An Explainable AI Fusion-Based Model for Enhanced Deepfake Detection Using Vision Transformer and InceptionResNetV1

Yousef A. Alsamaani[1,*], Murad A. Rassam[1]

[1]Department of Information Technology, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia

Emails: 441112423@qu.edu.sa; M.Qasem@qu.edu.sa

Abstract

Generative AI has made significant strides over the past few years, and this progress has accelerated the development of deepfake techniques, which can unfortunately be used for harmful purposes. It is essential to stay up-to-date with this advancement. In this paper, we present an explainable weighted average fusion deepfake detection system that combines Vision Transformer (ViT) and InceptionResNetV1 to improve classification accuracy. We also employed LIME and GradCAM++ to provide interpretability for the model decision. ViT utilizes self-attention modules to extract features, whereas InceptionResNetV1 employs convolutional layers to extract spatial features. Grad-CAM++ highlights the important regions influencing classification, and LIME examines the regional contributions. Together, these tools offer a deeper understanding of the model's decision-making process. Our fusion technique combines the outputs of both models by assigning specific weights that users can adjust interactively through the user interface. The use of these tools gives a better understanding of how the model classifies, which improves transparency and reliability in the models. The performance of the fusion strategy is tested with accuracy, precision, recall, and F1-score. Our proposed model achieves a classification accuracy of 99.19%, surpassing both ViT and InceptionResNetV1 when we evaluated them individually. To the best of our knowledge, this work represents the first deepfake detection model that combines Vision Transformer (ViT) and InceptionResNetV1 using a weighted averaging fusion approach with dual explainability techniques.

Keywords: Deepfake Detection; Machine Learning; Deep Learning; Detection Framework; Explainable Artificial Intelligence (XAI)

1. Introduction

Over the past few years, the rapid progress in artificial intelligence (AI) has introduced deepfake technology, which allows people to generate highly convincing but fabricated images and videos [1]. Deepfakes utilize sophisticated deep learning techniques to manipulate media in a manner that is often indistinguishable from the original. Though there are benefits to using technology, and it could be used in applications and fields such as entertainment, education, and graphics, there are many risks [2]. The ability to make fabricated visual media that is very close to the original one raises questions over what harmful things it could be used for. For example, misinformation, identity theft, and privacy breaches make detection of deepfakes an important discourse area [3].

One of the biggest challenges in detecting deepfakes is their sophistication, causing them to become widely adopted [4]. The earliest detection techniques were used to identify overt distortions or discrepancies in artificial media. However, as generative adversarial networks (GANs) and other deep learning methods continue to improve, deepfakes have become increasingly difficult to distinguish from authentic images [3]. This has rendered many traditional detection methods unsuccessful, which has created the need to come up with more complex techniques to detect tiny artifacts that might not be directly detectable by the human eye.

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

Machine learning models, particularly Convolutional Neural Networks (CNNs), have also made significant contributions to deepfake detection. CNN-based methods have the ability to extract features automatically from images and learn to identify patterns that distinguish real from manipulated media [5]. Yet, these models are prone to overfitting and poor generalization among various deepfake datasets, as well as ineffectiveness in interpreting their decision-making process. Insufficient explainability in deepfake detection is a major concern, as black-box models offer no insight into how or why an instance is classified as authentic or artificial. The opacity is limiting their credibility and applicability in practice, particularly in sensitive domains such as cybersecurity and digital forensics.

The latest advancements in computer vision have introduced alternative deep network architectures, such as ViT, which have surpassed other performance levels in image classification [6]. What makes ViT different from CNNs is that it takes an image in patches as a sequence to capture long-distance dependencies and spatial relationships in an image. For this reason, ViT is more effective in recognizing faint deepfake features that can be scattered across various regions [7]. Hybrid models that combine various neural network architectures have also emerged as popular due to their capacity to leverage strengths from multiple methods to offer higher overall performance [8].

Another important factor in handling deepfakes is transparency and interpretability. Techniques such as Explainable AI (XAI) have emerged to provide insight into how models make their decisions, thereby bolstering user confidence in them and enabling their use in real-world applications [9]. Techniques such as Grad-CAM++ and LIME provide visual and interpretive explanations to highlight the most important features contributing to a model's classification decision [10]. Implementing these techniques in deepfake detection is important for increasing accuracy, in addition to interpretability.

