Natural Resources for Human Health

eISSN: 2583-1194

Original Research

Received 28 May 2026
Revised 22 June 2026
Accepted 17 July 2026

Natural Resources for Human Health 2026, Vol. 6(9s): 1468–1482

KEYWORDS:
Image processing, Vision Transformer (ViT), Convolution neural network (CNN), Deep learning, Face detection.

Hybrid Deep Learning Model for Fake Image Detection with Advanced Face Object Segmentation Method

Saurabh Kumar Jain[1], Mohd. Akbar[2]

[1]Department of computer science and engineering, Integral university, Lucknow
sjain@student.iul.ac.in
[2]Department of computer science and engineering, Integral university, Lucknow
akbar@iul.ac.in

ABSTRACT: The fast development of deepfake technologies for image generation produces more and more realistic manipulated facial images that are harder to distinguish from the real content. In this paper, we propose a novel Hybrid CNN–Vision Transformer (HybCNNViT) framework for robust deepfake image detection by combining discriminative and generative learning mechanisms. The proposed method presents a new dual-stream architecture that separately encodes inner facial components (eyes, nose and mouth) and outer facial regions with Variational Autoencoder (VAE) and Autoencoder (AE) modules respectively. Also, a hybrid CNN-ViT and CNN-Swin Transformer architecture is used to learn local visual artefacts and global contextual dependencies. In addition, a dedicated preprocessing scheme such as background removal and inner-outer facial object segmentation is further integrated to improve feature relevance and reduce computational complexity. The experimental results on real and fake face datasets under easy, medium and hard manipulation scenarios show the effectiveness of the proposed framework. The proposed Inner–Outer Segment Face Object + HybCNNViT model attains superior accuracy of 86%–88% and outperforms the existing CNN–ViT, ViViT, and CNN–ViTEfficientNet approaches. The results indicate that the combination of facial-region segmentation, latent feature learning and hybrid Transformer architectures enhances the robustness and generalisation capability of deepfake detection significantly.

1. INTRODUCTION

The rapid development of technologies capable of altering audio, video, and image content has drastically transformed the creation of digital media [1]. At the same time, the public has greater access to the knowledge and resources needed to create and modify multimedia content. With low computational resources and the abundance of online tutorials and guidance materials [2–4], highly realistic synthetic images can now be created remarkably easily.

Deepfake technology is a digital manipulation technique that substitutes the face of a target person with another person's in aimage [5]. The process is usually achieved by integrating a synthetically generated face region into a genuine image or video frame [6]. The end result of such processes – a hyper-realistic synthetic media – is also often called "Deepfake." Besides malicious uses, Deepfake technology can be used for many other beneficial applications, such as generating realistic CGI, VR environments [7], AR, educational content, animation, art, and cinematic productions [8].

These advantages notwithstanding, the deceptive nature of Deepfakes has raised serious concerns about their misuse. As a consequence, a lot of research efforts have been dedicated to differentiating manipulated videos from real ones. Joshual et al. [9] noted that while many detection frameworks are effective under certain conditions, most existing approaches have limited generalisation capabilities. They found that most detection systems are built by using artefacts of specific methods of Deepfake generation, rather than learning characteristics of forgery that can be used more generally.

A number of researchers have tried to identify Deepfakes based on physiological inconsistencies. For instance, detection techniques based on irregular eye-blinking patterns are proposed by Yuezun et al. [10] and TackHyun et al. [11]. However, later studies by Konstantinos et al. [12] and Hai et al. [13] showed that such physiological cues can be artificially synthesised themselves. In particular, the framework proposed in [12] produces talking-head videos with realistic facial behaviours, such as natural eye blinking. Similarly, the model in [13] allows generating facial expressions from a single portrait image, making static photographs display emotions and simulated blinking movements.

The motivation of the current work is to address two major limitations of existing Deepfake detection systems identified in previous studies [9, 14], which are inadequate data preprocessing and insufficient generalisation ability. Polychronis et al. [14] found that many of the current Deepfake detectors focus mainly on architectural innovations while relatively ignoring the preprocessing strategies and their impact on the overall detection performance. Their results highlighted the importance of data preparation for reliable Deepfake classification. In the meantime, Joshual et al. [9] examined the generalisation capabilities of facial forgery detectors and concluded that the majority of existing methods are unable to maintain robust performance across various manipulation techniques. They defined generality as the ability to correctly identify multiple forgery types while still being robust to unseen manipulation techniques.

To tackle this challenge, Umur et al. [4] proposed FakeCatcher, a generalised Deepfake detection framework that leverages biological signals and internal representations generated during the image synthesis process. Their approach used a relatively simple three-layer Convolutional Neural Network (CNN) and a dataset of 3000 videos was used for evaluation. However, the authors do not provide detailed information about the preprocessing procedures applied to the dataset. Moreover, the results in [15] show that deeper CNN architectures tend to perform better than shallow networks in image classification problems. This observation suggests that there is still opportunity to develop a more robust generalised Deepfake detector, combining comprehensive preprocessing techniques with a deep neural architecture that can capture a wider range of forgery artefacts.

With these motivations, we propose a generalised Convolutional Vision Transformer (CViT) framework for Deepfake video detection. The proposed model combines the feature extraction capabilities of Convolutional Neural Networks with the contextual learning abilities of Transformer architectures. We consider our method to be generalised for three main reasons. First, the hybrid architecture captures local spatial information in CNN layers and global contextual dependencies in the Transformer attention mechanism simultaneously, which helps achieve richer feature representation. Second, equal importance is given to data preprocessing and model development, ensuring that image quality enhancement and artefact preservation contribute effectively to the learning process. Third, the framework is trained on a diverse set of face images taken from publicly available datasets, which enables it to detect Deepfakes created under different levels of image perturbations ranging from easy to moderate to highly challenging manipulation scenarios.

2 RELATED WORKS

The progress of technology in the field of artificial intelligence still has a significant impact on the process of digital picture creation and editing. Recent advances in generative learning frameworks allow the generation of very realistic synthetic images, which are hard to distinguish from real visual content. These improvements have many applications but also raise serious concerns about the proliferation of false information, breaches of privacy, truthfulness of content and public trust in digital media. Hence, the research efforts have been largely directed towards the development of sophisticated deepfake and AI-generated images and the development of efficient methods for the detection of altered visual information. This section provides an in-depth review of the recent advances in both fields, focusing on important technical breakthroughs, limitations, and new research challenges reported in the literature.

