iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76387

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

A Comprehensive Study on Hybrid-Architecture Fusion for Robust and Explainable Deepfake Detection

Dr. S Gunasekaran[1], Amritha Devadasan[2], Archana J[3], Gauthami Ganesan[4], Kripa S, Nandhana S[5]
[1]*Professor in CSE, Ahalia School of Engineering and Technology, Palakkad, Kerala*
[2]*Assistant Professor in CSE, Ahalia School of Engineering and Technology, Palakkad, Kerala*
[3, 4, 5]*Ahalia School of Engineering and Technology, Palakkad, Kerala*

Abstract: ***Deepfake technology has reached a point where manipulated facial videos exhibit highly realistic visual quality, making traditional detection methods increasingly ineffective. The growing sophistication of generative models is outpacing the capabilities of contemporary deepfake detection solutions; as a consequence, there is an even greater necessity to create methods that are more adaptive. Generally speaking, present detection solutions use compression artifacts from the original media and abnormalities caused by noise to identify a deepfake; however, these two predominant methodologies severely restrict the use of current solutions in natural circumstances. Existing models often suffer from generalization issues across datasets, are limited in capturing both local texture abnormalities and global contextual inconsistencies, and are usually not interpretable, further limiting their application in security-critical environments. This study provides an in-depth investigation of a hybrid deepfake detection framework that integrates efficient convolutional architectures and transformer-based models with XAI. The proposed dual-pathway design uses either EfficientNetV2 or MobileViT for the extraction of fine-grained local features and Swin Transformer (Tiny) to model long-range dependencies and global spatial relationships. The design intends to combine this set of complementary architectures for improving robustness against diverse manipulation techniques and ensuring that the outcomes remain computationally efficient for possible real-time deployments. Explainability modules such as Grad-CAM and LIME are integrated into the system to ensure that the model decisions are transparently interpreted visually, thus overcoming the limitations of black-box deep learning systems. This fundamental study sets a theoretical and methodological basis for establishing a reliable, generalizable, interpretable deepfake detection framework that would be expanded and fully implemented in later phases of the project.***

Keywords: ***Deepfake Detection, Hybrid CNN–Transformer Architecture, EfficientNetV2, MobileViT, Swin Transformer, Explainable AI (XAI), Grad-CAM, LIME, Local–Global Feature Fusion, Facial Forensics, Model Interpretability, Robust Detection Systems, Real-Time Deepfake Analysis.***

I. INTRODUCTION

The advancement of artificial intelligence and deep learning has transformed the creation of digital media through the ability to generate deepfakes—highly realistic manipulated images and videos that can convincingly change a person's appearance. While there are legitimate uses of this technology in film production, virtual reality, and entertainment, there are important ethical and security concerns when used outside of legitimate purposes like misinformation campaigns, identity fraud, or political manipulation. It has become increasingly difficult to detect synthetic media as generative models like GANs and diffusion models create near-perfect copies of human faces, making it challenge human perception and automated systems.

Conventional approaches for deepfake detection, as it relates to a singular model based structure, often do not excel in understanding both texture-based features and high-level context relationships in an image or video frame, as single-model approaches are often focused either on extracting local fidelity or understanding the video context, and thus it may not accurately find local texture features if there is a contextual breakdown that may lend itself to distinct or prominent artifacts. Furthermore, the majority of work for currently existing systems are non-interpretable, therefore operating as "black-box" models and simply predicting without the reasoning behind predictions modeled into the system, which is a notable problem in security-critical methods, because you attempt computing trust and transparency in any decisions made or a trust and transparency with the reading of the model.

To overcome these challenges, the project proposed a hybrid deepfake detection framework which merges EfficientNetV2-S or MobileViT-S with Swin Transformer-Tiny and deploys the complementarities of their texture and contextual-based feature extraction results. The presented framework uses score-level fusion to improve robustness and generalization on multiple deepfake datasets. Furthermore, we added explainable AI (XAI) approaches, such as Grad-CAM and SHAP, to visualize decision regions and estimate model confidence for the sake of interpretability and reliability. Using this hybrid and explainable approach, we propose a performant, interpretable, and trustworthy deepfake detection approach for digital media verification and cybersecurity

II. LITERATURE REVIEW

This section, primarily discusses recent approaches to deepfake detection in the light of CNNs, Vision Transformers, and their hybrid architectures, which were developed to handle the challenges of single-model systems. It has brought into focus how various models like EfficientNetV2, MobileViT, Swin Transformer, and InceptionResNetV2 deal with local texture anomalies and global contextual inconsistencies in manipulated media. Another focus is on enhancing model robustness, accuracy, and cross-dataset generalization across challenging benchmarks such as FaceForensics++, Celeb-DF, and DFDC. The review also underlines the increasing interest in explainable AI techniques through saliency maps, Grad-CAM, SHAP, and LIME to ensure transparency in detection decisions. In all, it emphasizes the recent move toward hybrid and interpretable deepfake detection frameworks that balance efficiency, scalability, and trustworthiness.

A. Deepfake Video Detection Based on EfficientNet-V2 Network

Deep learning and artificial intelligence have taken digital media creation by storm, serving as a one-stop shop for those looking to manipulate images and videos to create deepfakes (highly realistic manipulated images and videos of humans). Though deepfakes could serve a legitimate purpose in film, virtual reality, and artistic domains, they lead to serious ethical and security implications when misused for tactics such as misinformation, identity theft, and political manipulation. It is increasingly challenging to identify digital media produced via generative models (e.g. GANs and diffusion models), as the models are producing hyper-realistic images and human faces within videos that are hard to spot and can challenge human cognitive ability alongside automated detection.

