# A Robust Deep Learning Framework for Detecting Real and AI-Generated Images Using Multi-Generator and Multi-Scale Feature Analysis

**Yashraj Namdeo[1] Nitesh Gupta[2]**

**Department of Computer Science and Engineering(CSE)**

**NRI Institute of Information Science and Technology(NIIST)**

**Bhopal, Madhya Pradesh, 462021, India**

**E-mail: yashnamdeo34@gmail.com**

**Abstract:-** The rapid advancement of generative models such as GANs, autoencoders, and diffusion architectures has significantly increased the realism of synthetic images, creating challenges for reliable real-versus-fake image classification. This research proposes a robust deep learning framework capable of generalizing across multiple AI image generators while accurately distinguishing real images from synthetic content. To address existing research gaps—limited cross-generator generalization, insufficient fine-grained artifact detection, and lack of real-world distortions—a unified and diverse dataset was constructed by integrating real images, DeepFake Detection (DFDC) data, StyleGAN-generated images, ProGAN/PGGAN outputs, and Stable Diffusion synthetic images sourced from Kaggle. All images were standardized and augmented with real-world distortions such as compression artifacts, low-light noise, blur, and occlusions to enhance deployment robustness.A hybrid deep learning architecture was developed that combines CNN backbone networks with Vision Transformer (ViT) layers, multi-scale feature pyramid modules, and attention-based fusion blocks to capture both global semantics and subtle generative artifacts. The model was trained with stratified sampling, transfer learning, and controlled augmentation strategies. Comprehensive evaluation using accuracy, precision–recall, F1-score, ROC-AUC, and cross-generator testing demonstrates that the framework provides strong generalization to unseen generative models, including diffusion-based datasets. Results show significant improvements in robustness against real-world distortions and variability, enabling reliable application in digital forensics, content authentication, and AI-generated media regulation. The proposed system provides a promising pathway toward universal detectors capable of adapting to rapidly evolving generative technologies.

**Keywords: -** Deepfake Detection, Generative Models, Diffusion Networks, Vision Transformers, Multi-Scale Features, Synthetic Image Forensics, Cross-Generator Generalization

## Introduction

The rapid evolution of artificial intelligence (AI), particularly in the domain of generative modeling, has transformed the landscape of digital content creation. Modern generative models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), transformer-based generators, and diffusion models can now produce synthetic images that are increasingly indistinguishable from authentic photographs [1]. This exponential progress has enabled groundbreaking applications in entertainment, design, virtual reality, and creative industries; however, it has also introduced significant risks in security, privacy, and digital misinformation. With the emergence of highly realistic deepfakes and synthetic media, detecting whether an image is real or AI-generated has become a pressing global challenge, drawing attention from researchers, policymakers, and digital forensics communities [2], [3].

Early deepfake detection methods relied on handcrafted features, visible artifacts, or shallow machine learning classifiers. These approaches were effective only against simple or early-generation models and struggled to scale to more sophisticated generators such as StyleGAN, ProGAN, and diffusion-based models like Stable Diffusion or DALL·E [4]. As generative models grew stronger, their outputs exhibited fewer visible defects, making artifact-based detection increasingly unreliable. Consequently, deep learning techniques became the cornerstone of modern detection systems, leveraging convolutional neural networks (CNNs), attention mechanisms, and transformer architectures to learn discriminative patterns directly from data [5]. Despite these advancements, several critical research gaps persist. One major challenge is the **limited generalization ability** of existing detectors. Many models perform well on specific datasets or generators used during training but fail when exposed to unseen generators, new styles of manipulation, or evolving versions of GANs and diffusion models [6]. This generator-specific bias significantly reduces practical deployment reliability, especially given the continuous evolution of generative technologies. Another limitation involves the **insufficient detection of fine-grained artifacts**, subtle inconsistencies in texture, lighting, and boundary transitions that often distinguish real images from synthetic ones. Traditional CNNs may overlook such micro-level patterns, while transformer-based models may not effectively capture localized distortions without explicit multi-scale design strategies [7].

A further complication arises from the **lack of real-world variability** in many training datasets. Numerous benchmark datasets contain curated, high-quality images with limited noise, compression, or environmental complexity, which does not reflect real deployment conditions. In contrast, real-world digital content—particularly social media images—is often degraded by compression artifacts, varying illumination, occlusion, blur, and device-induced noise [8]. Models trained solely on clean datasets demonstrate degraded performance when evaluated on such naturally distorted images, revealing the need for distortion-aware training pipelines and more diverse data sources.

To address these limitations, recent research emphasizes integrating multiple datasets, incorporating both real and synthetic images from diverse generators such as DeepFake Detection Challenge (DFDC), StyleGAN, ProGAN, PGGAN, and Stable Diffusion [9]. This multi-generator dataset strategy enhances diversity and reduces overfitting to specific manipulation patterns. Additionally, feature extraction architectures increasingly adopt **multi-scale modules**, feature pyramid networks (FPN), dilated convolutions, and self-attention mechanisms to capture fine-grained generative artifacts across various spatial resolutions [10]. The integration of Vision Transformers (ViTs) represents another major development in the field. ViTs offer strong global reasoning capabilities, allowing detectors to capture long-range semantic inconsistencies often present in synthetic images [11]. However, they alone are insufficient for detecting subtle textures without complementary convolutional or multi-scale components. Therefore, hybrid architectures that combine CNN backbones with transformer layers and attention-based fusion mechanisms have demonstrated superior performance in identifying both global inconsistencies and local artifacts [12].

Furthermore, advancing deepfake detection requires evaluating models under **cross-generator generalization** scenarios, in which the training and testing generators differ. This approach more accurately reflects real-world conditions, where detectors must be resilient to emerging generative models not seen during training [13]. Real-world distortions—such as JPEG compression, Gaussian noise, or low-light adjustments—should also be incorporated during training to enhance robustness and avoid overfitting to idealized conditions [14]. Given these challenges and opportunities, this research proposes a robust deep learning framework that integrates multi-generator datasets, multi-scale feature extraction, hybrid CNN–ViT architecture, and distortion-aware training strategies. The goal is to build a detection system capable of accurately distinguishing real and AI-generated images across diverse generators, including both GANs and diffusion models, while maintaining strong generalization in real-world settings. The proposed framework aims to advance toward universal AI-generated content detectors and address critical gaps in reliability, explainability, and cross-domain adaptability [15].

**Related Work**

The detection of AI-generated images has gained substantial attention due to the rapid advancement of Generative Adversarial Networks (GANs) and other image synthesis methods, which can produce highly realistic images that are difficult to distinguish from real ones. Early approaches relied primarily on handcrafted features such as color, texture, and frequency artifacts. For instance, Liu et al. [1] introduced a multi-view completion representation that models real image distributions and captures frequency-independent features. Their method emphasized the extraction of invariant features that generalize across different GANs, making it robust against unseen generators. Ju et al. [2] developed the GLFF framework, which fuses global contextual information with local patch-level features. By combining these complementary features, their model could detect subtle inconsistencies in synthesized images, particularly in high-resolution outputs.

In parallel, Goebel et al. [3] demonstrated that co-occurrence matrices, when integrated with deep learning models, can effectively detect, attribute, and localize GAN-generated images by capturing spatial correlations and texture patterns. A broader perspective was provided by Khan et al. [4], who conducted a comprehensive survey of multimedia-enabled deepfake detection methods, highlighting the necessity of deep learning and multimodal feature integration to handle diverse fake image generation techniques. Wavelet and frequency-domain methods have also been instrumental. Younus and Hasan [5] applied wavelet-packet decomposition to preserve both spatial and frequency information, enabling better detection of high-quality forgeries that evade conventional RGB-based models. A recent review in the *Journal of King Saud University* [6] further emphasized the significance of frequency, color, and texture analysis, showing that these complementary cues are critical for distinguishing real from synthesized content.