2.1 Deepfake Image Generation

The development of generative artificial intelligence has had a significant effect on the realism, visual quality, and flexibility of synthetic picture generation. While many techniques exist, diffusion-based frameworks are the preferred paradigm, regularly outperforming traditional Generative Adversarial Network (GAN) designs in image quality, variety, and stability. Saharia et al. [17] presented Imagen, a powerful text-to-image generation model that uses diffusion processes and large-scale language understanding to generate photorealistic visuals. In a related work, Rombach et al. [23] introduced Latent Diffusion Models (LDMs) which operate in a compressed latent space and generate visually pleasing images while drastically reducing computational requirements

Table I: Summary of deepfake image generation and detection studies

Techniques/ Models	Major Contribution	Limitations
SEGAN [16], Imagen [17], PixArt-Σ [19], Latent Diffusion Models [23]	Improved diversity, photorealism, and high-resolution image synthesis.	High computational requirements and potential misuse for fake content generation.
OmniGen [18], FastComposer [20], Pick-a-Pic [21], Rectified Flow Transformer [22]	Personalized generation, unified frameworks, scalable image synthesis.	Limited interpretability and increasing difficulty for forensic detection.
ELA-CNN [25], ResNet50 [30], Noise Pattern Analysis [32], Synthbuster [34]	Effective detection of manipulated and diffusion-generated images.	Performance degrades on unseen generation models and compressed images.
DNA-Det [24], GPT-assisted Detection [26], Multimodal Models[27,31]	Architecture attribution, multimodal verification, good content analysis.	Dependence on datasets and limited generalization across domains.
CNN-RNN [28], Transfer Learning [29], DeepFake Doctor [33]	Robust detection of manipulated audio-video deepfakes.	High complexity and sensitivity to video quality variations.

Similarly, Chen et al. [19] proposed PixArt-Σ, a diffusion-transformer-based architecture designed to generate ultra high-resolution 4K pictures with improved computational efficiency and semantic consistency. These contributions together show the fast development of picture creation techniques and their growing ability to produce artificial material which looks like real images.

Recent works have been focused on improving visual realism, but also on increasing variety, personalisation and scalability in generative frameworks. To solve the problem of mode collapse that often occurs in GAN-based systems, Zuo et al. [16] proposed a conditional picture synthesis method based on conditional mutual information maximisation, which achieved enhanced output diversity and improved controllability. Xiao et al. [20] propose a system called FastComposer with localised attention processes, allowing for customisable multi-subject picture generation without further fine-tuning. In addition, Esser et al. [22] propose rectified flow transformers that outperform many state-of-the-art diffusion-based models in terms of textual understanding and image generation quality. These developments have opened up enormous possibilities for AI-driven picture synthesis, but they have also raised concerns about the unauthorised use, manipulation and dissemination of artificial visual information.

2.2 Deepfake Image Detection

The growing sophistication of synthetic picture generating methods makes the need for reliable and efficient deepfake detection solutions more pressing. As a result, deep learning-based methods have become a hot research topic and often outperform traditional picture forensic methods in detecting manipulated visual content. Rafique et al. [25] proposed a detection system to distinguish between real and fake photos that is based on Error Level Analysis (ELA), convolutional neural networks and machine learning classifiers. Similarly, Hamid et al. [30] proposed an architecture based on ResNet50 and used deep feature extraction to record the minute traces of manipulation resulting in highly accurate classification performance. These findings show that deep neural networks are capable of accurately recognising both digitally altered and AI-generated images.

Recent studies have focused on frequency-domain and multimodal analysis techniques for enhanced detection resilience and adaptability to different forging processes. Liu et al. [32] proposed an image verification system based on frequency domain noise characteristics as forensic evidence and demonstrated good detection performance on different picture production frameworks. Similarly, Bammey [34] developed a forensic method called Synthbuster that identifies diffusion-generated pictures by using frequency artefacts created during the diffusion process. Another notable contribution was made by Uppada

et al. [31], who proposed a multimodal detection framework that combines textual and visual information, enabling more comprehensive content analysis and surpassing unimodal approaches. While these approaches have been promising, it is still very difficult to guarantee consistent detection performance against generative models that have not been tested or are newly developed. This limitation emphasises the need for more flexible, future-proof and generalised forensic systems to cope with the fast-evolving image synthesis technology.

Although considerable progress has been made in recent years, many detection systems still suffer from poor generalisation across a wide range of datasets, modification methods and real-world deployment settings. Table I lists the main contributions and related drawbacks of the deepfake detection techniques currently in use for training and assessment. The following section presents the proposed methods to address the identified research gaps and improve the robustness and generalizability of deepfake image detection.

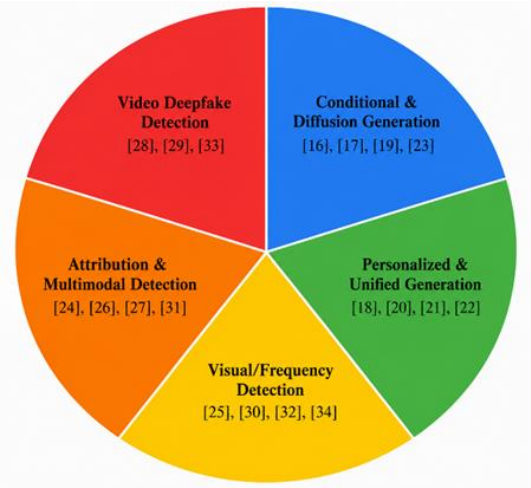

Figure 1: Summary for describing distribution research areas for different articles covered under related works.