Due to the increasing realism in deepfake content, alongside the current limitations in most deepfake detectors, there will always be a need for more effective and more interpretable solutions, and to advance the detection of deepfakes, superior deep learning architectures, fusion techniques, and explainability methodologies need to be examined. This work will aid in overcoming these challenges by introducing a resilient fusion approach to counteract the dangers associated with synthetic media.

2. Related Work

Deepfake detection field has also seen major breakthroughs with various combinations of deep learning frameworks, especially with convolutional neural networks (CNNs) and Vision Transformers (ViTs) [11]. Although these fusion methods have significantly improved classification performance, interpretability in deep models remains a long-standing issue. To overcome this problem, various studies have employed explainable AI (XAI) methods, which include Grad-CAM++ and LIME.

Kaddar et al. [12] was one of the first attempts to combine CCNs and ViTs for deepfake detection. They used a feature-level fusion method, merging Xception with ViT to combine CNNs' spatial feature extraction with transformers' long contextual modeling. The model performed on FaceForensics++ and DFDC with an accuracy of 89.7% on DFDC. One of the main limitations of their work is the lack of explainability, which makes it hard to understand the model decision-making and the process of the model, nor can you identify the important facial features determining classification results. The other limitation, in my opinion, is that the model used frame-based datasets to train, and upon being tested on unseen deepfake variations, their generalizability was lower. In comparison to this work, the lack of use of Grad-CAM++ or LIME in their work is one of their major shortcomings in terms of interpretability, which we address in this research work.

Similarly, using the CNN-transformers fusion, Khan et al. [13] presented an early fusion technique that combines three models, Xception, EfficientNet, and ViT, to enhance feature representation of their approach. They followed this method because they wanted to combine several feature extraction layers before passing the learned representations to the ViT backbone. They achieved an accuracy of 98.2% using DFDC, outperforming those who used CNN alone. However, the use of multiple CNN architectures requires high computational resources, which makes such an approach less desirable in applications that demand real-time. Furthermore, their work did not incorporate any XAI techniques, making decisions in their model unclear, while in this work, we apply weighted averaging fusion, which weighs contributions from both InceptionResNetV1 and ViT, with the use of Grad-CAM++ and LIME as explainable methods. This work fills in the interpretability gap in their work.

In a further effort to improve deepfake detection, Wang et al. [14] presented a multi-scale feature fusion framework that combines EfficientNet with ViT to enhance the detection performance in various resolutions of images. Their method achieved an accuracy of 97.6% when evaluated on FaceForensics++, and they utilized Grad-CAM++ as an XAI method on both models, providing insights into how manipulated facial pixels are detected by the model. However, with all that is good about its performance, they used a small dataset, which was one of the weaknesses that limited the generalization of their model to various forms of deepfake manipulations. Compared to this, our

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

approach, using both LIME and Grad-CAM++ in this work, means there is more thorough interpretability, with possibilities for both spatial heat map visualization as well as instance-centric feature importance analysis. Using the weighted averaging fusion technique in this work has also led to improved classification accuracy (99.19%) compared to Wang et al.'s.

Different from studies based mainly on fusion, Hasan Abir et al. [15] investigated explainability in deepfake detection without using fusion models. They utilized ResNet and DenseNet architectures and employed LIME to examine how models identified manipulated features in deepfake images. The research demonstrated that techniques based on XAI were capable of revealing inconsistencies on which deep learning models depend, thereby enhancing their interpretability. The performance, however, of their work lacked optimality since, in the absence of fusion models, they had a classification accuracy of 92.3%. The results of Hasan Abir et al. underscore the importance of explainability in deepfake detection, which is also the primary objective of this work.

Far from these traditional methods, this paper aims to achieve classification performance through an optimal weighted averaging fusion of ViT and InceptionResNetV1, with added interpretability via the use of Grad-CAM++ and LIME. In contrast to feature- and decision-level fusion methods, weighted averaging prevents overdependence on a single network. Additionally, the use of dual explainability methods provides thorough insight into decision-making by models, which makes this method especially suitable for forensic and security applications that are highly reliant upon transparency.