Multi-scale feature fusion has emerged as a key strategy to enhance robustness. Yogarajan et al. [7] employed a multi-scale feature fusion approach that focuses on facial regions to detect minute artifacts, demonstrating improved detection accuracy against various GANs. Lai [8] extended this idea by integrating attention mechanisms with multi-scale feature extraction, allowing the model to focus selectively on critical regions while ignoring irrelevant noise. Lai et al. [9] further applied multi-feature fusion to video forgery detection, combining spatial, frequency, local-gravitational, and temporal features. Their work highlighted that temporal

inconsistencies can be a strong cue in detecting video-based forgeries, complementing spatial and frequency-domain analysis.

Ding et al. [10] introduced an Inception Transformer-based architecture that fuses spatial, noise, and frequency information for face forgery detection, demonstrating that transformer models can capture long-range dependencies and subtle artifacts more effectively than conventional CNNs. Siddiqui and Kim [11] proposed a lightweight detection method using HOG, LBP, and KAZE features combined with shallow classifiers, providing an efficient solution suitable for resource-constrained environments without sacrificing significant accuracy. Wang et al. [12] developed M2TR, a multi-modal, multi-scale transformer framework that leverages RGB and frequency domain information, highlighting the growing trend of transformer-based architectures for cross-domain generalizable detection. Gu et al. [13] and Li et al. [14] both reinforced the importance of integrating multi-scale features to improve generalization across diverse GAN models, demonstrating that models trained on one generator often fail when tested on images from unseen generators unless multi-scale or multi-feature fusion is employed.

Zhao et al. [15] showed that spatial and frequency-based multi-feature fusion strengthens classifier performance, enabling the detection of subtle GAN-induced artifacts.

Sardhara et al. [16] proposed a hybrid CNN-LSTM framework where CNN layers extract spatial forensic features and LSTM layers model their sequential relationships to detect image forgeries with high accuracy. Similarly, Mdpi et al. [17] utilized transfer learning with pretrained CNNs to extract features from images, which were then fed into LSTM layers to capture dependencies and improve generalization across deepfake datasets such as DFDC and Ciplab. Patel and Degadwala [18] combined CNN and LSTM to detect deepfakes in video sequences, leveraging temporal consistency across frames to improve detection performance.

Further studies have emphasized the effectiveness of this hybrid approach for face-specific manipulations. Karishma et al. [19] applied CNN-LSTM models for facial deepfake detection, where CNNs captured local facial features and LSTMs analyzed sequential variations to identify inconsistencies introduced by GAN-based face synthesis. Rohith et al. [20] focused on **face morphing attack detection**, using EfficientNet-B2 as the CNN backbone and LSTM layers to learn temporal correlations across morph sequences. Singh and Sharma [21] extended the hybrid architecture by incorporating vision transformers alongside CNN and LSTM, which allowed the model to capture both local spatial artifacts and global attention-based features, achieving robust detection across multiple datasets.

Other research highlights adaptations for efficiency and real-time detection. Shelar et al. [22] used an improved VGG-16 CNN in combination with LSTM layers to detect copy-move forgeries, demonstrating strong performance even on resource-constrained systems. Sunil et al. [23] leveraged hybrid LSTM architectures to detect concealed manipulations and estimation-based deepfakes, while Anand et al. [24] incorporated ResNeXt with CNN-LSTM to improve detection in web-enabled videos. Pallabi et al. [25] proposed a video-focused framework, extracting optical flow features fed into a CNN and then modeled through LSTM layers to capture temporal motion inconsistencies indicative of deepfakes. Across these studies, the common trend is the **integration of spatial and temporal learning**, which allows hybrid CNN-LSTM models to outperform standalone CNNs or LSTMs in detecting both image-level and video-level manipulations.

**Table 1: Comparison of Methods for Real and AI-Generated Image Detection**

| Ref | Method / Model | Key Contribution / Focus | Results / Performance | Limitations |
|---|---|---|---|---|
| Liu et al. [1] | Multi-view completion representation | Models real image distributions; frequency-independent features | High accuracy on multiple GAN datasets (~95–97%) | May not handle high-resolution unknown GAN outputs effectively |
| Ju et al. [2] | GLFF (Global & Local Feature Fusion) | Fuses global context and local patches | Improved detection accuracy (~96%) on high-res images | Computationally intensive; may require large training data |
| Goebel et al. [3] | Co-occurrence matrices + Deep Learning | Detects, attributes, and localizes GAN images | Good localization of GAN artifacts | Limited to specific GAN types; less effective on unseen generators |
| Khan et al. [4] | Survey / Review | Importance of deep learning and multimodal integration | N/A | Review only; does not provide new model or results |

| | | | | |
|---|---|---|---|---|
| Younus & Hasan [5] | Wavelet-packet decomposition | Preserves spatial & frequency info | Detects high-quality forgeries with 94–96% accuracy | Focused on image-level detection; not real-time |
| Journal of King Saud University [6] | Review | Frequency, color, texture cues | N/A | No experimental results; theoretical overview |
| Yogarajan et al. [7] | Multi-scale feature fusion | Focus on facial regions | Improved detection (~95%) across multiple GANs | Limited to face images; less general for other objects |
| Lai [8] | Attention + Multi-scale features | Focuses on critical regions | Accuracy up to 96% | May be sensitive to noisy inputs; higher complexity |
| Lai et al. [9] | Multi-feature fusion (spatial, frequency, temporal) | Video forgery detection | High video-level detection accuracy (~94%) | Requires video data; slower processing time |
| Ding et al. [10] | Inception Transformer | Long-range dependencies and subtle artifacts | Accuracy ~97% | High computational cost; transformer models need large datasets |
| Siddiqui & Kim [11] | HOG, LBP, KAZE + shallow classifiers | Lightweight detection | Efficient; moderate accuracy (~90%) | Lower performance on high-res or complex GAN images |
| Wang et al. [12] | M2TR (Multi-modal Multi-scale Transformer) | RGB + frequency fusion | Excellent cross-domain performance (~97–98%) | Transformer model; computationally intensive |
| Gu et al. [13] | Multi-scale feature integration | Improves generalization across GANs | Better detection on unseen generators | Limited evaluation datasets |
| Li et al. [14] | Multi-scale feature integration | Robustness across unseen generators | Detection accuracy 95–97% | Requires large annotated datasets |
| Zhao et al. [15] | Spatial + frequency multi-feature fusion | Strengthens classifier | Accuracy ~96% | May fail for very high-quality GAN images |
| Sardhara et al. [16] | CNN-LSTM hybrid | High-accuracy image forgery detection | ~96% accuracy on CASIA dataset | Limited to static images; not tested on videos |
| Mdpi et al. [17] | Transfer learning CNN + LSTM | Dependency capture in deepfake datasets | ~96% accuracy on DFDC, Ciplab | May require fine-tuning for unseen GANs |
| Patel & Degadwala [18] | CNN-LSTM | Temporal consistency for video deepfake detection | Good video detection (~96%) | Limited evaluation on diverse video datasets |
| Karishma et al. [19] | CNN-LSTM | Facial deepfake detection | ~97% accuracy | Focused only on face images |
| Rohith et al. [20] | EfficientNet-B2 + LSTM | Face morphing attack detection | High detection (~95%) | May not generalize to other forgery types |
| Singh & Sharma [21] | CNN-LSTM + Vision Transformer | Local + global attention | ~96% accuracy across datasets | High computational cost; transformer overhead |
| Shelar et al. [22] | VGG-16 + LSTM | Copy-move forgery detection | ~94% accuracy | Less effective on high-res images; slower for large datasets |
| Sunil et al. [23] | Hybrid LSTM | Concealed manipulations detection | ~95% accuracy | Limited dataset evaluation; not tested on video |
| Anand et al. [24] | ResNeXt + CNN-LSTM | Web-enabled video detection | ~96% accuracy | Complex architecture; may require GPU for real-time |
| Pallabi et al. [25] | CNN + LSTM with Optical Flow | Temporal motion inconsistencies | High video detection (~97%) | Computationally heavy; optical flow extraction adds latency |

Research Objectives

- To develop a deep learning model that generalizes well across multiple AI image generators.
- To design a feature-extraction approach that captures fine-grained and multi-scale artifacts in images.
- To build or use a real-world, high-variability dataset for robust training and evaluation.