3. DATA PRE PROCESSING

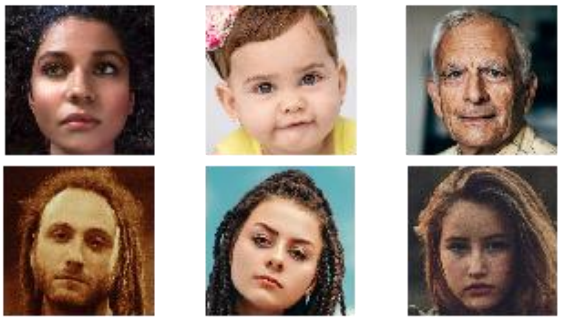

Figure 2: Images showing real face dataset

Figure 3: Images showing fake face dataset

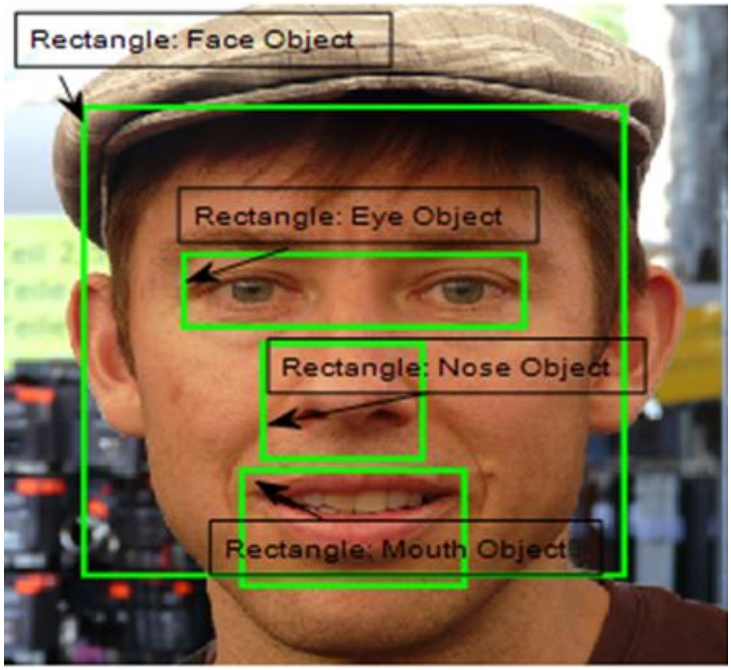

Figure 4a: Detected nose, eye and mouth objects (rectangle) from a sample of face image.

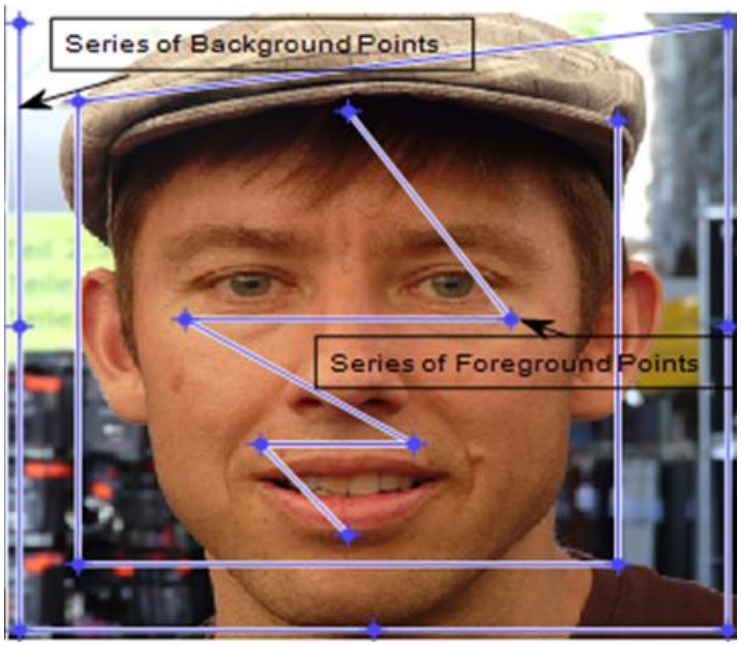

Figure 4b: Representation of the background area points for segmenting the face region of interest.

Examples of real and modified face photos collected in the dataset are shown in Figures 2 and 3. The dataset is divided into two categories containing 668 fake face photos and 779 real face photos. The dataset is then split into three difficulty levels for human recognition: Easy (240 photos), Medium (178 images) and Hard (240 images) to evaluate detection performance for different levels of visual complexity.

The modifications in the Easy category of the pictures are comparatively easy to spot as only one face feature is manually changed by professionals such as one eye, the nose or the mouth. Medium category images display more significant alterations that affect many regions of the face, especially the mouth and nose. In contrast, the images in the Hard category have both eyes changed, so visual recognition is much more difficult. The progressive framework enables to evaluate detection models on different levels of facial forgery complexity.

The data set was collected by the Computational Intelligence and Photography Laboratory (CIPLAB), Department of Computer Science, Yonsei University. It features a collection of great face photos expertly manipulated with picture editing techniques. The fake photos had altered the eyes, nose, mouth or full face area, and combined other facial features from many people. The editing process expertise makes the dataset a useful baseline to evaluate the robustness and generalisation capabilities of deepfake and face-forgery detection algorithms. The dataset is publicly available on the internet repository.

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

Fig 5a. Border Lines extraction

Fig 5b Removal of Background

Fig 5c. Image with Eye, Nose, Mouth object

Fig 5d. Image without Eye, Nose, Mouth object

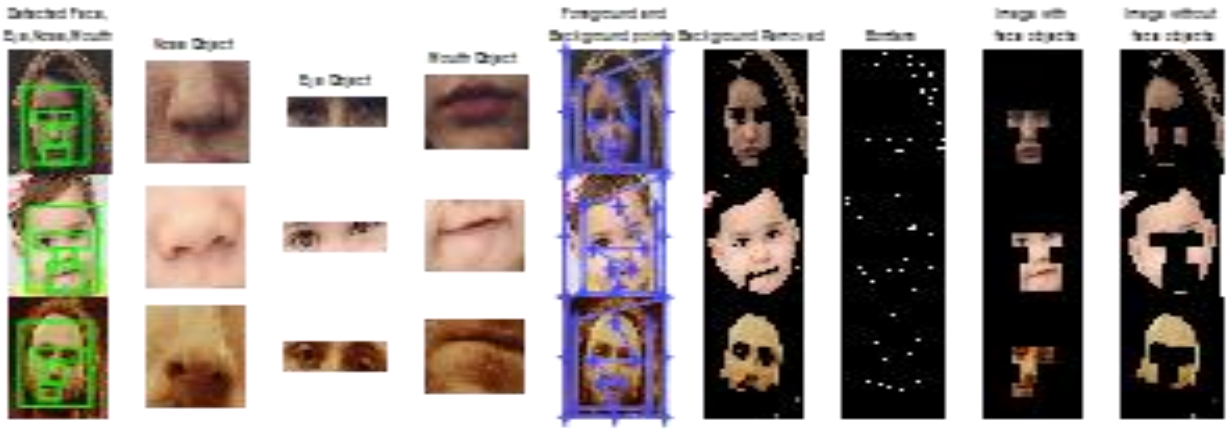

Figure 6. Data preprocessing and segmentation on real faces training data to obtain images with and without face.

