International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: *25 April 2026* ***Revised:*** *12 May 2026* ***Accepted:*** *30 June 2026*

TNVS-Net with Context-Aware Relation Network for Robust Face Forgery Detection across Varying Image Qualities

Dr.J.Suguna (Retd)

Associate Professor Department of Computer Science

Vellalar College for Women, Erode. sugunajravi@gmail.com

Mrs.M.Ramya

Assistant Professor Department of Computer Science

Vellalar College for Women, Erode. ramyasivanathan@gmail.com

ABSTRACT

Face forgery is the manipulation or synthetic generation of facial images or videos to falsely represent a person's identity, expressions, or actions. Deep Learning (DL) techniques have been extensively utilized for both generating and detecting such forged media. Several DL-based face forgery detection models have achieved significant performance on high-quality facial images and videos. However, in real-world scenarios, facial images are often compressed by social media platforms, which reduce image quality, obscure fine visual details, and introduces artifacts and noise that negatively affect detection accuracy. To overcome these challenges, a Multi-Scale Dual-Branch Network (MDCF-Net) was introduced, which integrates multi-scale Red Green Blue (RGB) features and frequency domain information for effective forgery detection. The extracted features are processed through a Spatial Pyramid Pooling (SPP) layer to generate fixed-length feature representations, followed by a softmax classifier for final prediction. Although MDCF-Net performs effectively on highly compressed facial images, its performance is limited when handling low-compressed or uncompressed datasets. Therefore, this paper proposes a Context-Aware Relation Network (CARN) branch integrated with the RGB domain branch and frequency domain branch of MDCF-Net. The proposed architecture, named Three-Node Variable-Scale Network (TNVS-Net), is designed to detect forged facial images across highly compressed, low-compressed, and uncompressed images. Experimental evaluation on three benchmark datasets demonstrates that the proposed TNVS-Net achieves superior performance compared to existing classical face forgery detection models.

Keywords: Face Forgery Detection, Deep Learning, Multi-Scale RGB Features, Frequency Domain Analysis, Context-Aware Relation Network, TNVS-Net

I. INTRODUCTION

Forgery refers to the deliberate creation, alteration, or imitation of documents, signatures, images, or digital content to deceive others. It is a significant cybercrime associated with financial fraud, identity theft, and misinformation. Among various forms of forgery, face forgery has emerged as a major digital threat, where facial images or videos are manipulated using techniques such as face swapping, facial reenactment, and synthetic face generation [1-4] to produce highly realistic fake media that is difficult to distinguish from authentic content.

The rapid advancement of Deep Learning (DL) has significantly improved face forgery generation using architectures such as Generative Adversarial Networks (GANs) and transformer-based models [5], [6]. Although deepfake technology has useful applications in entertainment and media production, it is increasingly misused for misinformation, fraud, and identity impersonation [7], [8]. Existing deepfake detection methods commonly employ Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and frequency-domain analysis to identify forgery artifacts; however, many models demonstrate limited robustness and poor generalization on compressed and unseen facial images [9]–[13]. Recent studies further indicate that combining RGB spatial features with frequency-domain representations and attention mechanisms can enhance forgery

detection performance

[14]–[18]. Motivated by these limitations, this paper proposes a Three-Node Variable-Scale Network (TNVS-Net) integrated with a Context-Aware Relation Network (CARN) branch, which jointly learns spatial, spectral, and contextual forgery representations for robust face forgery detection across highly compressed, low-compressed, and uncompressed facial images.

II. LITERATURE SURVEY

Despite significant advancements in deepfake detection, existing methods still face challenges such as poor generalization, reduced performance on compressed and uncompressed datasets, and high computational complexity. Most approaches independently utilize spatial or frequency-domain features, limiting their ability to effectively capture contextual forgery patterns. The following related works address these challenges.

Tran et al. [19] proposed a Meta Deepfake Detection (MDD) framework that leverages meta-optimization techniques to improve generalization across unseen deepfake domains. The approach is designed to learn transferable initialization parameters; however, it still exhibits a limitation in adapting dynamically during the testing phase, which restricts its robustness in real-world scenarios where data distributions continuously evolve.

Sun et al. [20] introduced the Fake Trajectory Detection Network (FTDN), which performs spatiotemporal analysis of facial forgery by modeling temporal inconsistencies across video frames. Although the method effectively captures motion-related artifacts, the inclusion of multiple dense fully connected layers significantly increases computational overhead, making the model less efficient for real-time or resource-constrained applications.

Alnaim et al. [21] utilized an Inception-ResNet-v2 architecture for deepfake detection using the DFFMD dataset. While the model benefits from deep hierarchical feature extraction, its performance remains limited due to insufficient training diversity and the inherent complexity of the dataset, leading to comparatively lower classification accuracy.

Ramadhani et al. [22] proposed an IViViT-based framework that integrates facial landmark information with attention mechanisms to enhance forgery detection. However, the heavy reliance on landmark-guided regions reduces the ability of the model to capture global spatial consistency, thereby affecting overall spatial coherence in detection results.