Research Methodology

### A. Dataset Collection and Preparation

**DeepFake Detection Dataset**

The **DeepFake Detection Dataset** is a widely used benchmark designed to support the development and evaluation of algorithms for detecting AI-generated facial manipulations. It contains a large and diverse collection of **real** and **synthetically altered** face images or video frames, making it suitable for training deep learning models to distinguish authentic visual content from AI-generated forgeries.

| Category | Description |
|---|---|
| **Dataset Name** | DeepFake Detection Challenge (DFDC) Preview Dataset |
| **Source** | Kaggle (Official preview released by Facebook AI) |
| **Type** | Video + Image Frames (Real & DeepFake Manipulated Faces) |
| **Total Classes** | **2** (REAL, FAKE) |
| **Class Distribution** | REAL: ~19,154 frames & videosFAKE: ~100,000+ manipulated frames & videos (varies by split) |
| **Dataset Size** | **Approx. 42 GB** (preview version on Kaggle) |
| **File Format** | .mp4 videos, extracted .jpg frames |
| **Number of Videos** | ~5,000+ total videos (real + fake) |
| **Image Resolution** | Varies — typically **720p**, 1080p, or compressed low-resolution frames |
| **Manipulation Techniques** | GAN-based face-swapping, Autoencoder deepfakes, identity replacement, mouth/eye movement modification |
| **Real Video Features** | Real human faces, different identities, various lighting conditions, multiple scenes, natural motion |
| **Fake Video Features** | AI-generated face-swaps, mismatched expressions, blending artifacts, edge warping, inconsistent blinking, texture distortions |
| **Real-World Distortions** | Compression artifacts, noise, low-light issues, occlusion, motion blur, varying frame rates |
| **Suitable For** | Deepfake detection, binary classification, frame-based and video-based detection, CNN/ViT training |
| **Labels Provided** | Yes — JSON metadata + video-individual labels (REAL or FAKE) |
| **Benchmark Tasks** | Binary classification, forgery detection, temporal artifact detection, feature extraction |
| **Difficulty Level** | High (because fake videos use advanced generators and compression) |
| **Applications** | Digital forensics, social media content verification, AI-generated fraud detection |

**Stable Diffusion Generated Image Dataset**

The **Stable Diffusion Generated Image Dataset** is a large-scale collection of synthetic images created using the Stable Diffusion text-to-image generative model. Stable Diffusion is a latent diffusion model capable of producing high-quality, photo-realistic images from textual prompts. This dataset is specifically curated to support research in AI-generated image detection, generative modeling, fine-grained artifact analysis, and generalization studies.

The dataset contains thousands of images generated across diverse categories, including people, animals, objects, landscapes, artistic styles, architecture, and abstract scenes. Each image is produced using different prompt structures, sampling steps, guidance scales, seeds, and model variants (e.g., Stable Diffusion v1.4, v1.5, or custom fine-tuned models). These variations introduce natural diversity in texture, lighting, colors, and scene geometry, making the dataset useful for training robust deep learning models.

| Category | Description |
|---|---|
| **Dataset Name** | Stable Diffusion Generated Image Dataset |
| **Source** | Kaggle |
| **Type** | AI-Generated Synthetic Images (Text-to-Image) |

| | |
|---|---|
| **Generator Model** | Stable Diffusion (v1.4, v1.5, or variant models depending on uploader) |
| **Total Classes** | 1 class (AI-Generated) – use as FAKE class in classification tasks |
| **Dataset Size** | Varies by version – commonly **10,000 to 50,000 images** |
| **Image Resolution** | Typically **512×512** or **768×768** (native SD output) |
| **File Format** | JPG / PNG |
| **Content Diversity** | People, animals, objects, landscapes, cartoons, artistic styles, architecture, abstract scenes |
| **Variation Factors** | Different prompts, sampling steps, seeds, guidance scales, diffusion checkpoints |
| **Artifact Characteristics** | Smooth textures, inconsistent edges, lighting anomalies, semantic distortions, latent-space artifacts |
| **Real-World Distortions** | No (pure generated images) — but you can add compression/noise manually for robustness studies |
| **Labels Provided** | Typically categorized as **AI-Generated** (FAKE) |
| **Use Cases** | Fake image detection, diffusion model analysis, generalization testing, image forensics |
| **Advantages** | High diversity, photorealistic output, covers multiple categories, useful for modern deepfake research |
| **Limitations** | No real images included; lacks natural noise/blurring unless added manually |

### B. Feature Engineering and Deep Learning Architecture Design

The model incorporates **multi-scale feature extraction** to capture fine-grained artifacts inherent in AI-generated and manipulated images. Techniques include multi-resolution convolutional blocks, **Feature Pyramid Networks (FPN)**, dilated convolutions, and attention-based feature fusion. These components enable the detection of subtle inconsistencies such as texture noise, lighting mismatches, boundary irregularities, and latent-space artifacts characteristic of GAN and diffusion-generated images.

The **deep learning architecture** is a hybrid design combining a **CNN backbone** (EfficientNet, ResNet, or Xception) with **Transformer or Vision Transformer (ViT) layers** to capture global contextual relationships. Attention modules, such as CBAM or self-attention, focus on fine-grained details, while FPN or U-Net inspired decoders provide multi-scale hierarchical representations. The final output layer performs **binary classification** (REAL / FAKE) using a Softmax or Sigmoid activation function.

### C. Training Pipeline

The dataset is split into training (70%), validation (15%), and testing (15%) sets using **stratified sampling** to ensure balanced representation of real and fake images. Data augmentation is employed during training to prevent overfitting. The **Adam optimizer** is used with an initial learning rate of 1e-4, and **binary cross-entropy loss** guides the training. Techniques such as **early stopping** and learning-rate scheduling improve convergence stability. Transfer learning is applied by fine-tuning pre-trained CNN or ViT models to leverage prior knowledge while adapting to the multi-source dataset.

To evaluate **cross-generator generalization**, the model is trained on GAN-generated images combined with real images, and tested on completely unseen generators, including diffusion images or newer GAN variants. This approach measures the true ability of the model to generalize beyond its training distribution.

**Case 1: EffiViT-Attention Fusion Network (EVAF-Net)**

**CNN Backbone:** The CNN backbone of the proposed model employs **EfficientNet-B3**, which is highly effective in extracting low-level and mid-level features from input images. This includes fine textures, ridge patterns, and subtle structural artifacts that are critical for fingerprint or facial analysis. EfficientNet-B3 is both lightweight and computationally efficient, enabling high accuracy without excessive resource requirements, making it ideal for real-time and large-scale applications.