In the changed facial images, various types of tampering are present, such as the replacement or change of one facial component or multiple facial regions at the same time. The eyes (one or both), nose, mouth and combinations of these (eyes and nose, nose and mouth, etc.) are often altered. They do this by taking these facial features and replacing them with similar

areas from other people to make composite face pictures. Such modified samples are shown as representative examples in Figure 2. The collection contains real and fake face photos with random orientations and rotations taken from frontal and side views. Beards, moustaches, haircuts, hats and other items that partially obscure salient facial features also contribute to substantial changes in face look. The dataset shows highly variable lighting conditions resulting in large variations in illumination. Thus, the dataset is a challenging benchmark for face fraud detection due to the high intra-class variability, irregularity and lack of visual consistency.

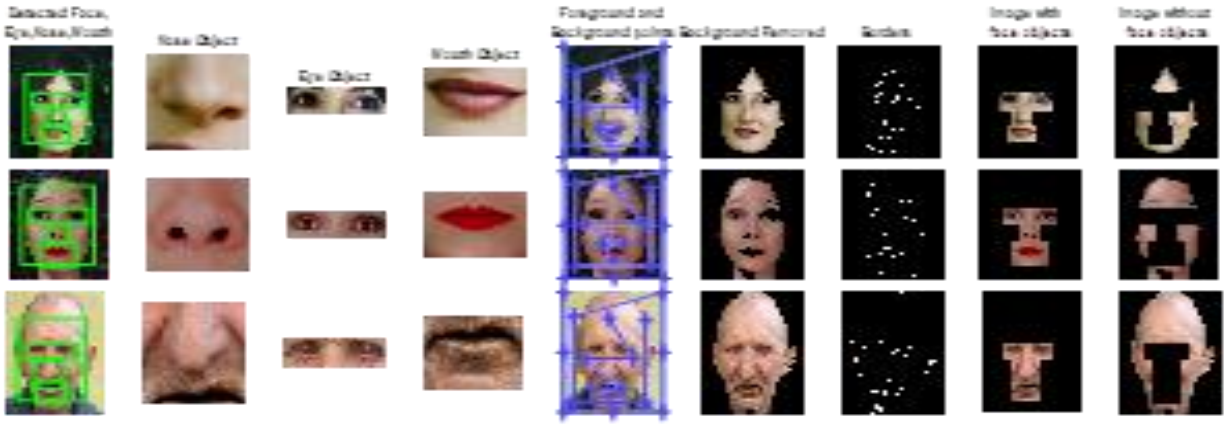

Figure 7. Data preprocessing and segmentation on fake faces training data to obtain images with and without face

In order to extract the efficient features each image from real face and fake face category is subjected to three stage of object identification. The procedure is capable of finding the bounding sections corresponding to mouth, nose, eyes and face as shown in the left part of Figure 3. During detection, many candidate regions may be identified as eye, nose or mouth objects, which are incorrect detections. To tackle this issue, a location-based selection technique is used for selecting the best face component. If multiple nose candidates are found, the bounding box whose center is closest to the midpoint of the lower boundary of the eye region is selected as the legitimate nose object. In the case of finding many mouth candidates, the mouth object is selected as the one closest to the observed nose position and below the nose bounding box. This technique of spatial filtering improves the precision and reliability of face component localisation.

The foreground facial area and the surrounding backdrop are automatically segmented after all the facial areas are successfully detected as shown in Figure 3 (right). A specialised MATLAB image-processing technique is employed to segment the facial region and remove background material and other superfluous visual information. The segmentation procedure is based on automatic detection of important facial landmarks such as the midpoint of the lower lip border, eye corner points, nose coordinates and the upper centre point of the detected face. These points in combination define the face contour used for foreground extraction.

Some representative examples of real and tampered facial photos after segmentation and preprocessing, ready to be fed into the ViT learning system, are shown in Figures 6 and 7. In some cases, important facial features such as mouth, nose or eyes are accidentally removed during the background removal step, leading to loss of important discriminative information. This problem is then solved by blending the segmented face picture with the independently derived facial component areas. This superimposition procedure retains essential information at the pixel level and enhances the quality of data provided to the learning model by restoring significant face features that would have been lost during automatic backdrop removal.

4. METHODOLOGY

The proposed deepfake detection system is based on the Hybrid CNN–Vision Transformer (HybCNNViT) architecture. The system learns discriminative visual artefacts through hidden feature connections and projects face images into latent feature spaces to classify movies as real or tampered. The architecture of the proposed HybCNNViT model is illustrated in Figure 8. It has four major parts: an Autoencoder (AE), a Variational Autoencoder (VAE), a Convolutional Neural Network (CNN), and a Vision Transformer (ViT) or a Swin Transformer. The two independently optimised subnetworks are called Network A and Network B. Network B is composed of a VAE, a CNN module and a Swin Transformer. Network A is composed of an AE, a CNN module and a Vision Transformer. Network A improves the certainty of deepfake class prediction by mapping

segmented facial images (excluding inner facial components such as mouth, nose, and eyes) into a latent feature (LF) space through the AE. In contrast, Network B uses a VAE to minimise the reconstruction error between the input image that contains only the inner face components (mouth, nose, and eyes) and the reconstructed picture. The AE and VAE modules generate the latent representations from the face pictures extracted from the image frames. These representations encode hidden dependencies and unique deepfake artefacts encountered during the training process. CNN–Transformer: A unique hybrid architecture is built by combining CNN and Swin Transformer models. In the hybrid system, the CNN is used as the main feature extraction backbone to extract representative picture features from the input data. Then, the Swin Transformer uses self-attention and hierarchical feature learning to extract localised feature representations and global contextual information from inner facial areas. Each subnetwork analyses a 224 × 224 RGB picture along with the latent feature representation of the AE (I A) or the VAE (I B) by means of a CNN–ViT or CNN–Swin configuration. The combination of CNN and Transformer architectures enables the AE and VAE models to learn well the relationships between the latent characteristics obtained from the outer and inner face areas, respectively.