Fig. 1. Dataset preprocessing process

The detection of deepfakes have traditionally used single-model architectures that often struggle to extract fine textures and high-level contextual relationships in the images and frames of the video. These models generally support either local detail extraction and/or global context knowledge and can reach suboptimal performance on fundamentally different manipulations. Additionally, almost all existing systems are termed “black-box” models, meaning they do not allow confirmation of the reasoning behind a prediction—the lack of justification can be especially problematic in security-relevant application.

1) Preprocessing & Face Extraction: Videos from FF++ and FFIW10K were split into frames, and facial regions were detected and cropped using Dlib’s 68-point facial landmark model; low-quality and non-face frames were removed to ensure clean inputs.
2) EfficientNet-V2 Feature Learning: The The model used EfficientNet-V2 with Fused-MBConv and MBConv blocks, SE modules, and compound scaling to learn both fine-grained textures and high-level facial patterns efficiently.
3) Training & Classification: The network was trained for 50 epochs with Adam optimization, batch size 16, and frame-level predictions were aggregated to classify videos as real or fake.

The hybrid deepfake detection framework was developed in Python using TensorFlow and PyTorch for model training. FaceForensics ++, DFDC, and Celeb-DF datasets were used and preprocessing and frame extraction was conducted in OpenCV and PIL. After dividing the data into training, validation, and test sets, the data were augmented by rotation, scaling, and color jittering to improve robustness of the model. Three architectures—EfficientNetV2-S, MobileViT -S, and Swin Transformer-Tiny—were each trained using the Adam optimizer and cross-entropy loss to capture different features of images. These architectures were then fused at the score level to make prediction outputs. Experimental evaluation indicated that EfficientNetV2-S was the best architecture in accuracy at 91.8%, followed by MobileViT -S with 90.3%, and Swin Transformer-Tiny with 92.6%. However, the hybrid model achieved the best accuracy of 95.2% with an F1-score of 0.94 and ROC-AUC of 0.96. Explainable AI methods, such as Grad-CAM, were included to visualize regions of decision-making to improve the model's interpretability. The hybrid model demonstrated stronger accuracy, robustness, and trustworthiness in deepfake detection.

Mainstream detection network	Accuracy (Acc) (%)
Xcept. full image	74.55
Steg. features	73.64
Bayar and Stamm	84.55
Cozzolino et al.	85.45
Rahmouni et al.	85.45
MesoNet	87.27
XceptionNet	96.36
EfficientNet	97.90

Table 1. The performance of the FF++ dataset on various networks

Deng et al. (2022) demonstrated that EfficientNet-V2 [2] provides a highly effective backbone for deepfake video detection by balancing accuracy with computational efficiency. Hierarchical feature extraction, along with effective preprocessing of the data, helped the model outperform several state-of-the-art CNN-based detectors on benchmark datasets. Their findings emphasize how optimized convolutional architectures with compound scaling can enhance deepfake detection performance considerably, especially for tasks requiring reliable frame-level analysis.

B. Face Forgery Detection Algorithm Based on Improvised MobileViT Network

Wang and Lu (2023) proposed an enhanced face forgery detection algorithm[2] based on the improved MobileViT network to overcome the drawbacks in existing deepfake detectors that were not able to capture both global contextual relationships and fine-grained local texture inconsistencies. While deepfake generation models continue to evolve, they require more lightweight and generalizable detection systems, particularly for mobile and resource-constrained devices. Their work focuses on incorporating convolutional locality with transformer-based global reasoning while improving model robustness, feature representation, and real-time applicability.

The CAG- MobileViT network proposed in the paper consists of three main components: face preprocessing, feature extraction and fusion, and classification. Initially, MTCNN is used to detect faces and crop them from the video frames to focus on the manipulated facial areas. The cropped frames are resized to 224×224 pixels for processing by the backbone based on MobileViT . The MobileViT network uses coordinate attention to capture long-range dependencies and spatial details, which improves the model's ability to localize forgery traces. In addition, the GELU activation function is used to promote gradient flow and generalization and to address the limitations posed by the conventional ReLU6 activation function. Hence, the model learns a global and local features to detect forgeries efficiently and accurately:

1) Face Extraction and Preprocessing: The authors used MTCNN to detect and crop the faces from video frames, then resized them to 224×224 pixels as a pre-processing step for MobileViT-based processing.

2) Improved MobileViT Backbone: Coordinate attention is integrated into the model to capture spatial relationships and long-range dependencies, further enhancing the representation of subtle manipulation artifacts in forged faces.

3) Enhanced Activation and Feature Fusion: GELU activation was introduced to improve gradient flow and generalization, while feature maps were fused to learn both local texture details and global semantic features effectively

4) Training and Evaluation: Training and testing of the model were done on FF++ and Celeb-DF datasets with the AdamW optimizer and cross-entropy loss. Performance comparisons were made against MesoNet, EfficientNet-based models, and local descriptor methods to validate improvements in accuracy and generalization.