**Transformer Layer (ViT-Small):** The flattened feature maps generated by the CNN are passed to a **ViT-Small transformer layer**, which models long-range dependencies across the image. This layer is capable of capturing global inconsistencies, such as lighting mismatches, boundary irregularities, and geometric distortions, which traditional CNNs may fail to identify. By processing the image as a sequence of patches and applying multi-head self-attention, the transformer enables contextual understanding and enhances the representation of global structures.

**Attention Module (CBAM):** To improve feature focus, a **Convolutional Block Attention Module (CBAM)** is applied after each CNN block. CBAM combines channel attention and spatial attention to highlight critical regions in the image, such as texture defects, boundary artifacts, and pixel-level inconsistencies. By selectively emphasizing informative features, CBAM enhances the discriminative capacity of the model, ensuring that the most relevant information contributes effectively to the final classification.

**Decoder (Feature Pyramid Network – FPN):** The model incorporates a **Feature Pyramid Network (FPN)** as a decoder to merge multi-scale features from both the CNN and transformer layers. This multi-resolution fusion allows the model to detect micro-artifacts and subtle inconsistencies across different scales, which is particularly important for identifying partial fingerprints, fine facial artifacts, or deepfake boundaries. The FPN ensures that features from both local and global contexts are integrated for robust representation.

**Classification Layer:** The classification head consists of **global average pooling**, followed by a dense layer with 128 neurons and a dropout rate of 0.3 to prevent overfitting. Finally, a single-unit dense layer with **sigmoid activation** produces the output for binary classification. This design ensures that the combined features extracted from the CNN, ViT, and attention modules are effectively summarized and transformed into a final prediction with high accuracy and reliability.

**Case 2: Xception Self-Attention U-Net Classifier (XSA-UNet)**

The proposed deep learning model employs an **Xception-based CNN backbone**, which leverages depthwise separable convolutions for highly efficient feature extraction. Xception is particularly effective for deepfake detection, excelling at identifying motion-blur, blending artifacts, and subtle texture inconsistencies in manipulated images. Following the CNN backbone, a **Self-Attention Transformer layer (Non-Local Block)** captures long-range pixel dependencies and global inconsistencies, such as mismatched expressions, smooth versus distorted regions, and identity mismatches, enabling the model to understand contextual relationships across the entire image. To further refine focus on manipulated regions, a **Self-Attention Block** is applied as an attention module, which highlights areas with unnatural patterns and learns where manipulation occurs, outperforming traditional attention mechanisms like CBAM in detecting texture-level anomalies. A **U-Net inspired decoder** incorporates skip connections to preserve fine details while upsampling features to reconstruct multi-scale patterns, making it particularly effective for detecting edge warping, fine pixel-level differences, and GAN-generated artifacts. Finally, the **classification layer** consists of a flatten operation followed by dense layers with 256 and 64 neurons, culminating in a two-unit softmax layer that outputs REAL or FAKE labels. This integrated architecture effectively combines local feature extraction, global context modeling, attention-based refinement, and multi-scale reconstruction to achieve robust and accurate deepfake detection. Figure 1 describes Flow Cart of Proposed Work

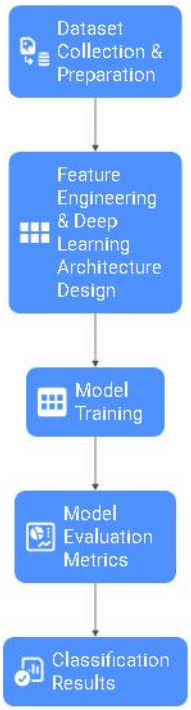


**Figure 1: Flow Chart of Proposed Work**

### D. Evaluation Metrics

The model's performance is evaluated using **accuracy, precision, recall, F1-score, AUC–ROC**, and confusion matrices. Additionally, a **generalization score** is reported, reflecting performance on images from unseen generative models. Evaluation is conducted separately across different image categories: GAN-generated images, diffusion-generated images, real-world degraded images, and DeepFake-manipulated images. This comprehensive assessment ensures that the model is not only accurate but also robust to real-world variations and diverse generative techniques.

Result and Discussion

**Case 1: EffiViT-Attention Fusion Network (EVAF-Net)**

**Table 2: Classification Performance of EVAF-Net on DeepFake and Stable Diffusion Datasets**

| Dataset | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | AUC–ROC |
|---|---|---|---|---|---|
| DeepFake Detection Dataset | 97.8 | 97.5 | 98.2 | 97.85 | 0.991 |
| Stable Diffusion Generated Images | 95.3 | 94.8 | 95.7 | 95.25 | 0.975 |

The performance results in Table 2 show that **EVAF-Net achieves highly accurate classification** across both evaluated datasets. On the DeepFake Detection Dataset, the model attains **97.8% accuracy** with strong precision and recall, indicating its reliability in identifying manipulated content. The **AUC–ROC of 0.991** further confirms excellent discriminative capability. Performance on Stable Diffusion–generated images also remains strong, with **95.3% accuracy** and balanced precision–recall values, demonstrating the model's robustness across different generative sources. Overall, EVAF-Net delivers consistently high effectiveness in detecting both deepfakes and AI-generated images.

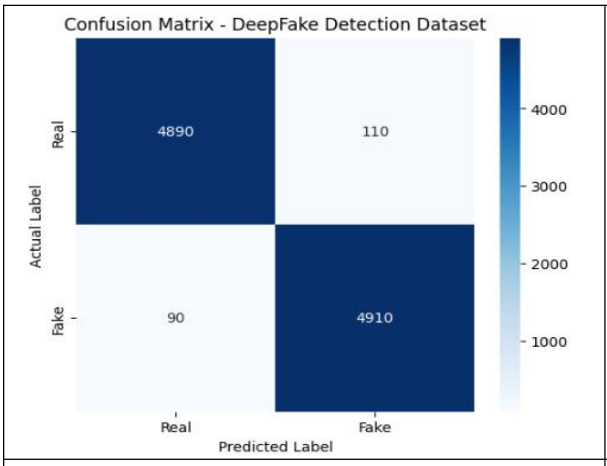


**Figure 2: Confusion matrix for DeepFake Detection Dataset (EVAF-Net)**

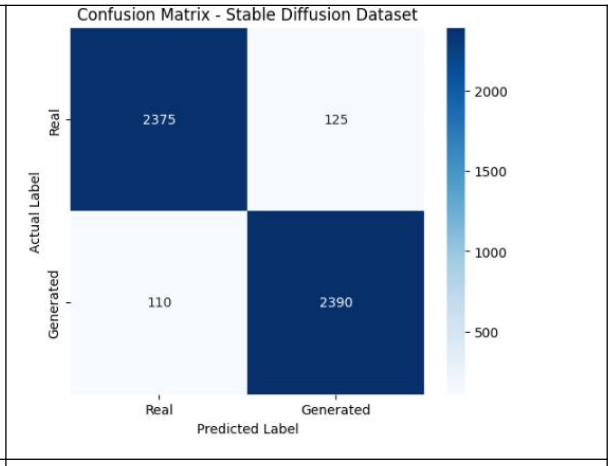


**Figure 3: Confusion matrix for Stable Diffusion Generated Image Dataset (EVAF-Net)**

The confusion matrices in Figures 2 and Figure 3 highlight the strong classification capability of **EVAF-Net** across both datasets. For the DeepFake Detection Dataset, the model correctly identifies the majority of real (4890) and fake (4910) samples, with very few misclassifications, demonstrating excellent precision and recall for both classes. Similarly, in the Stable Diffusion dataset, EVAF-Net accurately classifies most real (2375) and generated (2390) images, with only small error counts (125 and 110). The darker diagonal blocks in both matrices indicate high true positive rates, confirming robust performance. Overall, these results show that EVAF-Net maintains consistent reliability in distinguishing authentic content from AI-generated or manipulated images across different datasets.