Some latest research has established the hybrid architecture by connecting CNNs and Vision Transformers (ViTs) for the detection of deepfakes. Soudy et al. [16] proposed the fusion architecture of Vision Transformers-CNNs as a detection model with nearly 97% performance on benchmark datasets such as FaceForensics++ and DFDC. The architecture combines the spatial and contextual features in a well-balanced way but does not incorporate explainability methods.

Bar Cavia et al. [17] proposed an enhanced real-time detection system for unconstrained settings, which achieved approximately 95% on different video datasets. Making computational speed the priority, their system also does not have built-in explainable AI techniques; hence, high interpretability is not achieved.

Wodajo et al. [18] presented the Generative Convolutional Vision Transformer (GenConViT), wherein they combined the generative convolutional layer and ViT layer for video deepfake detection. The model achieved approximately 96.5% in accuracy when tested against baseline standards, with efficient detection but without the use of the dual XAI methods for further interpretability.

Unlike these prior works, our research uniquely combines ViT with InceptionResNetV1 through weighted average fusion, complemented by dual explainability techniques, Grad-CAM++ and LIME. The fusion enhances the classification capability while providing a deeper insight into the model's decision-making process, thereby filling the transparency gap in most current approaches.

3. Problem Visualization and Solution Concept

The core challenge of this paper is identifying deepfakes, which are becoming increasingly widespread and sophisticated. Even though traditional detection models are highly accurate in specific circumstances, they have one major drawback (i.e., poor interpretability of their predictions). Such models are "black box" in nature, giving predictions with no insight into their decision-making [19]. The opacity of such models restricts their usability in actual applications, as experts and users cannot have trust in the system if they don't know how it comes to its conclusions.

In deepfake detection, this non-interpretable nature of model predictions is particularly risky. Threat actors can exploit this vulnerability by creating highly realistic deepfakes that can evade detection by automated systems [20]. Therefore, there is a critical need for detection models that are not only accurate in making the right predictions but also interpretable, allowing experts to validate and refine the models.

To meet this challenge, we recommend the incorporation of XAI approaches in a two-model architecture for detecting deepfakes consisting of VisionTransformer (ViT) for feature extraction in general and InceptionResNetV1 for the analysis of local characteristics, where ViT and the local outliers by InceptionResNetV1, thereby providing a balanced and accurate detection process, retain the global structure pattern.

To increase interpretability, LIME is also included, as well as Grad-CAM++. Grad-CAM++ denotes the salient face regions in the output of InceptionResNetV1 in heatmaps, while LIME perturbs regions in the images in order to comprehend the contribution to the outcome of ViT in providing complementary visual as well as textual explanations. The explainability framework helps experts in the validation and improvement of model reliability.

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

The research problem is therefore focused on both the improvement in accuracy as well as clear visualization in deepfake detection frameworks. The description of the problem along with the methodology for the solution is given in Figure 1.

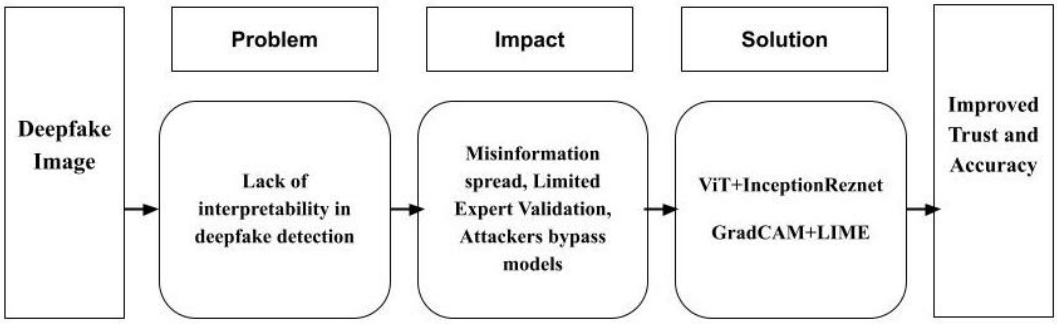

Figure 1. Conceptual Framework of the problem and solution

In the solution framework proposed in this paper, we consider both the problem of improving deepfake detection precision and extending model transparency. The lack of interpretability in modern deep learning models restricts both the scope of their usage and the level of trust in them since humans are unable to trust the system's outcome or adjust its performance. To circumvent these problems, we employ a two-model architecture in conjunction with explainable artificial intelligence methods.