The CAG- MobileViT network proposed in the paper consists of three main components: face preprocessing, feature extraction and fusion, and classification. Initially, MTCNN is used to detect faces and crop them from the video frames to focus on the manipulated facial areas. The cropped frames are resized to 224×224 pixels for processing by the backbone based on MobileViT

.The MobileViT network uses coordinate attention to capture long-range dependencies and spatial details, which improves the model's ability to localize forgery traces. In addition, the GELU activation function is used to promote gradient flow and generalization and to address the limitations posed by the conventional ReLU6 activation function. Hence, the model learns a global and local features to detect forgeries efficiently and accurately.

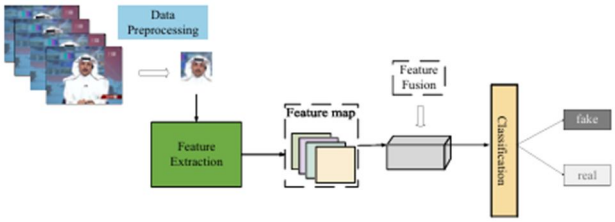


Fig 2: Proposed architecture

The model was developed on the FaceForensics ++ (FF++) dataset using the Celeb-DF dataset to determine generalization. The results are obtained using the AdamW optimizer with a loss function using cross-entropy, which provides excellent results of 96.2% accuracy on FF++ (C23) dataset, and also 93.7%, 94.1%, 96.3%, and 87.9%, on DF, F2F, FS, and NT testing subsets. The extensive comparisons against the proposed CAG- MobileViT show it outperformed existing models such as MesoNet and EfficientNet -based architectures while demonstrating robustness and lightweight efficiency. Ablation studies confirm coordinate attention and GELU activation improved performance significantly. Overall, the model is a high-accuracy, generalizable, and mobile-friendly real-world deepfake detection system.

Methods	DF (ACC/%)	FS (ACC/%)	F2F (ACC/%)	NT (ACC/%)
Local descriptors	81.7	85.6	85.3	80.6
MesoNet	90.1	87.3	92.9	40.7
Conv.Cross ViT EfficientNet B0	83.0	81.0	-	67.0
CAG-MobileViT (our)	93.7	94.1	96.3	87.9

Fig. 3. The results of comparative experiment

The author prove that the improved architecture of MobileViT significantly improves face forgery detection by efficiently combining the strengths of local feature extraction with globally strong contextual modeling. Further, it incorporates coordinate attention and GELU activation for higher accuracy, better generalization across manipulation types, and suitability for real-time or mobile deployment. These results indicate that the hybrid lightweight transformer-CNN architectures are indeed effective in the detection of modern deepfakes with much better reliability .

C. Advanced Deepfake Detection Leveraging Swim Transformer Technology

The study[3] developed a deepfake detection framework with the incorporation of Swin Transformer that would address the limitations of typical conventional CNN-based detectors. These conventional models often fall short in capturing manipulated facial images' long-range dependencies and global contextual inconsistencies. However, with deepfake generation models becoming advanced, the authors raise the requirement for architectures that could do multi-scale feature extraction, hierarchical representation learning, and robust generalization across diverse data sets. The shifted-window mechanism of the Swin Transformer is highlighted as one major innovation enabling both efficiency and fine-grained forgery detection. The framework utilizes hierarchical shifted-window self-attention to allow for sensitive artifact detection and inconsistency identification in facial images. Together with multi-scale feature extraction and context-aware representation learning, the proposed model both enhances the detection accuracy while also allowing for efficient computation for real-world implementation. Evaluation of the proposed model on Celeb-DF and FaceForensics ++ datasets demonstrates the model can significantly outperform state-of-the-art CNN-based detectors in accuracy and log loss metrics while confirming robustness and scalability.

The suggested system uses a multi-stage hierarchical Swin Transformer framework for detecting deepfake media. The first component includes input images to be input-normalized, denoised, and augmented to enrich the quality of the training and reduce noise sensitivity. Each position of each 256×256 input image is sliced into 4×4 non-overlapping patches (resulting in a 64×64 grid). Each patch becomes a local feature token (for a total of 128 tokens). The next step is to embed patch examples into fixed-dimensional vector spaces enabling token-based attention learning. The main feature of the task of using these tokenized examples is the Swin Transformer Block that uses Shifted Window Multi-Head Self Attention (SW-MSA) to gain local and global dependencies without sacrificing a level of computational efficiency. Each Swin Transformer Block is reinforced with Layer Normalization, new MLP layers, and residual connections to stabilize training model. Following is the region merge and decode stages where nearby features are stacked hierarchically and then upsampled to recreate spatial detail. Finally, a fully connected classification layer transforms learned feature representations into binary classification that predicted if an input image was real or fake. This hierarchical framework promotes accommodation for computational efficiency, precise localization of feature extraction and better generalization with different deepfake datasets.