Relation Network (CARN) jointly learns spatial, frequency, and contextual forgery features for robust face forgery detection under diverse real-world conditions.

III. PROPOSED METHODOLOGY

The proposed TNVS-Net is designed to detect face forgeries in highly compressed, low-compressed, and uncompressed facial images. The architecture consists of three major branches: a multi-scale RGB branch for extracting fine-texture spatial features using Transformers, a frequency filter branch with LFS for capturing spectral forgery artifacts, and a Context-Aware Relation Network (CARN) branch for learning contextual dependencies and semantic forgery patterns among facial regions. These complementary branches improve robust and generalized deepfake detection across varying image qualities.

3.1 Shallow and Deep Features

The proposed model employs Image Net pre-trained EfficientNet-B4 as the backbone network for efficient facial feature extraction. Given an input facial image $X \in \mathbb{R}^{H \times W \times C}$, where *H, W,* and *C* denote the image height, width, and number of channels, respectively. The network extracts shallow texture and deep semantic features from the pre-processed image. The resulting shallow feature map is $H \quad W$ Zhou et al. [23] developed the IIN-FTMM model to jointly learn forgery-specific and shared representations across datasets. Despite this design, significant variations in feature distributions between datasets hinder the model's ability to generalize effectively, leading to performance degradation under cross-dataset evaluation settings.

Soudy et al. [24] combined Convolutional Neural Networks (CNNs) with Vision Transformers (ViT) to leverage both local texture features and global contextual information for deepfake detection. However, the framework predominantly focuses on limited facial regions, which restricts its ability to capture holistic inconsistencies present in entire frames.

Zhou, J., Zhao, X., Xu, Q., Zhang, P., and Zhou, Z. (2024).[25] presented MDCF-Net, a Multi-Scale Dual-Branch Network for compressed face forgery detection. The paper mainly focused on detecting

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: 25 April 2026 *Revised:* 12 May 2026 *Accepted:* 30 June 2026

forged facial images and videos under compressed conditions. The network extracts spatial and frequency-domain features using dual branches and applies multi-scale feature fusion for effective forgery detection. The model achieved good detection accuracy, particularly for compressed images and videos.

Although recent approaches have improved through transformers, attention mechanisms, and frequency-domain learning, many methods still struggle to maintain robust performance across diverse compression conditions. Furthermore, the lack of effective integration of spatial, spectral, and contextual representations limits detection robustness and generalization. To address these limitations, the proposed TNVS-Net integrated with a Context-Aware represented as $T \in \mathbb{R}^{(4)\times(4)\times C}$, which is forwarded to the Multi-scale RGB, Frequency Filter, and CARN branches for spatial, spectral, and contextual forgery analysis [9], [10], [23], [24].

3.2 Multi-scale RGB Branch

The Multi-scale RGB Branch is designed based on the transformer self-attention mechanism [17], [18] to capture fine-grained contextual forgery features at multiple spatial resolutions. The input to this module is the shallow texture

$$\frac{H}{} \quad \frac{W}{}$$

feature map $T \in \mathbb{R}^{(4)\times(4)\times C}$, extracted from the

EfficientNet-B4 backbone. To obtain multi-scale contextual representations, the feature map is divided into patches at four scales: original, 1/2, 1/4, and 1/8. The patch representation at each scale is defined as $C_h = r_h \times r_h \times C$, where r_h denotes the spatial resolution of the patch at scale *h*.

Each patch sequence is independently processed through transformer self-attention layers to learn contextual semantic dependencies among facial regions. The contextual dependency weight between image patches is computed using the softmax-based attention mechanism [17] as Eq. (1).

$$p^h = Softmax\left(\frac{x \; x}{}\right) \qquad (1)$$

x $\sqrt{C_h}$

The contextual feature representation of each patch is then obtained by weighted aggregation of the value embeddings [18] as

h N

i j=1

h h

ij j(2)

The filtered spectrum is transformed back into the spatial domain using inverse FFT to generate the final frequency

The multi-scale contextual features extracted from

different resolutions are reshaped, concatenated, and transformed back into the original feature representation to generate the final RGB feature tensor represented as $f_{rgb} \in$

$H \; W$ _ _

$\mathbb{R}^{4 \times 4 \times C}$. This branch enhances contextual semantic learning by capturing forgery traces at multiple spatial resolutions, improving detection performance across compressed and uncompressed facial images.

Fig. 1 illustrates the overall architecture of the proposed TNVS-Net, which integrates the Multi-scale RGB branch, Frequency Filter branch, and Context-Aware Relation Network (CARN) branch for extracting spatial, spectral, and contextual forgery features from facial images.

feature map f_{freq} [11].

$$f_{freq} = \mathcal{Y}^{-1}(F) \qquad (5)$$

Where, $\mathcal{Y}^{-1}(F)$ denotes the inverse FFT. This process helps the model capture long-range frequency dependencies while preserving local spatial details for improved forgery detection.