These models incorporate the Vision Transformer (ViT) and the InceptionResNetV1, with both models having discernible strengths in detection. The Vision Transformer employs a transformer architecture with self-attention to capture holistic image characteristics and detect complex relations and patterns across the entire image. The InceptionResNetV1 addresses local characteristics and face distortions by utilizing convolutional layers to examine smaller regions in the image in-depth. The interplay between the two models enables extensive detection of the deepfake artifacts at the micro and macro levels.

For easier interpretability, the system incorporates Grad-CAM++ and LIME as XAI methods. The heatmaps from Grad-CAM++ indicate the regions of significance that impact the model's decision-making process, allowing one to comprehend how the InceptionResNetV1 model recognizes tampered attributes. Similarly, LIME perturbs localized patch areas of the input image to break down the ViT model's response, generating textual explanations for each prediction. The combined techniques provide an overall view of the decision-making processes of the models, assisting both technical and non-technical personnel in confirming the system's outputs.

This solution supports the research goals in terms of not only increased accuracy from leveraging dual models but also increased explainability. Such additions are crucial for facilitating adoption in practice, as they contribute to developing user trust in the detection. The use of both models, along with XAI techniques, addresses the issue of poor interpretability in AI-based detection models and establishes a robust framework for deepfake detection. The connection between problem identification and solution is presented in Table 1.

Table 1: Problem and Solution Concept

Problem Description	**Proposed Solution**
Lack of interpretability in AI models	Integration of XAI techniques (Grad-CAM++, LIME)
Difficulty in analyzing global and local features	Use of dual-model system (ViT & InceptionResNetV1)
Reduced trust in model predictions	Enhanced transparency through interpretable outputs

With such an organization of the solution, the research strikes a balance between performance and interpretability, satisfying both technical and user-oriented demands.

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

4. Implementation Steps of the Proposed XAI Fusion-Based Model

The implementation process is conducted in three major steps to build the deepfake detection system. It begins with data pre-processing and preparation to ensure that the data is correctly formatted for feeding into the model. Then, we configure the ViT and InceptionResNetV1 models and combine them through a fusion technique to improve classification performance. The second step is to implement explainable artificial intelligence as indicated in Figure 2.

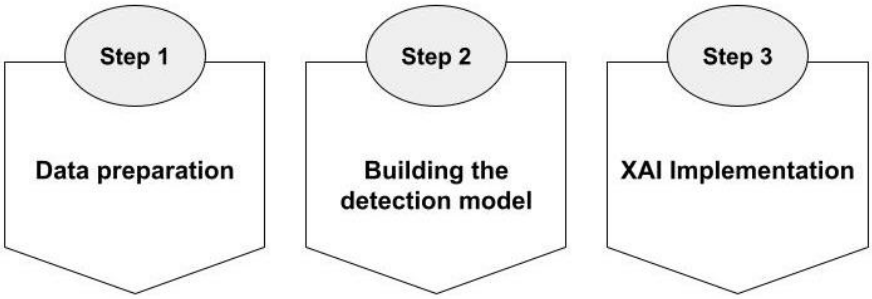

Figure 2. Overview of the implementation steps

4.1 Data Preparation

In this research, we employ two independent datasets for training and testing deepfake detection models. The first dataset used to test and train the InceptionResNetV1 model, as well as fine-tune the ViT model, is downloaded from Kaggle [21] and contains 140,000 images, divided equally between real and fake categories. The real images are drawn from Nvidia's Flickr dataset, while the fake images are generated using StyleGAN. The second database is used to test the pre-trained ViT model, maintaining the original framework of the dataset as designed by the model's creator. The database downloaded from Kaggle [22] comprises 190,335 images in the OpenForensics dataset, organized into training, validation, and test subsets. Each database is preprocessed separately to ensure compatibility with its respective model.