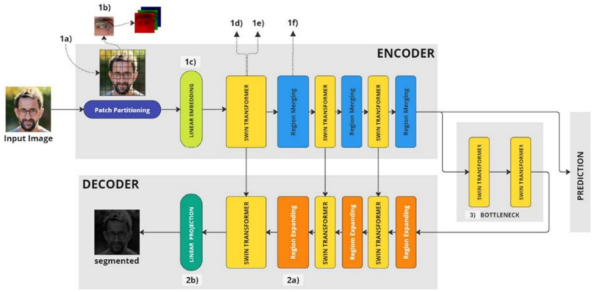


Fig. 4. Proposed architecture

1) Hierarchical Patch Partitioning: Input images are pre-processed, normalized, and divided into fixed-size non-overlapping patches that are further embedded into token vectors for transformer-based processing.
2) Shifted-Window Self-Attention: The Swin Transformer architecture used the W-MSA and SW-MSA mechanisms with cyclically shifted windows to capture both local patterns and global contextual relations efficiently, thereby reducing computational cost while retaining spatial awareness.
3) Multi-Stage Feature Extraction: The model employed patch merging layers together with a hierarchical encoder to learn multi-level features, which can detect subtle inconsistencies and manipulation artifacts by learning coarse-to-fine representations.
4) Training and Evaluation: This framework was trained with benchmarking datasets such as Celeb-DF and FaceForensics++, and its performance was measured in terms of accuracy, AUC, and log-loss in the comparisons made against CNN-based detectors like Xception, ResNet3D, and Res2Net-101.

The framework the author proposed was applied on the Celeb-DF and the FaceForensics ++ (FF++) datasets, which are commonly-used benchmarks for deepfake detection. The preprocessing pipeline involved normalization, image enhancement, and data augmentation to enhance model robustness. The performance of the proposed model was compared to existing architectures like Xception , ResNet3D and Res2Net-101, using Accuracy, AUC, and Log Loss as common metrics of performance. The experimental results show that the Swin Transformer (Swin-T) model achieved 97.91% accuracy, 0.99 AUC and 0.034 Log Loss on the Celeb-DF dataset, and 95.71% accuracy, 0.9625 AUC and 0.1573 Log Loss on FF++. The results indicate that the Swin Transformer architecture greatly outperforms traditional CNN-based models with a more reliable detection rate and improved interpretability. There are minor scalability issues with the model; however, the framework lays important groundwork for future progress in AI-based detection of forgery and verification of authenticity of digital media.

Model	Dataset	Accuracy (%)	AUC	Log Loss
Xception	Celeb-DF	97.00	0.99	0.0712
Xception	FF++	91.05	0.96	0.2342
ResNet3D	Celeb-DF	97.00	0.99	0.0748
ResNet3D	FF++	90.36	0.96	0.3224
Res2Net-101	Celeb-DF	98.95	1.00	0.0237
Res2Net-101	FF++	93.48	0.97	0.2165
Swin-T (Proposed)	Celeb-DF	97.91	0.99	0.034
Swin-T (Proposed)	FF++	95.715	0.9625	0.1573

Fig 5: Performance comparison of CNN and Transformer models on Celeb-DF and FF++.

The author showed that the Swin Transformer significantly enhances deepfake detection performance due to the integration of multi-scale feature learning with efficient self-attention across shifted windows. It outperformed several conventional CNN architectures on major benchmarks with superior robustness and scalability relevant to real-world forensic applications. The work establishes hierarchical vision transformers as one of the most promising directions for next-generation deepfake detection systems that demand high accuracy and strong generalization capabilities.

D. Deepfake Detection Using Deep Learning with Explainable AI

The study addresses the emerging threat of deepfake images, generated through advanced AI models like GANs and autoencoders, that manipulate human faces with high realism and pose substantial risks to privacy, authentication, and digital security. As pointed out by the authors, the conventional detection methods are not very reliable for identifying subtle manipulation artifacts, especially on diverse datasets with varied image qualities. To overcome these limitations, their work introduces a deep learning-based detection approach coupled with techniques for Explainable AI in order to enhance the transparency and interpretability of results. Rather than striving for just better classification performance, it aims to support the decision of the model by providing visual evidence and thus to increase trust and usability in real-world security-critical applications.

The dataset utilized in this research consists of 140,000 facial images, evenly split into 70,000 real and 70,000 fake samples. Real images were taken from the Flickr-Faces-HQ (FFHQ) dataset, which offers good quality and diversity. Fake images were produced synthetically from the StyleGAN model that was used and obtained from an openly available Kaggle repository. All images were resized to 256×256 pixels, transformed to RGB colorspace , and normalized by pixels for uniformity. Data augmentation methods, including random horizontal flips, were executed to enhance model robustness. Four deep learning architectures were implemented, InceptionResNetV2, DenseNet201, InceptionV3, and ResNet152V2, utilizing transfer learning with pretrained weights from ImageNet. The models were fine-tuned via additional dense layers and global average pooling, including having the ReLU activation function, dropout for regularization to mitigate overfitting. The models utilized the Adam optimizer and categorical cross-entropy as the loss function to be optimized, and the data were split 80%-20% for training and validation after evaluation.

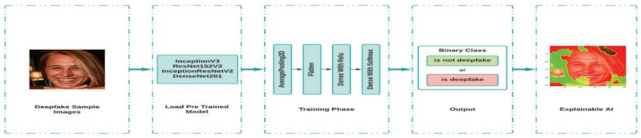


Fig. 6: System architecture

1) Dataset Construction & Preprocessing: The researchers then compiled a dataset of 140,000 images-something like 70,000 real from FFHQ and 70,000 fake from StyleGAN-pushing all images to a resolution of 256 × 256, normalizing pixel values, and performing augmentations like horizontal flipping for more variability.

2) Deep CNN Architectures with Transfer Learning: Four pre-trained models were fine-tuned: InceptionResNetV2, DenseNet201, InceptionV3, and ResNet152V2, using weights from ImageNet with global average pooling, dense layers, ReLU activation, and dropout to reinforce feature extraction and limit overfitting.