As illustrated in Fig. 2, the Learnable Frequency Selection (LFS) module extracts discriminative frequency information using sliding windows over the frequency feature map [11,28]. For each local window *q*, the frequency statistics are computed as Eq. (6).

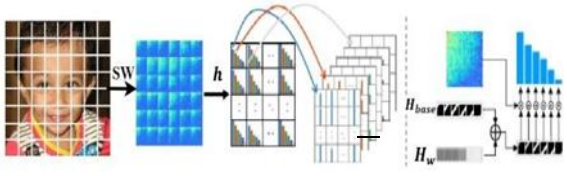

Fig. 2 LFS Module

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: 25 April 2026 *Revised:* 12 May 2026 *Accepted:* 30 June 2026

$\overline{u_x} = log_{10}\|D(q) \odot [H^x + \phi(H^x)]_1, x = \{1, \ldots . . , T\}$ (6)

$= base \quad w$

Where, $H^x$and H^x denote the base and learnable$basew$

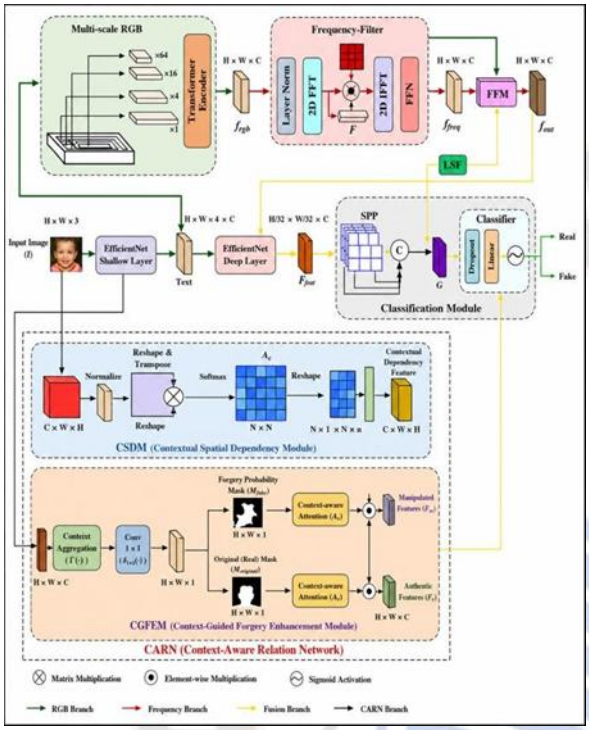

Fig. 1 Schematic Representation of the proposed model

3.3 Frequency-filter Domain

The spatial features determined after transformer encoding, filters for the x^{th} frequency band. The extracted local frequency features are reshaped and passed to EfficientNet-B4 for high-level forgery feature extraction. The LFS module processes a 299×299×3 image using 10×10 sliding windows with stride 2 and divides each frequency spectrum into 6 bands, generating compact localized frequency representations for effective forgery detection.

3.4 Feature Fusion Module

The Feature Fusion Module (FFM) [10] receives the RGB feature map f_{rgb} and frequency feature map f_{freq}. The module employs cross-modal multi-head attention to effectively fuse spatial and frequency-domain features. Initially, 1×1 convolution layers generate queries Q from RGB features and keys K and values V from frequency features. Each attention head computes feature interactions as defined by Eq. (7)

$$Q_x(f_{rgb})K_x(f_{freq})^T$$

the spatial feature map f_{rgb} is converted into the frequency

$$E_x = softmax\,(\sqrt{dk_x})\,V_x(f_{freq}) \qquad (7)$$

domain using a 2D Fast Fourier Transform (2D-FFT) to extract spectral forgery information [11].

$$H \quad W$$

$$F = y(f_{rgb}) \in \mathbb{R}^{4 \times 4 \times C} \qquad (3)$$

The obtained frequency representation is multiplied element-wise with a learnable weight matrix W for adaptive frequency selection [11].

$$F = W \odot F \qquad (4)$$

Where, $Q_x(f_{rgb}), K_x(f_{freq}) and\ V_x\,(f_{freq})$ denote the query, key, and value representations for the x[th] attention head. The module uses 3 attention heads to capture diverse feature relationships between RGB and frequency domains.

The outputs from all attention heads are concatenated to form the fused feature representation in Eq. (8).

$$M = Concatenate\,(M_1, M_2, \ldots \ldots, M_N) \qquad (8)$$

The concatenated features are passed through a convolution layer and added to the original RGB features

through a residual connection to obtain the final fused feature map in Eq. (9).

$$F_{out} = f_{rgb} + Conv(M) \qquad (9)$$

structure, instead of mapping the contextual dependency representation into feature space via fully connected layers (which could distort spatial information), the x^{th} row is reshaped into a two-dimensional spatial map $CD^{row} \in$

^{x}

Here, $Conv(M)$ consists of a 1×1 convolution layer, batch normalization, and Leaky ReLU activation. The fused feature map F_{out} is then forwarded to deeper convolutional layers for extracting high-level forgery representations.