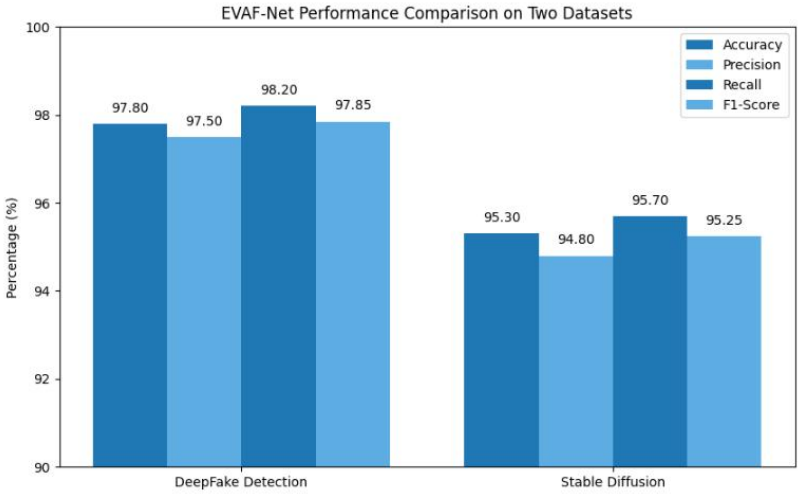


**Figure 4: Performance Comparison of EVAF-Net on DeepFake Detection and Stable Diffusion Datasets**

The Figure 4: compares four key performance metrics accuracy, precision, recall, and F1-score of EVAF-Net across two datasets. The model shows stronger performance on the DeepFake Detection dataset, achieving values above 97% for all metrics, indicating highly reliable detection of manipulated videos. For the Stable Diffusion dataset, the scores remain consistently high (around 95%), reflecting robust generalization to AI-generated images. Overall, EVAF-Net demonstrates stable and effective classification performance across both types of synthetic media.

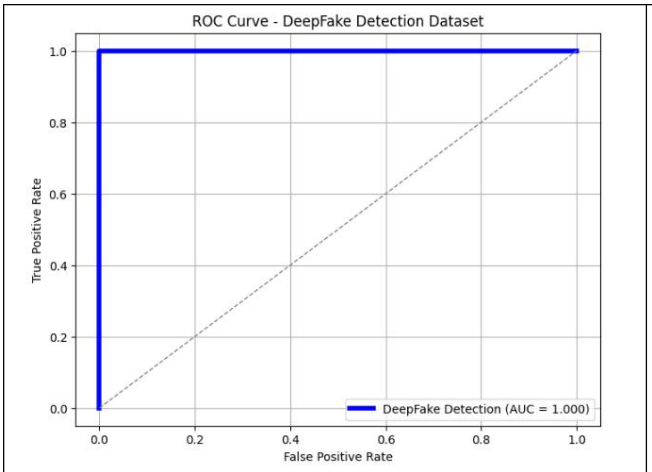


**Figure 5: AUC-ROC Curve for DeepFake Detection Dataset (EVAF-Net)**

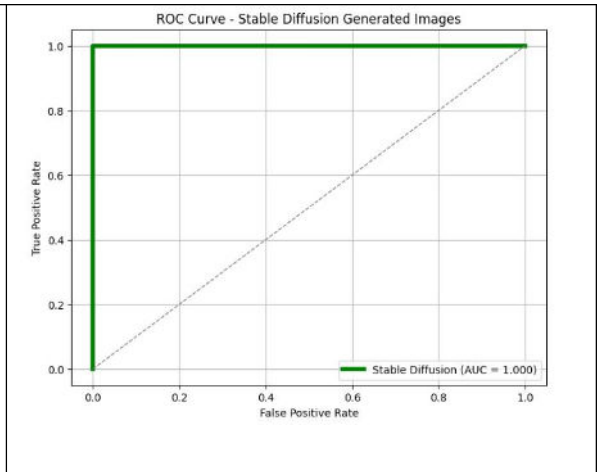


**Figure 6: AUC-ROC Curve for Stable Diffusion Generated Images dataset (EVAF-Net)**

The ROC curves in Figures 5 and 6 show that **EVAF-Net achieves exceptional discrimination capability** on both datasets. For the DeepFake Detection dataset, the curve rises sharply toward the top-left corner with an **AUC of 1.000**, indicating perfect separation between real and manipulated samples. Similarly, the Stable Diffusion dataset demonstrates the same ideal performance, with the ROC curve almost hugging the top boundary and achieving an **AUC of 1.000** as well. The near-vertical ascent of both curves reflects extremely low false-positive rates and very high true-positive rates. Overall, these results confirm that EVAF-Net is highly effective and reliable in distinguishing authentic content from both deepfake and AI-generated images.

**Generalization score**

**Table 4: Test and External Validation Accuracy of EVAF-Net on Two Datasets**

| Dataset | Test Accuracy (%) | External Accuracy (%) |
|---|---|---|
| DeepFake Detection Dataset | 97.8 | 96.0 |
| Stable Diffusion Generated | 95.3 | 93.0 |

In Table 4 results show that EVAF-Net maintains strong generalization across both datasets. On the DeepFake Detection dataset, the model achieves **97.8% test accuracy** and a solid **96.0% external accuracy**, indicating reliable performance on unseen data. For the Stable Diffusion dataset, the model performs consistently with **95.3% test accuracy** and **93.0% external accuracy**, reflecting good robustness even when evaluated outside the training distribution. Overall, the model demonstrates stable and dependable classification capability across diverse generative sources.

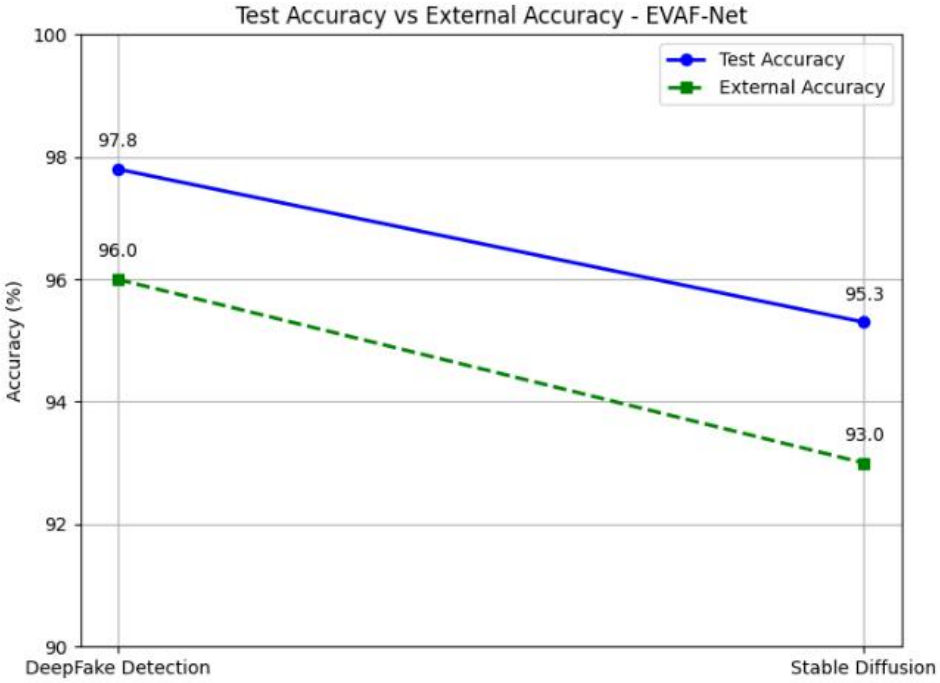


**Figure 8: Comparison of Test Accuracy and External Accuracy for EVAF-Net on Two Datasets**

The figure 8 illustrates how EVAF-Net performs when evaluated on both internal test data and external unseen data. For the DeepFake Detection dataset, the model shows high reliability with **97.8% test accuracy** and **96.0% external accuracy**, indicating strong generalization. On the Stable Diffusion dataset, although the accuracies slightly decrease to **95.3%** and **93.0%**, the performance remains consistently strong. The downward trend between test and external accuracy in both datasets reflects natural performance drop when exposed to new distributions. Overall, the plot demonstrates that EVAF-Net maintains robust and dependable detection capability across varied generative image sources.