4.1 Autoencoder and Variational Autoencoder

A VAE and an AE are made up of two main subnetworks, an encoder and a decoder. The Encoder in the AE architecture encodes an input image$X \in R^{H' \times W' \times C'}$to a latent representation space $X \in R^{H' \times W' \times C'}$ denote the height and width of the created feature map respectively, and K is the number of output feature channels. The Decoder then reconstructs the latent representation $Z \in R^{H' \times W' \times C'}$ to generate an output picture $X \in R^{H' \times W' \times C}$. The encoder is generated by five convolutional layers with gradually increasing channel depths from 3 to 256. Every layer uses a 3x3 kernel with stride 2, GeLU activation, and Max-Pooling with a 2x2 kernel and stride 2. The encoding process results in a compressed latent feature (LF) representation of size 256x7x7. The Decoder consists of five transposed convolutional layers, where the channel size reduces from 256 to 3 using 2x2 kernels with a stride of 2. Deconvolutional layers are followed by GeLU activation. The last Decoder output IA reconstructs the H×W×C feature space corresponding to the original input picture. The proposed implementation has a spatial resolution of 224x224x3 for IA.

The VAE mainly aims to get informative latent embeddings from images with only eyes, nose and mouth. The input image is then reconstructed by minimising the reconstruction error using stochastic sampling in the latent space. The VAE Encoder maps an input image X to a probabilistic latent space Z–RK where Z–N(μ,σ2) and μ, σ2 are the mean and variance of the learned latent distribution respectively. The VAE Encoder is built with four convolutional layers, with the number of channels spanning from 3 to 128. It has a 3x3 kernel with stride 2, followed by Batch Normalisation (BN) and GeLU activation. The resulting Encoder output is a one dimensional vector of size 12544 which encodes the latent probability features. The decoder consists of four transposed convolutional layers with channel sizes between 256 and 3, 2x2 kernels and a stride of two. GeLU activation is applied after every transposed convolution process. The Decoder output IB is the reconstructed version of the input picture with dimensions H/2xW/2xC . IB dimensions in the implemented configuration are 112 x 112 x 3.

The full architectural layout is illustrated in Table III. For the AE and VAE designs, the convolutional layers were chosen considering the trade-off between the available computing resources, the memory restrictions, the performance of the models, the extensive experimental evaluation, and the overall training efficiency during the development of the proposed framework.

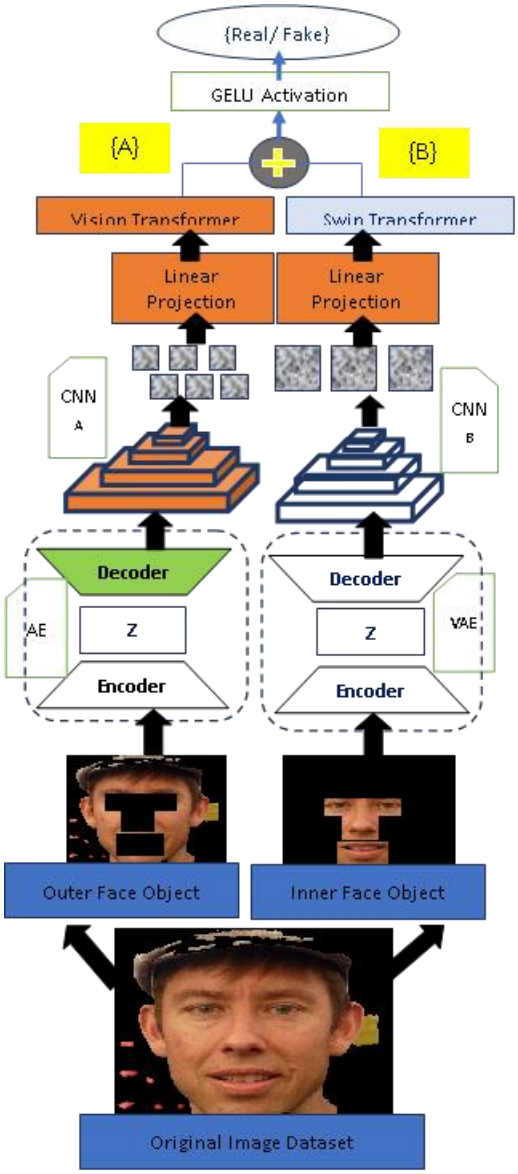

Figure 8. Proposed framework.

4.2 CNN-Swin Hybrid

The CNN–Swin Transformer architecture is an integrated CNN–Transformer architecture that utilises the complementary strengths of ConvNeXt and Swin Transformer architectures for deepfake detection. ConvNeXt is a convolutional based network which has shown strong performance in image recognition tasks by learning discriminative visual representations through a series of convolutional operations. The Swin Transformer, however, uses an attention-driven feature modelling strategy to capture both local image properties and long-range contextual dependencies. This combination of two architectures allows the framework to benefit from the feature learning ability of CNNs and the contextual awareness provided by Transformer models [39-42].

In the proposed framework, the CNN part acts as the main backbone for representation learning, and the Swin Transformer performs the contextual refinement of the extracted features.

The CNN–Swin architecture first extracts high level representations from the inner facial regions of the input image. Then, these representations are passed through a HybridEmbed module, which converts the extracted features into a compact latent embedding. The generated embedding vector is then input into the Swin Transformer for further feature analysis. The CNN backbone consists of the hierarchical convolutional blocks to learn informative feature representations from the input facial

images and the latent features (LFs) produced by the VAE. We leverage pre-trained CNN and Swin Transformer models trained on the ImageNet dataset for better feature generalisation and convergence.