3.5 Context-Aware Relation Network (CARN) Branch

The proposed Context-Aware Relation Network (CARN) captures contextual dependencies and inter-region relationships to enhance facial forgery

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: *25 April 2026* ***Revised:*** *12 May 2026* ***Accepted:*** *30 June 2026*

detection. Unlike conventional methods that separately process spatial and frequency features, CARN learns context-aware representations across facial regions, improving robustness under highly compressed, low-compressed, and uncompressed conditions. The extracted contextual features are fused with RGB and frequency-domain features to form a robust representation for classification, thereby improving accuracy and generalization across diverse datasets.

To model these dependencies effectively, two multi-scale modules are introduced: the Contextual Spatial Dependency Module (CSDM) and the Context-Guided Forgery Enhancement Module (CGFEM), which learn context-aware and forgery-sensitive spatial representations.

3.5.1 Contextual Spatial Dependency Module (CSDM)

The Contextual Spatial Dependency Module (CSDM) is designed to capture contextual dependencies and semantic interactions among spatial feature representations before performing pixel-wise classification. Similar contextual attention mechanisms have been effectively utilized [17],

[18] in transformer-based feature learning and multi-scale vision models. Given an input feature map $F_x \in \mathbb{R}^{C\times H\times W}$, where C, H and W denote the number of channels, height, and width respectively, the spatial feature representations are first normalized. The normalized feature map is then

reshaped to $Z \in \mathbb{R}^{C\times N}$, where, $N=H\times W$ represents the total number of spatial feature locations in the feature maps. To model contextual feature dependencies, a contextual dependency map $CD \in \mathbb{R}^{N\times N}$ is computed by multiplying Z and Z^T, where T represents the transpose. This results in a matrix that encodes contextual dependencies among semantically correlated spatial features. The resulting contextual dependency map is then normalized along its second dimension using the softmax function, as described by Eq. (10).

$\mathbb{R}^{H\times W}$.

Given that input images typically feature face-centered and structurally consistent content, a lower-resolution representation of contextual spatial dependencies is adequate for effective face forgery detection, while also helping to reduce computational overhead. The Contextual Dependency Feature $F^x \in \mathbb{R}^{C\times H\times W}$ for the x^{th} pixel is then calculated by applying average pooling and a few convolutional layers to CD^{row}. The final feature map is updated using Eq. (11).

$$F_{ud} = I(f_{1\times1}(F_{input})F^x \quad) \qquad (11)$$
$$SRel$$

Where, $f_{1\times1}$ is a 1 × 1 convolution used to integrate features across channels, and I represents the integration operation between the original spatial representations and the context-enhanced feature representations.

This module captures contextual spatial dependencies by computing adaptive contextual interactions among semantically correlated facial regions, enabling the model to identify subtle forgery inconsistencies more effectively.

3.5.1 Context-Guided Forgery Enhancement Module

The feature map F_{ud} from the CSDM is given as input to the CGFEM along the shallow and deep features to aggregate representations from different facial regions using a spatial attention mechanism [18]. A fake mask is computed $M_{fake} \in \mathbb{R}^{C\times H\times W}$, by performing contextual forgery probability estimation on F_{ud}. Specifically, F_{ud} which is fed into 1 × 1 convolution layer and an adjacent activation layer as in Eq. (12).

$$M_{fake} = Sigmoid\ (\delta_{1\times1}(\Gamma(\mathrm{F_{ud}}))) \qquad (12)$$

Where *Sigmoid* (.) denotes the sigmoid activation function, Γ (.) denotes the contextual feature aggregation operation, and δ1×1 represents the 1×1 convolution projection layer. The sigmoid function maps each spatial response value into the range [0,1], representing the probability that the corresponding spatial region contains manipulated content. Based on the M_{fake} , the CARN framework effectively enhances manipulated contextual region representations. Instead of utilizing M_{fake} as the attention map, the context-aware attention map A_C is generated for manipulated facial regions based on the M_{fake} with normalization by Eq. (13).

$CD_{=}$ exp(ϕ$(Z_x)^T$×ψ(Z_y))(10)

exp(M_{fake})

N

y=1

exp(ϕ$(Z_x)^T$×ψ(Z_y))A_C $\overset{SRel}{=}$ N

(13) x

$\exp(M \quad)$

Where, $Z^T_x \in \mathbb{R}^{C\times1}$, denotes the feature vector of the x^{th} pixel, $CD_{x,y}$ represents the contextual dependency between the i^{th} and j^{th} spatial feature representations, and ϕ and ψ are contextual projection functions. To preserve the spatial In A_C, the value of each spatial location is positively correlated with the likelihood of facial manipulation, enabling the model to focus on contextually significant manipulated regions. Additionally, A_C provides a

significant advantage compared to M_{fake}. When the values in M_{fake} are close to zero, directly weighting the feature representations using Mfake may excessively suppress contextual feature responses. In contrast, A_C adaptively assigns higher attention weights to regions more likely to contain forgery-related contextual inconsistencies, regardless of the absolute probability values. Finally, the manipulated contextual representation F_m is computed by, The procedural steps of the proposed TNVS-Net algorithm are illustrated through the architectural flow presented in Fig. 1.