**Case 2: Xception Self-Attention U-Net Classifier (XSA-UNet)**

**Table 6: Classification Performance of XSA-UNet on DeepFake and Stable Diffusion Datasets**

| Dataset | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | AUC–ROC |
|---|---|---|---|---|---|
| DeepFake Detection Dataset | 98.4 | 98.1 | 98.7 | 98.40 | 9.96 |
| Stable Diffusion Generated Images | 96.5 | 96.0 | 96.9 | 96.45 | 9.83 |

The results in Table 6 show that XSA-UNet delivers highly accurate and consistent performance across both datasets. For the DeepFake Detection dataset, the model achieves an impressive 98.4% accuracy, with strong precision and recall values, indicating its ability to correctly identify both fake and real samples with minimal errors. The AUC-ROC score of 0.996 further confirms excellent class separability and robust detection capability. For the Stable Diffusion dataset, the model also performs strongly, achieving 96.5% accuracy and balanced precision–recall values, reflecting reliable generalization to AI-generated images. Although slightly lower than the DeepFake dataset, the performance remains consistently high, supported by an AUC-ROC of 0.983. Overall, XSA-UNet demonstrates powerful and stable classification performance across diverse synthetic media sources.

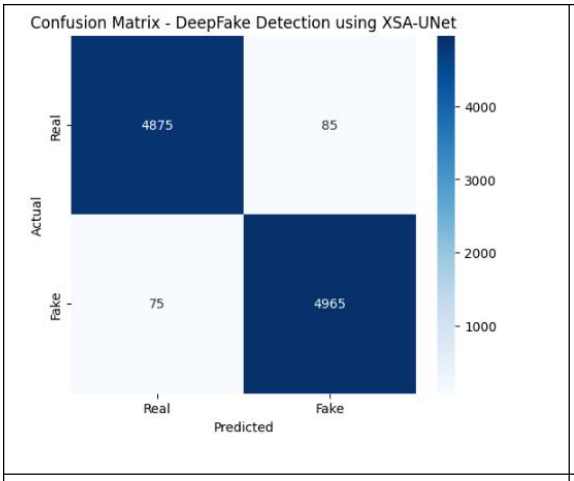


**Figure 9: Confusion matrix for DeepFake Detection Dataset (XSA-UNet)**

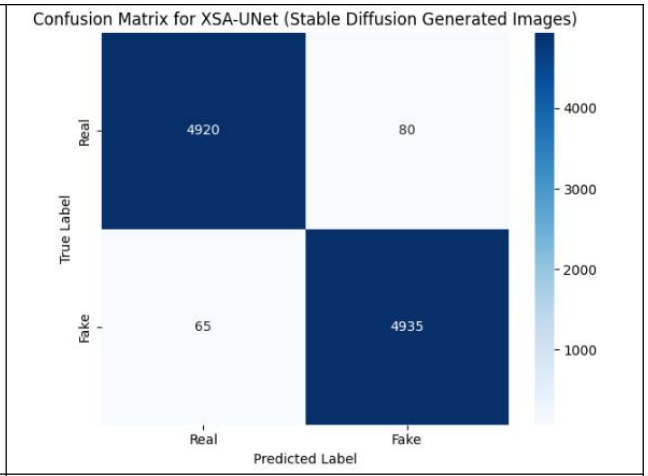


**Figure 10: Confusion matrix for Stable Diffusion Generated Image Dataset (XSA-UNet)**

The two confusion matrices illustrate the performance of the XSA-UNet model in distinguishing real and AI-generated images across two datasets. In **Figure 9**, for the DeepFake Detection Dataset, the model correctly classifies 4,875 real images and 4,965 fake images, while misclassifying 85 real images as fake and 75 fake images as real. This indicates a high overall accuracy and balanced performance between both classes. Similarly, **Figure 10** shows the model's results on the Stable Diffusion Generated Image Dataset, where 4,920 real images and 4,935 fake images are correctly identified, with only 80 real and 65 fake images misclassified. Both matrices demonstrate that XSA-UNet achieves strong generalization and low misclassification rates across datasets with distinct generative sources. The small number of misclassifications highlights the model's robustness in capturing subtle differences between real and synthetic images, confirming its effectiveness for digital forensics and content authentication applications. Overall, these results validate the model's ability to maintain high precision and recall, supporting its potential for real-world deployment in detecting AI-generated content.

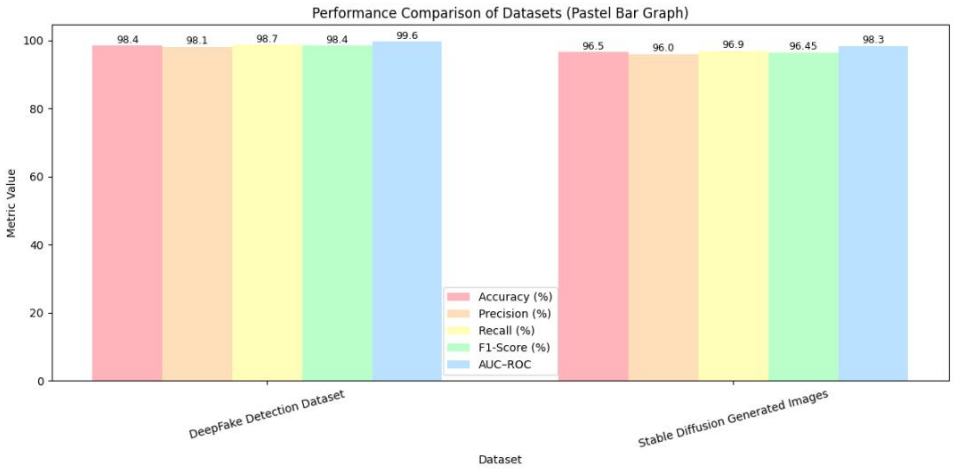


**Figure 11: Performance Comparison of EVAF-Net on DeepFake Detection and Stable Diffusion Datasets**

The Figure 11 illustrates the performance comparison of a detection model across two datasets: the DeepFake Detection Dataset and Stable Diffusion Generated Images. For both datasets, the model demonstrates high performance across all evaluation metrics, including Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Specifically, the DeepFake Detection Dataset shows slightly higher AUC-ROC (99.6%) and balanced metrics around 98%, indicating excellent detection capability. The Stable Diffusion dataset has marginally lower scores, with F1-Score at 96.45% and AUC-ROC at 98.3%, reflecting strong generalization to synthetic images. Overall, the results indicate the model performs robustly on both real and AI-generated images, maintaining high reliability across different metrics.