4.1.1 Dataset for InceptionResNetV1 Model

For training and evaluation of the InceptionResNetV1 model, we use three subsets of this dataset: 100,000 for training, 20,000 for validation, and another 20,000 for testing. Each of these subsets has an even split between real and fake images to ensure that the training is based on equally weighted data. Table 2 presents the structure of the dataset, and Figure 3 provides a visualization of the distribution of real and fake images in the InceptionResNetV1 dataset.

Table 2: Dataset structure for the InceptionResNetV1 model.

Subset	***Real Images***	***Fake Images***	***Total***
Training	50000	50000	100000
Validation	10000	10000	20000
Testing	10000	10000	20000

Before feeding the images to the InceptionResNetV1 model, a sequence of preprocessing operations is applied to ensure compatibility and conformity to the model's requirements. The first step is to resize all images to 256 × 256 pixels to normalize the input dimensions. Data augmentation operations are also performed in the form of random horizontal flipping, rotation of up to 15 degrees, and color jittering, which applies random variations in brightness, contrast, and saturation. Such operations are necessary to enhance the model's robustness concerning various face orientations and lighting variations. The images are also normalized, with a mean of (0.485, 0.456, 0.406) and a standard deviation of (0.229, 0.224, 0.225), to ensure pixel intensity values conform to the expected input format of the model.

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

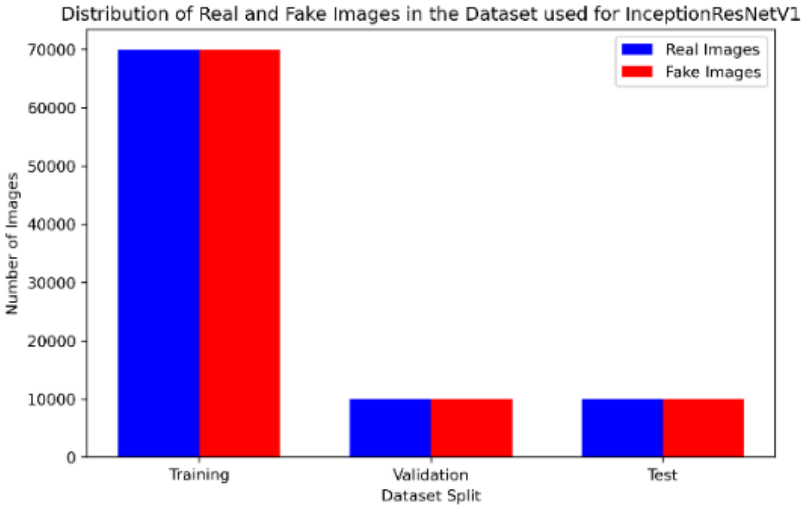

Figure 3. Distribution images for the InceptionResNetV1 dataset

The InceptionResNetV1 is trained with this dataset with a batch size of 32, an initial learning rate of 0.001, and for a total of 50 epochs. The Adam optimizer is employed for enhanced weight update, and cross-entropy loss is used as the binary classification objective function. For the training process, we used a local workstation with an RTX 2080 GPU, an Intel i9-9900K CPU, and 48 GB of system RAM. The Conda environment is used for training the model, utilizing the TensorFlow and PyTorch frameworks optimized for GPU use.

4.1.2 Dataset for Vision Transformer (ViT) Model

Unlike InceptionResNetV1, this ViT model is first assessed as pre-trained to facilitate comparison with the fine-tuned one later. A pretrained model assessment will be carried out using the original dataset presented by the developer of the original model, which is available in another Kaggle repository. This dataset follows the original author's experimental setup and is used solely for pretrained model assessment in this study. The dataset contains 190,335 images derived from OpenForensics, with 140,002 used for training, 39,428 for validation, and 10,905 for testing. The structure and distribution of the dataset for the ViT model are presented in Table 3, while Figure 4 illustrates, visually, how the actual versus fake images are distributed in the pre-trained ViT set.