Alogrithm: TNVS-Net for detecting highly compressed, low or uncompressed face forgeries.

Input: Collected facial images that consist of real and fake face images *Output:* Detecting face forgeries from all compressed

$$1 \quad N \quad x \qquad x\pi F = -\sum_N Ax = ① \, Fc \qquad ud(14)$$

scenarios (Highly compressed, low or uncompressed)

1. Divide the input image into patches of size

In addition, the contextual representation of authentic facial regions F_{real} and the corresponding context-aware attention map A_{real} are determined in a similar manner, where the original mask $M_{original}$ is employed instead of

M_{fake}.

The CGFEM improves forgery detection by applying context-guided attention to emphasize manipulated and authentic regions, similar to attention-guided forgery localization approaches proposed in recent deepfake detection studies [9]. It generates normalized attention maps from the forgery probability mask to highlight informative forged areas while suppressing irrelevant context, enabling better discrimination of subtle artifacts and improving robustness across different compression levels and manipulation types.

3.6 Classification layer

To capture multi-scale contextual information, a Spatial Pyramid Pooling (SPP) Layer [27] is employed after the entire feature extraction and fusion process in the network. It pools the deep feature maps F_{feat} into containers of different sizes, concatenating the resulting vectors to form a fixed-length output suitable for subsequent fully connected layers and classifiers. This is crucial for maintaining multi-scale contextual spatial representations between features at different scales. An SPP layer with L levels [27] can be defined as,

2. Compute proximity between image patches using self-attention using Eq. (1)

3. Determine output of query image patch by weighted sum of value embeddings by Eq. (2)

4. Reshape and concatenate outputs from multiple patch scales into a unified representation.

5. Apply 2D-FFT on texture map to obtain frequency domain representation using Eq. (3)

6. Apply learnable frequency filter via element-wise multiplication by Eq. (4)

7. Apply inverse FFT to obtain spatial-domain frequency features using Eq. (5)

8. Compute LFS in defined frequency bands using Eq. (6)

9. Fuse RGB and frequency features using cross-modal multi-head attention using Eq. (7)

10. Concatenate outputs of attention heads as per Eq. (8)

11. Apply 1×1 convolution and residual connection to obtain final fused feature by Eq. (9)

12. Compute contextual dependency map using contextual feature interactions and softmax normalization as in Eq. (10)

13. Integrate contextual dependency features with the input feature map using convolution as in Eq. (11)

14. Compute forgery probability mask for manipulated regions using sigmoid activation as in Eq. (12)

15. Generate normalized context-aware attention map for manipulated regions using Eq. (13)

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: 25 April 2026 Revised: 12 May 2026 Accepted: 30 June 2026

16. Compute contextual manipulated feature representations

$SPPLayer\ (F) = \oplus^{L}$ pool (F) (15)

using the context-aware attention map as in Eq. (14)

17. pply Spatial Pyramid Pooling (SPP) on extracted

The contextual feature representations extracted (F_{fake} , F_{real}) from the CARN branch, F_{feat} from SPP layer and LSF derived F_{freq} are fed into the G of SPP which adaptively fuses the multi-scale feature representations into a unified feature vector at different level as by Eq. (16).

$SPP(G) = SPP\ (\emptyset(F_{feat} \oplus F_{freq} \oplus F_{real} \oplus F_{fake}))$ (16)

Subsequently, the classifier performs binary classification (real or fake) using a Dropout layer, a fully connected layer, and a softmax activation function for effective forged face detection. The model remains robust under both compressed and uncompressed conditions, enabling reliable deepfake detection in real-world scenarios.features by Eq. (15)

18. Concatenate contextual, frequency-domain, and multi-scale feature representations and feed them into the classifier as in Eq. (16)

19. Classifier outputs final prediction: Real or Fake face

IV. RESULT AND DISCUSSION

4.1 Dataset Description

The collected dataset is originally in video format. For the experimental evaluation of the proposed method, the videos are converted into individual image frames to enable frame-level analysis, facilitate the extraction of spatial features, and support the use of image-based deep learning models for face forgery detection. Table 1 summarizes the datasets used for evaluating the proposed TNVS-Net model.