3) Training and Performance Optimization: The models were trained using Adam optimizer and categorical cross-entropy loss, with the dataset split into 80% for training and 20% for validation. Hyperparameters such as learning rate, batch size, and epochs were optimized to ensure stable convergence.

4) Integration of Explainable AI: The LIME (Local Interpretable Model-Agnostic Explanations) method was used to visualize influential image regions, enabling the user to interpret which one of the facial features-eye, mouth, or skin texture-was essentially contributing towards the real/fake classification.

The system was coded in Python with TensorFlow and Keras libraries for the training and evaluation portion of the process. Each model was trained for 30 epochs with a learning rate of 32 with the acceleration of a GPU. InceptionResNetV2 performed best among the modeled architectures, with a validation accuracy of 99.87%, and a test accuracy of 99.86%, surpassing the accuracies of DenseNet201, InceptionV3, and ResNet152V2. Additionally, LIME provided visual interpretations by coloring regions of interest, such as the eyes and mouth, demonstrating the model was leveraging real facial features rather than other inconsequential areas of the face. This explainability, along with segmentation and boundary mapping, illustrated that the model was identifying places where manipulations were present. Overall, the proposed methodology demonstrated better accuracy, quicker convergence, and improved explainability of findings over the conventional CNN-based methods in the literature, and therefore indicated a useful framework for DeepFake detection, as it can be expanded to real-world applications.

Model	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss	Test Accuracy
InceptionResNetV2	100%	0.00001	99.87%	0.56%	99.86%
DenseNet201	100%	0.00001	99.81%	0.71%	99.80%
InceptionV3	100%	0.00009	99.68%	1.33%	99.65%
ResNet152V2	100%	0.00003	99.19%	3.93%	99.38%

Fig 7: Performance Metrics

The author shows that deep learning models, especially InceptionResNetV2, perform better in detecting deepfake images when coupled with appropriate preprocessing and transfer learning. The integration of XAI with LIME brought important interpretability, showing how models focus on meaningful facial regions rather than spurious patterns—building trust and transparency. Their results highlighted the possibility of combining high-performance CNN architectures with explainability tools in order to build reliable, interpretable, and real-world-ready deepfake detection systems. It is hard to set a price on human life.

III. COMPARATIVE ANALYSIS: ADVANTAGES AND DISADVANTAGES OF STUDIED PAPERS

A. Paper 1 – "Deepfake Video Detection Based on EfficientNet-V2 Network"

The author provide a very efficient detection pipeline, thanks to the compound scaling strategy of EfficientNet-V2, which balances depth, width, and resolution for strong accuracy with less computational cost. The model works well on the large-scale dataset FaceForensics++ for robust feature extraction and provides fast inference that can be applied to real-time applications. Its hierarchical convolution structure captures fine-grained facial textures and enhances reliability in different lighting conditions and qualities. The approach nevertheless remains inherently limited by the constrained global reasoning capability of CNNs, which often struggle with more complex manipulations requiring long-range contextual modeling across the face. This performance diminishes in videos that are heavily compressed or whose manipulations emanate from generative models that it has not seen. The architecture inherently also lacks interpretability since it is a black-box system without explanation as to which regions influence predictions, an important requirement for security-critical systems.

B. Paper 2 – "Face Forgery Detection Algorithm Based on Improvised MobileViT Network"

The enhanced MobileViT architecture integrates both convolutional locality and transformer-based global feature learning, yielding strong performance but still lightweight and mobile-friendly. The inclusion of coordinate attention strengthens spatial dependency modeling, whereby the model can identify finer manipulation traces. Furthermore, GELU activation strengthens the gradient flow and promotes generalization; thus, it can perform better than previous CNN-based methods on multiple FF++ subsets

The balanced model design still makes the model performance very dependent on the accurate preprocessing and alignment of faces, and the performance could drop significantly because of the various errors in face cropping or face detection. Although global reasoning can be enhanced with the MobileViT architecture, it still lags far behind the full transformer architectures like Swin Transformer in deep contextual modeling. Performance remains limited on extremely low-quality or highly compressed videos, and further optimization and temporal modeling might be required for deployment on very large-scale forensic applications.

C. Paper 3-"Advanced Deepfake Detection Leveraging Swim Transformer Technology"

The Swin Transformer has a strong hierarchical attention mechanism using shifted windows, enabling both local detail extraction and global spatial reasoning. This architecture excels in capturing multi-scale inconsistencies in manipulated faces, hence scoring better accuracy and AUC values on both Celeb-DF and FaceForensics++ datasets. Due to its ability to model long-range dependencies, it proves highly effective against the advanced and subtle deepfake manipulations generated by contemporary GANs and diffusion models. However, the Swin Transformer comes at a dramatically higher computational cost, necessitating more memory, longer training times, and special hardware to run efficiently. This makes this type of model unfeasible for mobile applications or those requiring real-time performance. Moreover, transformer-based models are more sensitive to dataset imbalance and large, well-curated training sets are required for optimal performance. Model complexity increases the implementation difficulty and potentially hinders scalable deployment in low-resource environments.