Table 3: Dataset structure for pretrained ViT model.

Subset	***Real Images***	***Fake Images***	***Total***
Training	70001	70001	140002
Validation	9787	19641	39428
Testing	5413	5492	10905

Because ViT is based on transformer architecture, we need to use a different preprocessing pipeline from that used by convolutional models. The images are all resized to 224 × 224 pixels, as this is the default receptive size of the original ViT-Base-Patch16-224 model [23]. Images are also normalized using a mean of (0.5, 0.5, 0.5) and a standard deviation of (0.5, 0.5, 0.5) to ensure that the input distribution is compatible with the pretrained model. Pretrained ViT is only tested on the test set to maintain consistency with its original baseline. This is done to have a direct test of how ViT can perform before fine-tuning.

In the present paper, we fine-tuned the ViT model using the same deepfake data as the InceptionResNetV1 model, which was explained earlier, to compare and test both models on a single dataset. The data set, as explained earlier, contains 100,000 training images, 20,000 for validation, and 20,000 test images, with an even split between real and fake images.

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

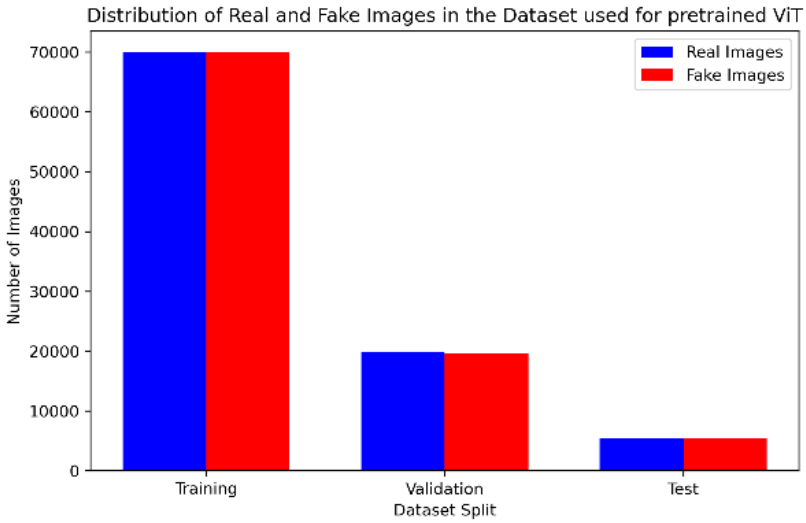

Figure 4. Distribution images for pretrained ViT dataset

4.2 Building the Detection Models

The deepfake detection system used in this work is founded upon a fusion strategy that combines two different models: InceptionResNetV1 and Vision Transformer (ViT). The two models bring complementary strengths to bear upon the task of classification, with InceptionResNetV1 emphasizing fine-grained feature extraction from faces and ViT taking advantage of its transformer structure to learn global image dependencies. This section describes the architecture components of each of these models and their respective configurations, as well as their function in the overall detection system.

4.2.1 InceptionResNetV1 Model

InceptionResNetV1 is a network specifically dedicated to face recognition, combining Inception modules with block connections for effective multi-scale feature extraction from images. Three inception blocks (A, B, and C) with different kernel sizes (1×1, 3×3, and 5×5) for convolutional filters are present in this network's architecture. Inception-like layers of the network enable it to extract fine details at different spatial resolutions, making it highly effective for fine-grained inconsistency detection in deepfakes.

We trained InceptionResNetV1 on data described in Section 4.1.1. The final classification layer was modified to enable its use in a binary classification task, where real and fake images need to be distinguished. Training was done with a batch size of 32, a learning rate of 0.001, and with an Adam optimizer for effective convergence. Optimization of classification accuracy was done based on cross-entropy loss as an objective function, for which the network was trained for 50 epochs. Training was monitored based on performance indicators, such as accuracy and loss, using a validation set.

After being trained, we tested the model using the test set, employing standard performance metrics such as accuracy, precision, recall, and the F1-score to assess its classification performance. The performance results from this evaluation gave insight into how well such a model could discriminate based on localized facial features to identify whether an image is real or not.