Table 1 Dataset Description

Dataset	*Key Characteristics*	*Manipulation Methods*	*Size & Sources*
FaceForensics++ [13]	Multiple compression levels with high-quality manipulations	(DF), (F2F), (FS), (NT)	15,000 videos from 977 YouTube sources
Celeb-DF (v2) [26]	Realistic DeepFakes with diverse demographics	(DF)	590 original + 5,639 fake videos from YouTube
WildDeepfake (DF-W) [27]	Real-world and unconstrained DeepFake scenarios	(DF)	7,314 face sequences from 707 online videos

TN (True Negative): Correctly predicted authentic images

- FP (False Positive): Authentic images incorrectly predicted as forged
- FN (False Negative): Forged images incorrectly predicted as authentic

Area Under the Curve (AUC):

AUC is a performance metric that measures a model's ability to distinguish between two classes (e.g., PPD vs. Non-PPD) across all possible classification thresholds. AUC is computed using the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR)

4.2 Experimental Settings

To evaluate performance improvements, both existing and

$TPR = \frac{TP}{TP+FN}$, $FPR = \frac{FP}{FP+TN}$

proposed models are implemented using Python on three different datasets. Each video is split into image frames, and the dataset is divided into 80% for training and 20% for testing which is depicted in table 2.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: *25 April 2026* ***Revised:*** *12 May 2026* ***Accepted:*** *30 June 2026*

Table 2 Data split of 80% and 20%

Dataset	*Train (Real)*	*Test (Real)*	*Train (Fake)*	*Test (Fake)*
FaceForensics ++	8,000	2,000	16,000	4,000
Celeb-DF	3,556	890	22,556	5,639
DF-W	13,480	3,370	13,592	3,397
Total	*25,036*	*6,260*	*52,148*	*13,036*

Each dataset provides a balanced mix of authentic and manipulated samples to ensure robust model evaluation.

Table 3 outlines the key hyperparameters selected for training the TNVS (Transformer-based Network for Visual Spoofing) model, which are optimized to enhance the accuracy and stability of face forgery detection during the training process.

Table 3 Hyper parameters for TNVS

Parameters	*Optimal Range*
Training rate	0.01
Dropout rate	0.5
Number of epochs	200
Batch size	256
Optimizer	Adam
Loss	Cross-entropy

4.3 Performance Evaluation Metrics

The model's performance in Automated Spoof Detection (ASD) is evaluated using four key metrics namely Accuracy, Precision, Recall, and F1-Score [28].

These parameters are used to evaluate the classification performance of forged and authentic facial images.

- TP (True Positive): Correctly predicted forged images
- AUC score is obtained by integrating the ROC curve:

$$AUC = \int_0^1 TPR(FPR)d(FPR)$$

4.4 Training Vs. Testing Accuracy

An epoch in machine learning represents one full pass through the training dataset. As the number of epochs increases, the model typically improves in learning patterns and generalizing to new data. This chart compares training and testing accuracy at Epochs 20 and 100 across three datasets—FaceForensics++, Celeb-DF, and DF-W.

Epoch 20: The model has only seen the data 20 times, so it's still in an early learning phase. Accuracy is relatively low (e.g., 46% training and 22% testing on FaceForensics++).

Epoch 100: After 100 passes through the data, the model has learned significantly more. Accuracy is much higher (e.g., ~99% for both training and testing), indicating better performance and generalization.

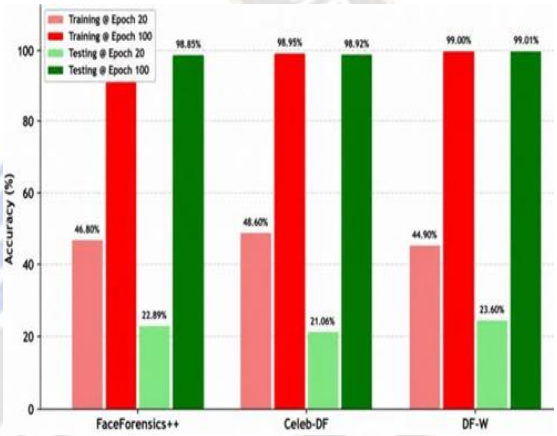

Fig. 3 Training vs Testing Accuracy at 20 and 100 across Datasets

Fig. 3 illustrates the training and testing accuracy of the proposed TNVS-Net on FaceForensics++, Celeb-DF, and DF-W datasets at epochs 20 and 100. The model shows significant improvement in both training and testing performance across all datasets, achieving nearly 99%

accuracy at epoch 100. The integration of the CARN branch enhances contextual dependency learning and semantic forgery representation, enabling robust generalization across compressed and uncompressed facial datasets.

4.5 Performance Analysis

The proposed TNVS-Net is compared with VGG19, MesoNet, EfficientNet-B4, and XceptionNet using Accuracy, Precision, Recall, F1-Score, and AUC on the FaceForensics++, Celeb-DF, and DF-W datasets. Fig. 4(a)–

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: *25 April 2026* ***Revised:*** *12 May 2026* ***Accepted:*** *30 June 2026*

(c) present the AUC analysis of the proposed and existing models. The integration of the Context-Aware Relation Network (CARN) enhances contextual dependency learning and semantic forgery localization, enabling effective discrimination between real and forged facial images. Experimental results demonstrate that TNVS-Net consistently outperforms existing models with higher AUC values, improved robustness, and better generalization under diverse real-world deepfake scenarios.