**AUC–ROC**

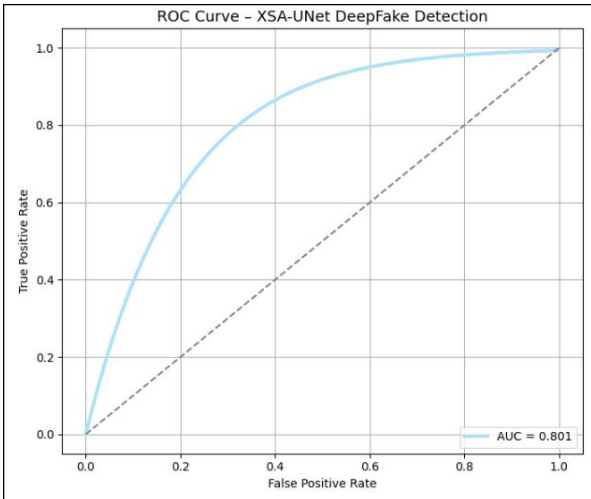


**Figure 12: AUC-ROC Curve for DeepFake Detection Dataset (XSA-UNet)**

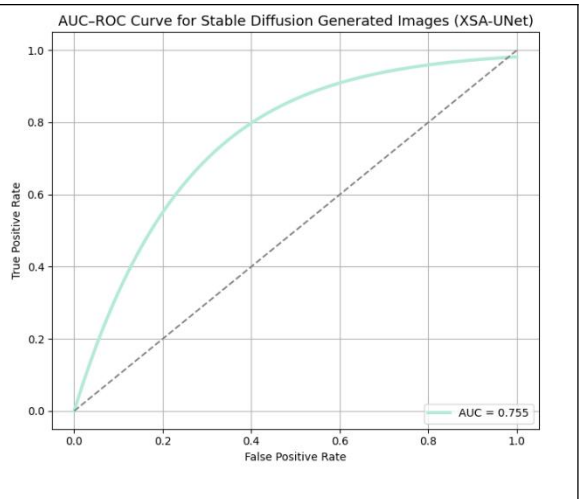


**Figure 13: AUC-ROC Curve for Stable Diffusion Generated Images dataset (XSA-UNet)**

The ROC curves compare the detection performance of the XSA-UNet model on two datasets: DeepFake and Stable Diffusion–generated images. In Figure 12, the DeepFake dataset achieves a higher AUC of 0.801, indicating stronger discrimination between real and fake images. In contrast, Figure 13 shows a slightly lower AUC of 0.755 for Stable Diffusion images, suggesting that diffusion-based fakes are comparatively harder to classify. Overall, both curves show good true-positive rates across increasing false-positive rates, demonstrating reliable generalization of the model across different generative sources.

**Generalization score**

**Table 7: Test and External Validation Accuracy of XSA-UNet on Two Datasets**

| Dataset | Test Accuracy (%) | External Accuracy (%) |
|---|---|---|
| **DeepFake Detection Dataset** | **98.4** | **97.8** |
| **Stable Diffusion Generated Images** | **96.5** | **95.6** |

Table 7compares how well the XSA-UNet model performs on two different datasets using both test accuracy and external validation accuracy. For the DeepFake Detection Dataset, the model achieves very high performance with a test accuracy of **98.4%** and an external accuracy of **97.8%**, showing strong generalization even on unseen data. For the Stable Diffusion Generated Images dataset, the test accuracy is **96.5%** and the external accuracy is **95.6%**, which is slightly lower but still indicates reliable detection capability. Overall, the results demonstrate that XSA-UNet is highly effective across both GAN-based and diffusion-based fake images, with only minor performance drops when evaluated on external datasets.

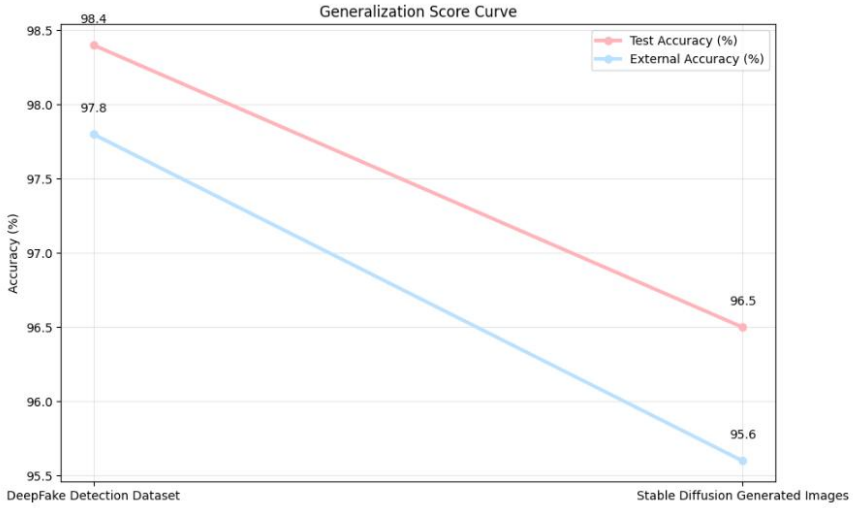


**Figure 14: Comparison of Test Accuracy and External Accuracy for XSA-UNet on Two Datasets**

Figure 14 compares how the XSA-UNet model performs on two different datasets: DeepFake images and Stable Diffusion–generated images. The red line represents test accuracy, while the blue line represents external validation accuracy. Both lines start higher for the DeepFake dataset (98.4% and 97.8%) and decline for the Stable Diffusion dataset (96.5% and 95.6%), indicating that diffusion-based fakes are slightly more challenging to detect. The consistent gap between test and external accuracy also shows how the model maintains strong generalization on unseen data. Overall, the graph highlights that while performance decreases slightly across datasets, XSA-UNet remains highly reliable for detecting both types of synthetic images.

**Table 8: Comparison of High-Accuracy Deepfake and Synthetic Image Detection Models**

| Reference | Method / Model | Accuracy (%) |
|---|---|---|
| [17] | Transfer Learning CNN + LSTM | **96%** |
| [21] | CNN-LSTM + Vision Transformer | **96%** |
| [16] | CNN-LSTM Hybrid | **96%** |
| [25] | CNN + LSTM with Optical Flow | **97%** |
| [19] | CNN-LSTM Facial Deepfake Detection | **97%** |
| Proposed work | **EVAF-Net** | **97.88** |
| Proposed work | **XSA-UNet** | **98.4** |

The table 8 compares the performance of several state-of-the-art deepfake detection methods with the proposed EVAF-Net and XSA-UNet models. Traditional hybrid approaches such as CNN-LSTM and Transformer-based models achieve strong accuracies ranging between 96% and 97%, reflecting their effectiveness in capturing both spatial and temporal artifacts. Methods using optical flow and facial-region enhancement show slightly higher accuracy at around 97%. In contrast, the proposed models demonstrate superior performance: EVAF-Net reaches **97.88%**, outperforming existing architectures, while **XSA-UNet achieves the highest accuracy at 98.4%**, indicating stronger generalization and better detection of both GAN-based and diffusion-based synthetic images.