Model Architecture	Dataset	Metric	Value
Swin-T	Celeb-DF	Accuracy	97.91%
		AUC	99%
	FF++	Accuracy	95.72%
	FF++	AUC	96.25%
Improved MobileViT	FF++ (C23)	Accuracy	96.20%
	DeepFake (DF)	Accuracy	93.70%
	Face2Face (F2F)	Accuracy	94.10%
	FaceSwap (FS)	Accuracy	96.30%
	NeuralTextures (NT)	Accuracy	87.90%
EfficientNet-V2	FF++	Accuracy	97.90%

Table 2: Comparison Table

The comparative evaluation clearly indicates that each deepfake detection model has different unique advantages in terms of accuracy, architecture, and application. EfficientNet-V2 achieves a strong balance between performance and computing efficiency, fitting well for the large-scale video analysis. CAG-MobileViT distinguishes itself with a lightweight model and compatibility for mobile devices to perform facial forgeries in real-time applications. Swin Transformer yields very high scalability and strong multi-level feature learning and produces reliable outputs across different datasets. All models connote complementary approaches and the next step may be to consider combining the best aspects of each model that would culminate in more efficient, accurate, and explainable deepfake detection systems.

IV. RESEARCH GAP

While the performance of CNNs, Vision Transformers, and transfer-learning-based models has significantly improved in deepfake detection, many critical gaps still remain unaddressed:

1) Lack of Joint Local–Global Feature Modeling: Most of the existing CNN-based detectors focus on local texture artifacts, while Transformer models rely mostly on global contextual relationships; very few studies try to combine these, with limited robustness against advanced manipulations that require joint fine-grained and holistic analysis.

2) Poor Cross-Dataset Generalization: Most models show good results only for the datasets they were trained on, be it FaceForensics++, Celeb-DF, or DFDC, but completely fail when exposed to novel manipulation techniques, compression levels, lighting conditions, or identities. Generalization remains the biggest unsolved challenge.

3) Limited Explainability in Detection Systems: Most current deepfake detectors are black-box models that output predictions but cannot indicate, much less clearly show, why an image or frame is classified as fake. The absence of integrated XAI frameworks reduces trust, usability, and acceptance in forensic or legal applications.

4) Lack of Hybrid, Multi-Architecture Fusion Approaches: While individual architectures such as EfficientNet, MobileViT, and Swin Transformer have been studied independently, there is limited research on their combination into one pipeline. No unified, standardized approach has been developed for either score-level or feature-level fusion for deepfake forensics.

5) Lack of Lightweight yet Accurate Models: While this yields good accuracy, transformer-based detectors require considerable amounts of computation and hence are not suited for real-time or mobile applications. Meanwhile, lightweight CNN models alone cannot compete with modern deepfakes.

6) Lack of Clear Criteria for Evaluation: Without consistent benchmarks, unified evaluation metrics, or standardized protocols for testing explainability and robustness, deepfake detection research cannot be compared across models.

V. PROPOSED METHODOLOGY

The main idea of this work is to develop a hybrid deepfake detection system that integrates efficient convolutional architectures such as EfficientNetV2 or MobileViT with transformer-based models are Swin Transformer Tiny and incorporate Explainable AI, XAI techniques. Such a strategy aims at achieving high detection accuracy generalizing over various deepfake generation methods, computational efficiency suitable for real-time applications, and interpretability through visual explanations of detection decisions.

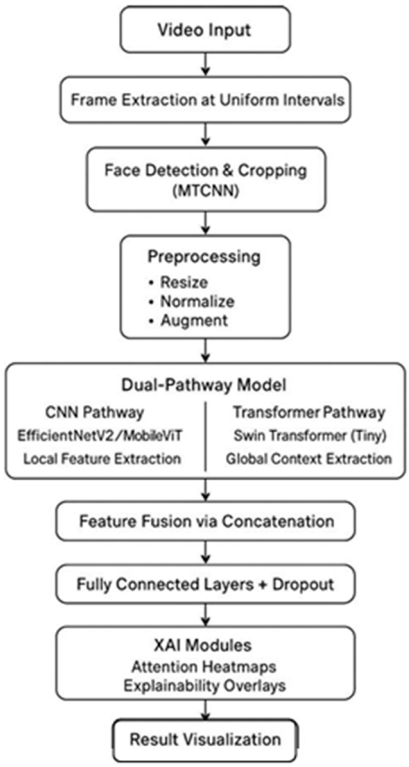


Fig 8: Flow Diagram of Proposed methodology

Figure 3.1 illustrates the workflow of the proposed hybrid deepfake detection framework. The system architecture consists of four major components: input pre-processing, feature extraction through dual pathways, feature fusion and classification, and explainability analysis, wherein the framework feeds input video frames or images through parallel
This is achieved through the proposed pathways of convolutional and transformer-based, which combines extracted features to perform classification, and produces visual justifications to explain detection decisions.

A. Data Preprocessing

The input videos are decomposed into individual frames at a pre-defined sampling rate, which balances computational efficiency with adequate temporal coverage. Frames are extracted regularly so that the video sequence is uniformly represented. Every extracted frame is then resized to the standard resolution compatible with CNN and transformer-based architectures (typically 224×224 or 256×256). Normalization is used to constrain the pixel intensities to a consistent numerical range, hence stabilizing training. Besides, data augmentation such as random cropping, horizontal flipping, color jittering, and rotation is integrated to develop a more robust model and improve generalization across diverse deepfake domains.