4.2.2 Vision Transformer (ViT) Model

The Vision Transformer (ViT) used in this work is based on a self-attention architecture, which enables it to process images in terms of fixed-size patches rather than through standard convolutional operations. This enables the model to extract long-distance relationships among regions in images, making it highly suitable for identifying deepfakes based on discrepancies in images.

The pretrained ViT model in this paper was evaluated in its pretrained state before fine-tuning it. The model was originally obtained from Hugging Face and was initially trained on ImageNet-21k, a large-scale dataset comprising over 14 million images across 21,843 classes. The model creator later fine-tuned it on ImageNet-1k, which includes 1,000 image categories [20].

Input to ViT is an array of non-overlapping patches of size 16×16, which are embedded in a higher-dimensional space before being processed by multi-head self-attention. The model comprises various transformer encoder

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

layers, which are composed of self-attention and feed-forward network elements. The prediction is made using the [CLS] token, which aggregates global feature representations from all patches.

To evaluate the pretrained ViT model, it was tested using test data from its dataset in Section 4.1.2, based on the same preprocessing pipeline as its original training. No extra fine-tuning was performed at this point to ensure that the test results demonstrate the model's ability to identify deepfakes based on its pre-training knowledge, for comparison with the subsequent fine-tuned model.

Pretrained ViT is fine-tuned over the same InceptionResNetV1 training data to maintain comparability in the evaluation. In contrast to the setup in which only ViT would be evaluated, this work trains ViT for 10 epochs with a learning rate of 0.0001, a batch size of 32, and the Adam optimizer. The training, validation, and test data are all drawn from the same distribution with an equal number of real and generated images.

Figure 5 illustrates the architecture of both models, with their respective key components being convolutional layers, inception modules, and a final classification layer in the case of the InceptionResNetV1 model, as well as key processing stages of patch embedding, transformer encoders, and a final classification layer in the case of the ViT model.

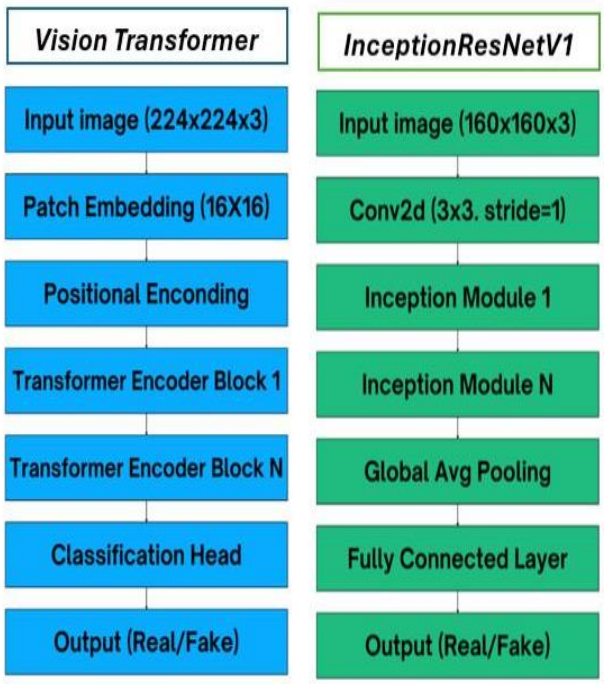

Figure 5. The architecture of Vision Transformer and InceptionResNetv1

4.2.3 Fusion Mechanism in the Proposed Model

After developing and training individual models, the next essential aspect of this research is to combine the outputs of the models to establish a reliable decision-making process. We developed the integration process to use the complementary merits of ViT and InceptionResNetV1, aiming to provide improved reliability in detecting deepfakes.

The fusion technique presented in this paper is based on weighted averaging, where we combine the results from both models (i.e., ViT and InceptionResNetV1) using a mathematical equation rather than complex rule-based guidelines. This technique provides contributions from both models without adding extra decision rules. This helps maintain interpretability and simplicity in the model.