The ROC curves show the relationship between True Positive Rate (TPR) and False Positive Rate (FPR). TNVS-Net achieves higher AUC values and curves closer to the top-left corner, demonstrating superior capability in distinguishing forged and authentic facial images through the integration of RGB, frequency-domain, and context-aware semantic features.

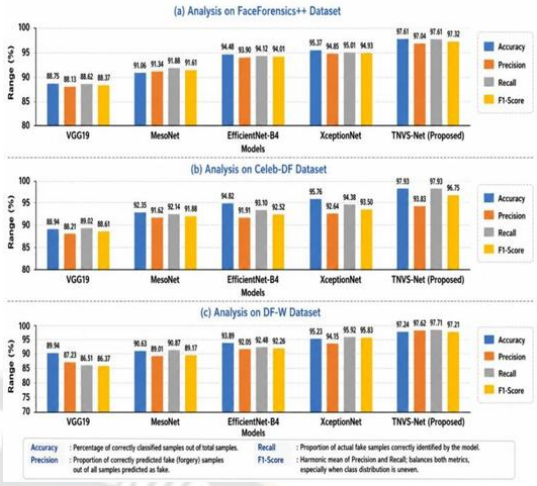

Fig.5. Performance analysis of the proposed TNVS-Net with existing models in terms of accuracy, precision, recall, and F1-score.

4.6 Computational Metrics

The TNVS-Net model demonstrates excellent computational efficiency alongside high detection accuracy across all datasets

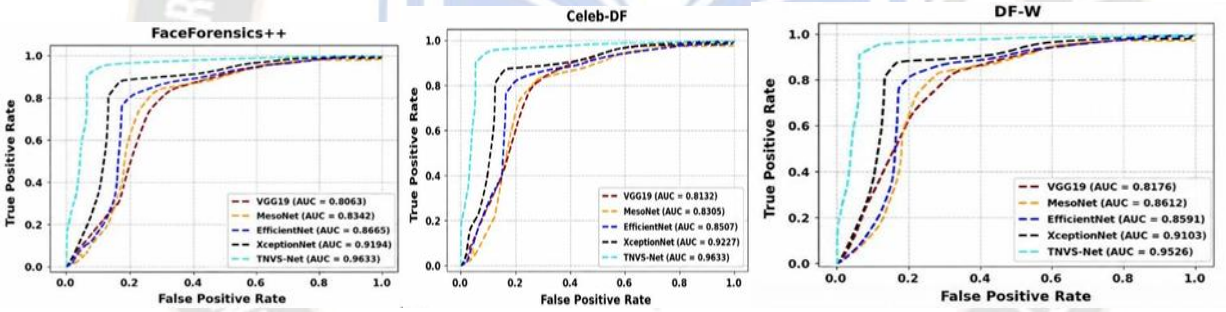

Fig. 4 (a)-(c) AUC analysis for proposed TNVS-Net and existing models

Table 4. Computational efficiency of the proposed model

Datasets	**Training time (sec)**	**Testing time (sec)**	**Detection Accuracy (%)**
FaceForensics ++	1420	85	98.85
Celeb-DF	1735	92	98.92
DF-W	1260	78	99.01

DF-W achieved the highest accuracy with the lowest computational cost, while Celeb-DF required the most training time, likely due to its complexity. Overall, TNVS-Net effectively balances detection accuracy and computational performance across varying datasets. Although the proposed TNVS-Net includes an additional CARN branch, the model maintains efficient computational performance while significantly improving detection robustness.

5. CONCLUSION

In this paper, TNVS-Net is proposed for detecting compressed, low or uncompressed face forgeries. Initially, the pre-trained EfficientNet-B4 backbone is used to extract the shallow and deep features from the input facial images. A multi-scale RGB branch utilizes a transformer encoder to capture hierarchical texture representations, while a frequency filter branch applies a 2D-FFT and LFS to expose subtle spectral anomalies indicative of forgery. These spatial and frequency features are fused using a FFM that combines multi-head attention and convolutional refinement to

enhance complementarity. To further improve spatial consistency, CARN is introduced, comprising a CSDM to model contextual semantic dependencies among spatial feature representations and a CGFEM to localize manipulated regions using an adaptive context-aware attention mask. Finally, a SPP layer aggregates multi-scale contextual features into a fixed-length vector, which is classified via fully connected layers and a softmax activation. Experimental results demonstrate the model's superior accuracy compared to classical face forgery detection methods.