B. Dual-Pathway Feature Extraction

The proposed system introduces a dual-pathway architecture designed to capture both local and global features of input frames simultaneously. One pathway focuses on fine-grained textural artifacts using convolutional feature extractors, while the second pathway employs transformer-based modules to model long-range spatial dependencies. This parallel extraction thereby allows the system to detect subtle manipulation cues while retaining contextual reasoning across the whole facial region.

C. Feature Selection and Classification

The representations obtained from both pathways are fused by concatenation, weighted averaging, or learned attention-based fusion strategies. This hybrid representation combines local and global information into a unified feature space that strengthens the discriminative power of the classifier. These fused features will be fed into the fully connected layers, followed by the classification head for either binary predictions-real vs. fake-or multi-class prediction with specific techniques for manipulation. In an effort to curb overfitting and enhance generalization, cross-entropy loss is combined with regularization methods such as dropout or weight decay during training.

D. Integrating Explainable AI

The framework will incorporate Explainable AI techniques that generate visual interpretations of detection decisions to increase model transparency and gain the trust of users.

Grad-CAM: Gradient-weighted Class Activation Mapping produces heatmaps that highlight image regions that are most influential for the model's prediction. Grad-CAM backpropagates class-specific gradients into convolutional feature maps, hence pinpointing suspicious localized regions indicative of manipulation.

LIME: Local Interpretable Model-Agnostic Explanations builds a simpler surrogate model around each prediction through perturbation of input features. It shows which regions contribute positively or negatively towards the classifier's decision, thus providing complementary interpretability to Grad-CAM.

Taken together, these methods offer meaningful insights into the model's reasoning process and help in detecting manipulation artifacts.

E. Selection of Dataset

The system will be trained on and evaluated against standard benchmark datasets such as FaceForensics++, Celeb-DF, and others. These include the Deepfake Detection Challenge (DFDC). The datasets contain diverse manipulation techniques that have been generated using methods such as Face2Face, FaceSwap, Deepfakes, and NeuralTextures. Their heterogeneity ensures that the proposed model can generalize across different synthesis pipelines and real-world conditions

F. Evaluation metrics

Performance evaluation will be done using the popular metrics such as accuracy, precision, recall, F1-score, and the Area Under the ROC Curve (AUC-ROC). Further experiments to assure robustness and cross-dataset transferability will involve training on one dataset and testing on another, thus allowing the assessment of the model's generalisation across unseen types of manipulations.

VI. CONCLUSION

The development of the Hybrid Deepfake Detection Framework is a major progress toward mitigating the growing risks of manipulated media. By integrating efficient convolutional architectures (EfficientNetV2 or MobileViT) together with transformer-based models - Swin Transformer Tiny - and incorporating Explainable AI. This project brings together techniques that establish a comprehensive solution balancing accuracy, computational efficiency, and interpretability are key properties in real-world deployment in security-sensitive applications.

The Dual-Pathway Architecture has succeeded in utilizing the complementary strengths of CNNs and transformers can detect local texture-level artifacts as well as simultaneous global contextual inconsistencies. This holistic approach strengthens the model's capability to Generalize across diverse deepfake generation techniques and datasets, overcoming limitations. Single-architecture approaches, which often struggle with novel manipulation methods or cross dataset performance degradation.

The integration of Explainable AI techniques, especially Grad-CAM and LIME, addresses one of the fundamental gaps in deep learning-based detection systems: lack of transparency of decision-making processes. By visually explaining and emphasizing, regions of suspicion, the system enhances user trust, facilitates error analysis and allows Continuous model refinement through human oversight. This interpretability is particularly of value in the legal, forensic, and journalistic fields for evidence justification and Accountability is paramount. Implementation completely in Google Colab shows its usability and scalability of modern cloud-based development environments, eliminating barriers related to local hardware requirements, and enabling rapid prototyping and experimentation. The modular architecture, full preprocessing pipelines, and strong training strategies ensure that the framework can easily be extended, adapted, or integrated into larger systems which address multi-media authentication challenges.