The generated feature maps are fed to the HybridEmbed module where the CNN backbone extracts features. This module takes as input feature maps from the CNN, reshapes them into sequential representations, and maps them into an embedding space of size 768. The HybridEmbed module is made up of a 1x1 convolutional layer that reduces the channel dimension of the backbone output to the required embedding size. The transformed feature maps are then flattened and transposed to obtain a sequence of feature vectors, which are fed into the Swin Transformer to learn the hierarchical feature dependencies and contextual relationships.

The proposed HybCNNViT framework comprises two Hybrid CNN–ViT models in Network A receiving the input image of size 224×224×3 and a latent feature representation $I_{\Delta}$of the same size produced by the A.E. Each branch produces a feature vector of size 1,000, and the generated vectors are fused by concatenating them.

The combined representation is then passed through a linear mapping layer with output dimension of 2 to produce classification scores for real and fake categories. Network B has the same architecture as Network A, but replaces the AE with a VAE. Network B produces classification confidence scores and also reconstructs an image of 224×224×3 size. The outputs of Networks A and B are then ensembled by averaging their scores to obtain the final real/fake classification result.

In summary, the proposed HybCNNViT framework provides a unified hybrid architecture for deepfake detection, combining the representation learning advantages of pre-trained CNN-ViT and CNN-Swin models with the generative modelling power of AE and VAE networks. The discriminative and generative learning mechanisms are combined together to allow the framework to effectively detect the fine-grained manipulation artefacts and hidden visual inconsistencies which are characteristic of deepfake images.

Algorithm: Hybrid CNN-ViT

1. Input real and fake face image dataset
2. Apply preprocessing to image remove noise and unwanted object.
3. Remove background objects.
4. Generate image of inner (IFO) and outer face object (OFO).
5. Pass through Auto encoder/decoder AE for OFO inputs.
6. Pass through Variational Auto encoder VAE for IFO inputs.
7. Apply CNN to the AE and VAE output
8. Generate 8x8 size patches for OFO and 64x64 patches from IFO input parallel pipeline,
9. Pass o/p of both pipelines through Vision transformer (ViT) for OFO and Swin X^{mer} for IFO.
10. Pass the results of X^{mers} through activation function.
11. Combine the inferences of both pipelines and produce final result as (Real/Fake)

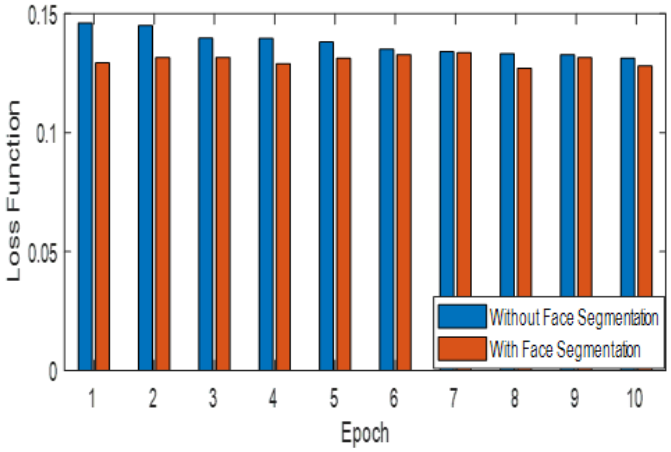

Figure 9. Loss function during training process using face data without and with face segmentation.

5. RESULTS

Table II. Performance analysis under different setup model in terms of recall.

	Recall		
	Easy	Mid	Hard
ViT is applied only	0.8126	0.8072	0.7954
Background Removal+ ViT only	0.8667	0.8541	0.8401
Background Removal+ Hybrid CNN ViT	0.8672	0.8531	0.8437
Inner outer Segment face object + Hybrid CNN-ViT	0.8732	0.8654	0.8544

Table III. Performance analysis under different setup model in terms of F1-score.

	F1 score		
	Easy	Mid	Hard
ViT is applied only	0.8286	0.8201	0.8132
Background Removal+ ViT only	0.8619	0.8322	0.8128
Background Removal+ Hybrid CNN ViT	0.864	0.8475	0.8263
Inner outer Segment face object + Hybrid CNN-ViT	0.8766	0.8702	0.8661

Table IV. Table II. Performance analysis under different setup model in terms of precision.

	Precision		
	Easy	Mid	Hard
ViT is applied only	0.7932	0.7841	0.77.89
Background Removal+ ViT only	0.8571	0.8437	0.8372
Background Removal+ Hybrid CNN ViT	0.8621	0.8577	0.8541
Inner outer Segment face object + Hybrid CNN-ViT	0.8768	0.8691	0.8602

Table V. Performance analysis under different setup model in terms of accuracy

	Accuracy		
	Easy	Mid	Hard
ViT is applied only	80.47%	78.36%	77.21%
Background Removal+ ViT only	84.55%	81.79%	79.27%
Background Removal+ Hybrid CNN ViT	87.21%	86.35%	85.83%
Inner outer Segment face object + Hybrid CNN-ViT	88.35%	87.21%	86.53%

The segmented face image real and fake faces from the training dataset are used to develop an efficient deepfake image classification system using the proposed HybCNNViT architecture. The network parameters are optimised through a specific training approach during the learning process. The model weights are iteratively updated during training as the number of epochs increases to minimise the optimisation loss function. Fig. 9. The convergence behaviour of training loss for 10 learning epochs. The bar graph compares two training scenarios, with the blue bars showing the loss values obtained without face segmentation and the other bars showing the variation of loss values in epochs 1-10 with face-region segmentation.The model is trained with segmented face photos and the loss becomes smaller and smaller in each epoch, which is obvious. Furthermore, the data clearly indicate a consistent reduction in the loss function as training proceeds, irrespective of whether segmentation is employed or not. However, the segmented-face configuration has better convergence properties and lower loss values during the training period. The loss function calculates the difference between the predicted output and the ground-truth output, which indicates the degree of agreement between the model's prediction and the corresponding target labels. Therefore, the smaller the loss value, the better the learning and prediction effect of the model. Given N input samples, y as the actual output, and $\hat{y}$ as the expected output, the loss function $L(y, \hat{y})$ can be written as follows:

$$L(y, \hat{y}) = \frac{1}{N}\sum_{i-1}^{N}(y_i - \hat{y}_i)^2 \qquad (1)$$

The classification accuracy, precision, recall and F1-score assessment is also based on the confusion matrix along with the True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) statistics. These performance metrics are computed for four experimental configurations: ViT only, Background Removal + ViT, Background Removal + Hybrid CNN–ViT, and Inner–Outer Segment Face Object + Hybrid CNN–ViT. The corresponding results for accuracy, precision, recall and F1-score are shown in Tables 2, 3, 4 and 5, respectively. The performance evaluation is performed on three types of datasets (easy, medium and hard) with different levels of picture alteration. The Inner – Outer Segment Face Object + Hybrid CNN – ViT method consistently achieves the best detection performance across all considered datasets among all evaluated configurations.

Table 7 presents the comparison of the proposed framework with the state-of-the-art methods relevant to deepfake detection. The last row in the table shows the proposed method called Inner–Outer Segment Face Object + Hybrid CNN–ViT. The first three rows summarise the competing methods such as Convolutional ViT, ViViT, and Convolutional Cross ViTEfficientNet B0. The Convolutional ViT method utilises the FaceForensics++ and DeepFake Detection Challenge (DFDC) datasets to detect manipulated face images using a Vision Transformer to extract features from CNN layers. This method is effective but has a high computational cost and a poor generalisation performance facing unseen facial changes.

Table VI. Comparison to state of art methods

Author	**Method**	**Accuracy**
Wodajo et al. [35]	CNN-ViT	67%
A. Arnab et al. [36]	ViViT	84%
Davide Coccomini et al. [37]	CNN-ViTEfficientNet	80%
Proposed Work	Inner outer Segment face object + Hyb- CNN-ViT	86% to 88%

The Video Vision Transformer (ViViT) approach [22] uses only transformer-based learning and was evaluated on the Kinetics dataset. But for it to reach acceptable performance levels it needs a lot of training data and a lot more processing power. EfficientNet [23] is another prominent method that utilises a Vision Transformer to learn temporal and relational features on DFDC and Celebrity DeepFake Dataset (CelebDF) after using EfficientNet to extract spatial representations. Conv. Cross ViT Even if it works well, the more complex model design brings more processing lag in the detection phase.

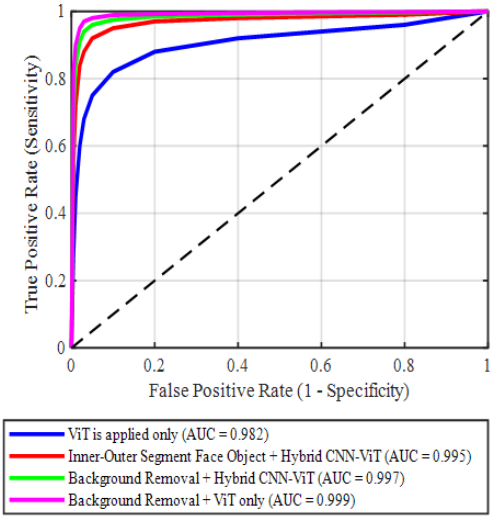

Figure 10. ROC curve for different experimental setup model configuration.

The proposed framework removes background regions and non-facial objects before feature extraction and classification to overcome these drawbacks. This means a significant reduction in model complexity and computing overhead during both training and inference, since the input data and corresponding redundancy are halved. Also, by removing unnecessary background data, the model can focus on face areas with significant manipulation artefacts, which leads to better discriminative feature representations. Thanks to these benefits, the proposed approach achieves better detection accuracy from easy to hard manipulation levels, from 86% to 88%. These results demonstrate the success of the proposed Inner-Outer Segment Face Object + Hybrid CNN-ViT architecture for deepfake image recognition, achieving better results than the competitive approaches, namely Convolutional ViT (67\%), ViViT (84\%) and Conv. Cross ViTEfficientNet (80\%).

6.CONCLUSIONS

A novel HybCNNViT framework for deepfake image detection was proposed in this paper. The framework employs AE, VAE, CNN, ViT and Swin transformer in a dual-stream architecture. Feature learning was improved and redundancy was removed through background removaland Inner–Outer Face Segmentation. Facial inputs segmentation led to experimental results with better convergence and lower loss function values. The Inner–Outer Face Segmentation+ HybCNNViT configuration performs best on Easy, Medium and Hard datasets. The proposed model achieved accuracy values from 86% to 88% and outperformed CNN–ViT, ViViT and CNN–ViTEfficientNet methods. The higher Precision, Recall and F1-score values also further validated the effectiveness of the framework. The combination of Latent Featureslearning and hybrid CNN-Transformer architectures allowed to reliably detect deep fake artefacts. Furthermore, BR greatly reduced computational complexity and improved classification performance. Future work is directed towards Deepfake Generated Imagedetection and Cross-Dataset Evaluation, lightweight real time deployment and better generalisation against unseen deep fake techniques.

ACKNOWLEDGEMENT

We would like to express our appreciation for the assignment of the Manuscript Communication Number [IU/R&D/2026-MCN0004941] per the guidelines provided by the university's studies and research departments. This identifier facilitates communication and tracking of our research throughout the publication process.

REFERENCES

[1] B. Chesney and D. Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 2019.

[2] A. Khodabakhsh, R. Ramachandra, K. Raja, and P. Wasnik, "Fake Face Detection Methods: Can They Be Generalized?" in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, 2018, pp. 1–6.

[3] P. Korshunov and S. Marcel, "DeepFakes: A New Threat to Face Recognition? Assessment and Detection," *arXiv preprint arXiv:1812.08685*, 2018.

[4] J. Brandon, "Terrifying High-Tech Porn: Creepy 'Deepfake' Videos Are on the Rise," 2018.

[5] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," pp. 1–7, 2018.

[6] L. Zheng, Y. Zhang, and V. L. L. Thing, "A Survey on Image Tampering and Its Detection in Real-World Photos," *Signal Processing: Image Communication*, vol. 58, pp. 380–399, 2018.