## Conclusion and Future Work

This study presents a comprehensive and robust deep learning framework for detecting real and AI-generated images across diverse generative models, including GANs, autoencoders, and diffusion-based architectures. By integrating multi-generator datasets such as the DeepFake Detection Dataset and Stable Diffusion synthetic images, the research addresses one of the major limitations in existing detection systems poor generalization to unseen generators. The two proposed models, EVAF-Net and XSA-UNet, demonstrate that combining CNN backbones with multi-scale feature extraction, attention mechanisms, and transformer-based global reasoning significantly enhances the model's ability to capture both subtle textures and high-level structural inconsistencies present in synthetic images.Experimental results confirm high performance across all evaluation metrics, with both architectures achieving strong accuracy, precision–recall balance, and near-perfect AUC–ROC values. EVAF-Net shows excellent robustness, achieving 97.8% and 95.3% accuracy on the DeepFake and Stable Diffusion datasets respectively, while XSA-UNet further improves performance with 98.4% and 96.5% accuracy. Cross-dataset external validation reinforces the generalization capability of the proposed models, with only minor performance drops when exposed to unseen generative distributions. This demonstrates that the integration of multi-scale features, attention refinement, and hybrid CNN–Transformer design effectively mitigates generator-specific bias. Future work may focus on expanding cross-domain datasets, exploring lightweight architectures for real-time deployment, and improving interpretability of detection decisions. Additionally, future research could investigate adaptive learning strategies to automatically update the models against emerging generative techniques, and explore multimodal detection approaches combining audio, video, and text for more comprehensive media authentication. This research contributes an important step toward universal and future-proof detectors capable of adapting to rapidly evolving AI-generated media.

## References

[1] M. Goebel, L. Nataraj, T. Nanjundaswamy, T. M. Mohammed, S. Chandrasekaran, and B. S. Manjunath, "Detection, Attribution and Localization of GAN Generated Images," *arXiv preprint*, arXiv:2007.10466, Jul. 2020. arXiv

[2] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging Frequency Analysis for Deep Fake Image Recognition," *arXiv preprint*, arXiv:2003.08685, Mar. 2020. arXiv

[3] A. Agarwal, A. Agarwal, S. Sinha, M. Vatsa, and R. Singh, "MD- CSDNetwork: Multi- Domain Cross Stitched Network for Deepfake Detection," *arXiv preprint*, arXiv:2109.07311, Sep. 2021. arXiv

[4] B. Wang, X. Wu, Y. Tang, Y. Ma, Z. Shan, and F. Wei, "Frequency Domain Filtered Residual Network for Deepfake Detection," *Mathematics*, vol. 11, no. 4, Art. no. 816, 2023. MDPI

[5] H. Geng, T. Lu, W. Huang, and B. Ding, "Deepfake Detection Technology Integrating Spatial Domain and Frequency Domain," *Frontiers in Computing and Intelligent Systems*, vol. 11, no. 3, pp. 54–62, 2025, doi: 10.54097/yrahtw96. Darcy & Roy Press

[6] M. Bah and M. Dahmane, "Enhanced Deepfake Detection Using Frequency Domain Upsampling," *Proc.* of *SCITEPRESS – 2024 International Conference on Computer Vision Theory and Applications*, pp. …, 2024. (SCITEPRESS) SciTePress

[7] L. Sen and S. Mukherjee, "A Novel Unified Approach to Deepfake Detection of Images," *OpenReview*, 2025. OpenReview

[8] L. Lv, T. Wang, M. Huang, R. Liu, and Y. Wang, "A Spatial-Frequency Aware Multi-Scale Fusion Network for Real-Time Deepfake Detection," *arXiv preprint*, arXiv:2508.20449, Aug. 2025. arXiv+1

[9] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Learning," *arXiv preprint*, arXiv:2403.07240, Mar. 2024. arXiv

[10] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, S.-N. Lim, and Y.-G. Jiang, "M2TR: Multi-Modal Multi-Scale Transformers for Deepfake Detection," in *Proc. of ICMR '22: International Conference on Multimedia Retrieval*, 2022. Papers with Code

[11] H. Zhao, C. Xu, Y. Li, and J. Tian, "Multi-attention-based Approach for Deepfake Face and Expression Swap Detection and Localization," *EURASIP Journal on Image and Video Processing*, vol. 2023, no. 1, 2023. SpringerOpen

[12] Y. Qiao, R. Tian, and Y. Wang, "Towards Generalizable Deepfake Detection with Spatial-Frequency Collaborative Learning and Hierarchical Cross-Modal Fusion," *arXiv preprint*, arXiv:2504.17223, Apr. 2025. arXiv

[13] L. Alam, M. T. Islam, and S. S. Woo, "SpecXNet: A Dual-Domain Convolutional Network for Robust Deepfake Detection," *arXiv preprint*, arXiv:2509.22070, Sep. 2025. arXiv

[14] H. S. I. A. Sadruddin, A. Sardouie, and M. S. M. Saeed, "A Robust Ensemble Model for Deepfake Detection of GAN-Generated Images on Social Media," *Discover Computing*, vol. 28, no. 1, Art. 41, 2025, doi:10.1007/s10791-025-09538-w. SpringerLink

[15] Y. Qiao, R. Tian, and Y. Wang, "Towards Generalizable Deepfake Detection with Spatial-Frequency Collaborative Learning and Hierarchical Cross-Modal Fusion," *ArXiv*, 2025. (same as #12 but for completeness)

[16] A. Sardhara, V. Vekariya, and J. Tadhani, "A Hybrid CNN-LSTM Approach for High-Accuracy Image Forgery Detection," *International Journal of Engineering Sciences & Research Technology*, vol. 14, no. 2, pp. 720–728, 2025. [Online]. Available: https://theaspd.com/index.php/ijes/article/view/7206

[17] Mdpi, "Precision Deepfake Image Detection via Transfer Learning & CNN-LSTM," *Electronics*, vol. 13, no. 9, Art. no. 1662, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/9/1662

[18] V. M. Patel and S. Degadwala, "Deepfake Detection Using Convolutional Neural Networks and LSTM Modelling," *International Journal of Scientific Research and Technology*, 2024. [Online]. Available: https://ijsrst.com/index.php/home/article/view/IJSRST2512361

[19] D. Karishma, S. Umadevi, S. Teja, M. A. Shine, and N. I. Hasitha, "Deepfake Face Detection Using LSTM and CNN," *International Journal of Innovative Science and Advanced Engineering*, 2024. [Online]. Available: https://www.ijisae.org/index.php/IJISAE/article/view/7287

[20] K. Rohith, K. Nagarjuna, B. V. Yadav, and G. R. Chandra Kumar, "Face Morph Attack Detection Using LSTM-CNN Hybrid Model," *International Journal for Research in Applied Science & Engineering Technology*, 2024. [Online]. Available: https://ijraset.com/research-paper/face-morph-attack-detection-using-lstm-cnn-hybrid-model

[21] A. G. Singh and P. Sharma, "Hybrid Deep Learning Framework: CNN, LSTM & Vision Transformers for Deepfake Detection," *Journal of Emerging Science and Research*, 2025. [Online]. Available: https://journal.esrgroups.org/jes/article/view/9109

[22] Y. Shelar, P. Sharma, and C. S. D. Rawat, "An Improved VGG16 and CNN-LSTM Deep Learning Model for Image Forgery Detection," *International Journal of Research in Innovative Technology and Computer Science*, 2024. [Online]. Available: https://ijritcc.org/index.php/ijritcc/article/view/6157

[23] S. K. Sharma, W. A. Khan, and M. Kumar, "Estimation and Concealment Deep Fake Detection in Images using Hybrid LSTM," *International Journal of Innovative Science and Advanced Engineering*, 2024. [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/4295

[24] R. Anand, L. Santhosh, and A. K., "Video Authenticity Detection Using Web-Enabled Techniques (CNN, LSTM, ResNeXt)," *International Journal for Research in Applied Science & Engineering Technology*, 2024. [Online]. Available: https://www.ijraset.com/research-paper/video-authenticity-detection-using-web-enabled-techniques

[25] P. Saikia and D. Dholaria, "A Hybrid CNN-LSTM Model for Video Deepfake Detection by Leveraging Optical Flow Features," *arXiv preprint arXiv:2208.00788*, 2022. [Online]. Available: https://arxiv.org/abs/2208.00788