REFERENCES

1. Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., Zhai, G., Yang, J., Shen, C., and Tao, D., "Deepfake Generation and Detection: A Benchmark and Survey," *arXiv preprint arXiv:2403.17881*, 2024.
2. He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., & Liu, Z. (2021). ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4360–4369).
3. Zhuang, W., Chu, Q., Yuan, H., Miao, C., Liu, B., & Yu, N. (2022). Towards intrinsic common discriminative features learning for face forgery detection using adversarial learning. In 2022 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1–6). IEEE.
4. Guo, Z., Liu, Y., Zhang, J., Zheng, H., & Shan, S. (2024). Face forgery detection with elaborate backbone. arXiv preprint arXiv:2409.16945.
5. Mirsky, Y., and Lee, W., "The Creation and Detection of Deepfakes: A Survey," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–41, 2021.
6. Kumar, M., & Sharma, H. K. (2023). A GAN-based model of deepfake detection in social media. Procedia Computer Science, 218, 2153–2162.
7. Nguyen, V. D., Mejri, N., Singh, I. P., Kuleshova, P., Astrid, M., Kacem, A., Ghorbel, E., and Aouada, D., "LAA-Net: Localized Artifact Attention Network for Quality-Agnostic and Generalizable Deepfake Detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 17395–17405.
8. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., and Wei, Y., "Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Learning," *arXiv preprint arXiv:2403.07240*, 2024.
9. R. Yang, K. You, C. Pang, X. Luo, and R. Lan, "CSTAN: A Deepfake Detection Network with CST Attention for Superior Generalization", Sensors, vol. 24, no. 22, p. 7101, 2024.
10. X. Li, Y. Wang, Z. Zhang, and H. Chen, "Refining Localized Attention Features with Multi-Scale Relationships for Enhanced Deepfake Detection in Spatial-Frequency Domain," Electronics, vol. 13, no. 9, p. 1749, 2024.
11. Li, J., Xie, H., Li, J., Wang, Z., & Zhang, Y. (2021). Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6458–6467).
12. Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1–7). IEEE.
13. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1–11).
14. Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A., "CNN-Generated Images Are Surprisingly Easy to Spot... for Now," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8695–8704.
15. Coccomini, D. A., Messina, N., Gennaro, C., and Falchi, F., "Combining EfficientNet and Vision Transformers for Video Deepfake Detection," *Journal of Imaging*, vol. 8, no. 3, 2022.
16. Nguyen, H. H., Yamagishi, J., & Echizen, I. (2024). Exploring self-supervised vision transformers for deepfake detection: A comparative analysis. In 2024 IEEE International Joint Conference on Biometrics (IJCB) (pp. 1–10). IEEE.

17. Chen, C. F. R., Fan, Q., & Panda, R. (2021). CrossViT: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 357–366).

18. Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y.-G., & Li, S.-N. (2022). M2TR: Multi-modal multi-scale transformers for deepfake detection. In Proceedings of the International Conference on Multimedia Retrieval (pp. 615–623).

19. Tran, H. H., Nguyen, H. H., Yamagishi, J., and Echizen, I., "Meta Deepfake Detection," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, 2021, pp. 1–8.

20. Sun, K., Liu, H., Ye, Q., Liu, M., and Gao, Y., "Fake Trajectory Detection Network for Deepfake Video Detection," *IEEE Transactions on Multimedia*, vol. 25, pp. 1042–1056, 2023.

21. Alnaim, N. M., Almutairi, Z. M., Alsuwat, M. S., Alalawi, H. H., Alshobaili, A., & Alenezi, F. S. (2023). DFFMD: a deepfake face mask dataset for infectious disease era with deepfake detection algorithms. *IEEE Access*, *11*, 16711-16722.

22. Ramadhani, K. N., Munir, R., & Utama, N. P. (2024). Improving Video Vision Transformer for Deepfake Video Detection using Facial Landmark, Depthwise Separable Convolution and Self Attention. *IEEE Access*.

23. Zhou, Q., Zhou, Z., Bao, Z., Niu, W., and Liu, Y., "IIN-FFD: Intra-Inter Network for Face Forgery Detection," *Tsinghua Science and Technology*, vol. 29, no. 6, pp. 1839–1850, 2024.

24. Soudy, A., Afifi, M., and Abdelhamed, A., "Exploiting Vision Transformers and CNNs for Deepfake Detection," *IEEE Access*, vol. 12, pp. 31542–31555, 2024.

25. Zhou, J., Zhao, X., Xu, Q., Zhang, P., & Zhou, Z. (2024). MDCF- Net: Multi-Scale Dual-Branch Network for Compressed Face Forgery Detection. IEEE Access, 12, pp. 58740 – 58749.

26. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 3207–3216.

27. Zi, B., Chang, M., Chen, J., Ma, X., & Jiang, Y. G. (2020, October). Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2382-2390).

28. Altuncu, E., Franqueira, V. N. L., & Li, S. (2024). "Deepfake: Definitions, Performance Metrics and Standards, Datasets, and a Meta-Review," *Frontiers in Big Data*, vol. 7, 2024.