[7] A. J. Bose and P. Aarabi, "Virtual Fakes: Deep-Fakes for Virtual Reality," in *Proc. IEEE 21st Int. Workshop Multimedia Signal Processing (MMSP)*, 2019, pp. 1–1.

[8] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals," *arXiv preprint arXiv:1901.02212*, 2019.

[9] J. Brockschmidt, J. Shang, and J. Wu, "On the Generality of Facial Forgery Detection," in *Proc. IEEE 16th Int. Conf. Mobile Ad Hoc and Sensor Systems Workshops (MASSW)*, 2019, pp. 43–47.

[10] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI-Generated Fake Face Videos by Detecting Eye Blinking," *arXiv preprint arXiv:1806.02877*, 2018.

[11] T. Jung, S. Kim, and K. Kim, "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern," *IEEE Access*, vol. 8, pp. 83144–83154, 2020.

[12] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic Speech-Driven Facial Animation with GANs," *International Journal of Computer Vision*, vol. 128, pp. 1398–1413, 2020.

[13] H. X. Pham, Y. Wang, and V. Pavlovic, "Generative Adversarial Talking Head: Bringing Portraits to Life with a Weakly Supervised Neural Network," *arXiv preprint arXiv:1803.07716*, 2018.

[14] P. Charitidis, G. Kordopatis-Zilos, S. Papadopoulos, and I. Kompatsiaris, "Investigating the Impact of Pre-processing and Prediction Aggregation on the DeepFake Detection Task," *arXiv preprint arXiv:2006.07084*, 2020.

[15] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[16] Zuo, Z., Li, A., Wang, Z., Zhao, L., Dong, J., Wang, X., & Wang, M. (2024). Statistics enhancement generative adversarial networks for diverse conditional image synthesis. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(7), 6167-6180.

[17] Saharia C, Chan W, Saxena S et al (2022) Photorealistic text-to-image diffusion models with deep language understanding. In: Annual conference on neural information processing systems, 28 November–9 December 2022

[18] Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., ... & Liu, Z. (2025). Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13294-13304).

[19] Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., ... & Li, Z. (2024, September). Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *European Conference on Computer Vision* (pp. 74-91). Cham: Springer Nature Switzerland.

[20] Xiao, G., Yin, T., Freeman, W. T., Durand, F., & Han, S. (2025). Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, *133*(3), 1175-1194.

[21] Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., & Levy, O. (2023). Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, *36*, 36652-36663.

[22] Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., ... & Rombach, R. (2024, July). Scaling rectified flow transformers for high-resolution image synthesis. I*nternational conference on machine learning*.

[23] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684-10695).

[24] Yang, T., Huang, Z., Cao, J., Li, L., & Li, X. (2022, June). Deepfake network architecture attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 4, pp. 4662-4670).

[25] Rafique, R., Gantassi, R., Amin, R., Frnda, J., Mustapha, A., & Alshehri, A. H. (2023). Deep fake detection and classification using error-level analysis and deep learning. *Scientific reports*, *13*(1), 7422.

[26] Mohammed, A. (2024). Deep Fake Detection and Mitigation: Securing Against AI-Generated Manipulation. *Journal of Computational Innovation*, *4*(1).

[27] Wu, G., Wu, W., Liu, X., Xu, K., Wan, T., & Wang, W. (2023, July). Cheap-fake detection with LLM using prompt engineering. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (pp. 105-109). IEEE.

[28] Suratkar, S., & Kazi, F. (2023). Deep fake video detection using transfer learning approach. *Arabian Journal for Science and Engineering*, *48*(8), 9727-9737.

[29] Klemt, M., Segna, C., & Rohrbach, A. (2025). DeepFake Doctor: Diagnosing and Treating Audio-Video Fake Detection. *arXiv preprint arXiv:2506.05851*.

[30] Hamid, Y., Elyassami, S., Gulzar, Y., Balasaraswathi, V. R., Habuza, T., & Wani, S. (2023). An improvised CNN model for fake image detection. *International Journal of Information Technology*, *15*(1), 5-15.

[31] Uppada, S. K., Patel, P., & B, S. (2023). An image and text-based multimodal model for detecting fake news in OSN's. *Journal of Intelligent Information Systems*, *61*(2), 367-393.

[32] Liu, B., Yang, F., Bi, X., Xiao, B., Li, W., & Gao, X. (2022, October). Detecting generated images by real images. In *European conference on computer vision* (pp. 95-110). Cham: Springer Nature Switzerland.

[33] Suratkar, S., & Kazi, F. (2023). Deep fake video detection using transfer learning approach. *Arabian Journal for Science and Engineering*, *48*(8), 9727-9737.

[34] Bammey, Q. (2023). Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, *5*, 1-9.

[35] Wodajo, D., &Atnafu, S. (2021). Deepfake Video Detection Using Convolutional Vision Transformer. *arXiv e-prints*, arXiv-2102.

[36] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6836-6846).

[37] Coccomini, D. A., Messina, N., Gennaro, C., & Falchi, F. (2022, May). Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing* (pp. 219-229). Cham: Springer International Publishing.

[38] Sharma, R., & Banoudha, A. (2013). Implementation of N-Bit Divider using VHDL. International Journal of Research and Development in Applied Science and Engineering, 3(1).

[39] Sahay, S., Banoudha, A., & Sharma, R. (2014). On the use of ANFIS for predicting groundwater Levels in Hard Rock Area. Int. J. Res. Dev. Appl. Sci. Eng.

[40] Z. A. Ansari, M. M. Tripathi, and R. Ahmed, "The role of explainable AI in enhancing breast cancer diagnosis using machine learning and deep learning models," Discovery Artificial Intelligence, vol. 5, art. no. 75, 2025, doi:10.1007/s44163-025-00307-8.

[41] Roshan Jahan, Saleha Mariyam, 2025, Robust Facial Recognition using Deep Learning with MTCNN, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 13, Issue 06 (July 2025).

[42] Alam, M., Alourani, A., Ali, A., & Ahamad, F. (2025). Symmetry-Aware Feature Representations and Model Optimization for Interpretable Machine Learning. Symmetry, 17(11), 1821.