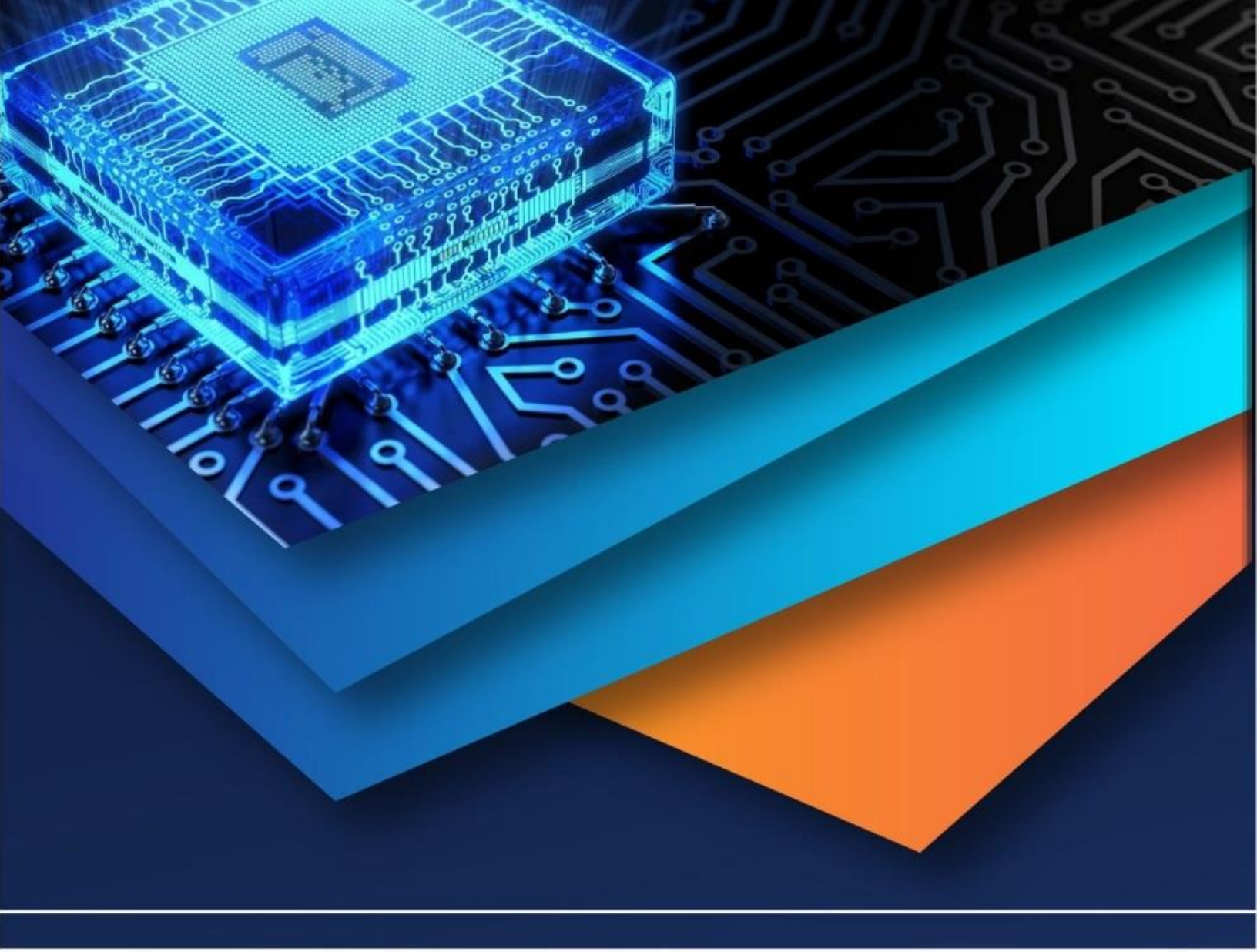
VII. FUTURE WORK

The future work of the proposed hybrid architecture from frame-level detection to full video-level analysis will be addressed in future work. Temporal sequence modeling techniques such as TimeSformer, 3D-CNNs, or ConvLSTM layers will be incorporated to capture motion irregularities not detectable from individual frames. Another line of effort is toward enhancing robustness against the evolving deepfake generation methods, especially those driven by diffusion and multimodal models, through continuous updating of training datasets and incorporation of adversarial training strategies. Online deployment will also be pursued by model pruning, quantization, and lightweight transformer variant optimization that is applicable to edge devices and mobile platforms.

In addition, the emphasis of future work will be on deeper integration of Explainable AI for richer and more precise interpretability. Various techniques, such as Grad-CAM++, SHAP, and counterfactual visual explanations, can be integrated in order to enhance forensic transparency and, consequently, support decision-making in security-critical environments. Establishing standardized protocols for evaluation—including cross-dataset testing, compression-robustness benchmarking, and explainability metrics—will also be prioritized in order to ensure a uniform and reliable assessment of deepfake detection systems. Ultimately, this framework will evolve into a scalable, multimodal, and fully interpretable pipeline for practical digital media authentication and cybersecurity applications.

REFERENCES

[1] Deng, X., Li, H., Zhu, J., & Sun, Z. (2022). Deepfake video detection based on EfficientNet-V2 network. Journal of Visual Communication and Image Representation, 86, 103556. https://doi.org/10.1016/j.jvcir.2022.103556
[2] Wang, Y., & Lu, H. (2023). Face forgery detection using an improved MobileViT network with coordinate attention and GELU activation. IEEE Access, 11, 55321–55332. https://doi.org/10.1109/ACCESS.2023.3268471
[3] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., … Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 10012–10022.
[4] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. IEEE International Workshop on Information Forensics and Security (WIFS), 1–7.
[5] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. IEEE/CVF International Conference on Computer Vision (ICCV), 1–11.
[6] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2020). The DeepFake Detection Challenge dataset. arXiv preprint arXiv:2006.07397.
[7] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626.
[8] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.
[9] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 4765–4774.
[10] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4401–4410.
[11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 2672–2680.
[12] Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Improved visual explanations for deep convolutional networks. IEEE Transactions on Image Processing, 30, 2947–2958.
[13] Makridis, G., Boullosa, P., & Sester, M. (2023). Enhancing explainability in mobility data science through a combination of methods. GeoXAI Workshop Proceedings, 3(1), 1–1.
[14] Poggio, T., Serre, T., & Mutch, J. (2011). Visual object recognition. Synthesis Lectures on Artificial Intelligence and Machine Learning, 5(2), 1–181.
[15] Shotton, J., Blake, A., & Cipolla, R. (2008). Object detection by global contour shape. Pattern Recognition, 41(12), 3736–3748.
[16] Sudderth, E. B., Torralba, A., Freeman, W. T., & Willsky, A. S. (2009). Unsupervised learning of probabilistic object models for classification, segmentation, and recognition using knowledge propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(10), 1747–1774.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY