residual-dense Swin transformer with attention based UNet++. Through the assistance of suitable techniques, it illustrates that the detection of image forgery is effectively processed through four stages. The procedure commences with input image where raw data is provided to the system for forgery detection. This is followed by preprocessing where resize and normalization takes place to ensure consistent input for the model, which helps in training and inference stability. In feature extraction, it extracts high level features from the input image using a hybrid model.

The extracted features are then segmented to produce a pixel-wise segmentation mask indicating forged images. A predicted pixel mask highlighting suspected forged regions in the image. Then the mask is subjected to a threshold by the detection module and if exceeded, it marks the image as forged. The model is implemented in Python and it outperforms higher accuracy, precision, recall and FI-score of 0.98%.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Sruthi Anand and Saranya V.

Methodology: Sruthi Anand and Saranya V.

Investigation: Sruthi Anand.

Discussion of results: Sruthi Anand and Saranya V.

Writing – Original Draft: Sruthi Anand.

Writing – Review and Editing: Sruthi Anand and Saranya V.

Resources: Saranya V.

Supervision: Saranya V.

Approval of the final text: Sruthi Anand and Saranya V.

VII. REFERENCES

- [1] M. A. Elaskily, H. A. Elnemr, A. Sedik, M. M. Dessouky, G. M. El Banby, O. A. Elshakankiry, A. A. Khalaf, H. K. Aslan, O. S. Faragallah, and F. E. Abd El-Samie, "A novel deep learning framework for copy-moveforgery detection in images," *Multimedia Tools and Applications*, vol. 79, no. 27, pp. 19167-19192, 2020.
- [2] S. R. Gadu, "A Hybrid Generative Adversarial network with Quantum U-NET for 3D spine X-ray image registration," *Healthcare Analytics*, vol. 4, pp. 100251, 2023.
- [3] H. Chen, C. Chang, Z. Shi, and Y. Lyu, "Hybrid features and semantic reinforcement network for image forgery detection," *Multimedia Systems*, vol. 28, no. 2, pp. 363-374, 2022.
- [4] F. M. Al_Azrak, A. Sedik, M. I. Dessowky, G. M. El Banby, A. A. Khalaf, A. S. Elkorany, and F. E. Abd. El-Samie, "An efficient method for image forgery detection based on trigonometric transforms and deep learning," *Multimedia Tools and Applications*, vol. 79, no. 25, pp. 18221-18243, 2020.
- [5] R. Karthickmanoj, S. Aasha Nandhini, D. Lakshmi, R. Rajasree, and T. Ananth Kumar. "Transforming deep learning-based skin disease detection and classification with AI-based diagnostic framework in the era of Industry 5.0," 2025.
- [6] Y. Liu, X. Li, J. Zhang, S. Li, S. Hu, and J. Lei, "Hierarchical Progressive Image Forgery Detection and Localization Method Based on UNet," *Big Data and Cognitive Computing*, vol. 8, no. 9, pp. 119, 2024.
- [7] H. Zhang, J. Zeng, and J. Yang, "Multi-Scale Edge-Guided Image Forgery Detection via Improved Self-Supervision and Self-Adversarial Training," *Electronics*, vol. 14, no. 9, pp. 1877, 2025.
- [8] K. Aishwaryameenakshi, S. H. Banu, A. K. Priya, and S. Chitrakala, "Image Indexing using Edgel Index for Sketch-based Image Search and Retrieval," *International Journal of Wisdom Based Computing*, vol. 1, no. 3, pp. 134-138, 2011.
- [9] D. P. Timothy, and A. K. Santra, "Detecting Digital Image Forgeries with Copy-Move and Splicing Image Analysis using Deep Learning Techniques," *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 5, 2024.
- [10] A. Diwan, and A. K. Roy, "Cnn-keypoint based two-stage hybrid approach for copy-move forgery detection," *IEEE Access*, vol. 12, pp. 43809-43826, 2024.
- [11] V. Balaji, R. J. Malar, K. S. Kavin, S. R. Rose, T. R. Permila, and B. Pandian, "PSO Optimized RBFNN Classifier for Lung Cancer Identification System," In *2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT)*, vol. 1, pp. 1189-1194, 2024.
- [12] J. Priyanka, A. A. Sundaresan, L. R. Kalabarige, and R. A. Venkata, "Advancements in image forgery detection through deep learning methods," In *AIP Conference Proceedings*, vol. 3291, no. 1, pp. 030030, 2025.
- [13] O. Aparna, P. Rani, T. Ramya, T. Priyanka, N. Sundari, P. G. K. Sirisha, R. Ramesh, and D. Anand, "Detection of Frauds in Deep Fake Using Deep Learning," *Engineering Proceedings*, vol. 66, no. 1, pp. 48, 2024.
- [14] G. E. Lio, M. D. L. Bruno, F. Riboli, S. Nocentini, and A. Ferraro, "Fast and Robust Speckle Pattern Authentication by Scale Invariant Feature Transform algorithm in Physical Unclonable Functions," *arXiv preprint arXiv:2504.21041*, 2025.
- [15] I. Sedeeq, "Image Forgery Detection Using Histogram-Oriented Gradients (HOG)," *Iraqi Journal of Science*, pp. 2048-2058, 2025.
- [16] S. Bokefode, J. Sarwade, K. Sakure, S. Bankar, S. Janrao, and R. Patil, "Using a Clustering Algorithm and a Transform Function, Identify Forged Images," *International Research Journal of Multidisciplinary Studies*, vol. 5, no. 1, pp. 781-789, 2024.
- [17] Y. Chen, M. C. Chang, and X. Li, "Leveraging Semantic Segmentation for Image Manipulation Detection and Localization," In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 95-101, 2024.
- [18] I. T. Ahmed, F. A. Abdulazeez, and B. T. Hammad, "Employing Different Algorithms of Lightweight Convolutional Neural Network Models in Image Distortion Classification," *Computers*, vol. 13, no. 10, pp. 268, 2024.
- [19] A. Alzahrani, "Digital image forensics: An improved DenseNet architecture for forged image detection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13671-13680, 2024.
- [20] R. R. Choudhary, S. Paliwal, and G. Meena, "Image forgery detection system using VGG16 UNET model," *Procedia Computer Science*, vol. 235, pp. 735-744, 2024.