Each model separately processes the distinct input image and produces a probability score indicating the level of confidence in classifying the picture as real or fake. Since we used different architectures and optimization strategies in training the models, the raw measures of confidence had to be normalized for interoperability before fusion. The resultant decision score was calculated as a weighted sum of the output by individual models as depicted in Equation (1):

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

$$P_{\text{final}} = w_1 P_{\text{ViT}} + w_2 P_{\text{Inception}}$$

where: P_{ViT} represents the classification probability given by the ViT model, $P_{\text{Inception}}$ and represents the output classification probability given by the InceptionResNetV1 model, and $w1$ *and* $w2$ are the weights assigned to each model's output. Figure 6 on the following page illustrates the fusion mechanism in the proposed model.

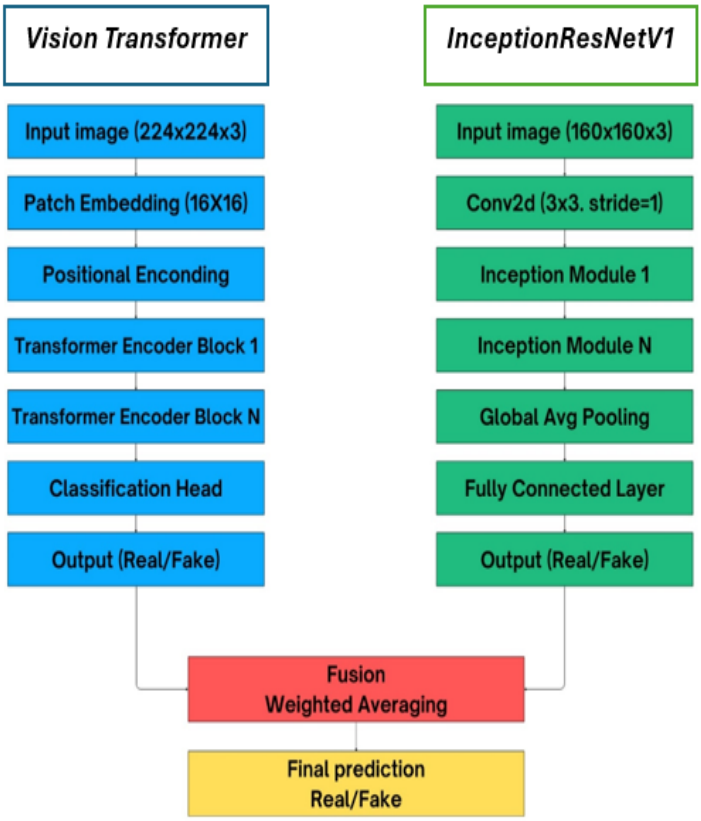

Figure 6. An overview of the fusion mechanism in the proposed model

4.3 Implementation of XAI Techniques in the Proposed Model

In addition to improving the interpretability of the deepfake detection system, this paper combines Explainable AI techniques, namely Grad-CAM++ and LIME. Both techniques provide both visual and textual interpretations of the model's decision, offering insight into which regions of an image contributed to the classification. The use of XAI is also crucial in deepfake detection, as understanding how models work is essential in identifying manipulated content. The implementation of Grad-CAM++ and LIME is presented in this section, along with a discussion of how both contribute to the model's explainability.

4.3.1 Grad-CAM++ for Convolutional Neural Networks

Grad-CAM++ is another gradient-based visualization technique that highlights the important regions in an image that contribute most to a model's decision-making [24]. It is an extension of Grad-CAM, refining weight allocations for activation maps to generate more detailed heatmaps. We utilized Grad-CAM++ with the InceptionResNetV1 model to leverage its capability to localize discriminative features in deepfakes.

The process to implement Grad-CAM++ is systematic. The model first takes an input image to produce feature maps from its convolutional layers. The gradients of the predicted class concerning these feature maps are calculated, and then a weighted aggregate of the activations is used to generate a heat map. The heat map is then overlaid on the original picture to highlight the regions that contributed to the classification.

Figure 7 shows the visualization of Grad-CAM++ over an example deepfake image, with red-colored zones indicating highly important regions for the model's prediction.

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

Figure 7. Grad-CAM++ visualization

Adding Grad-CAM++ to the deepfake detection system offers two significant advantages. It first allows for qualitative analysis of model choices, so researchers can check whether the network is concentrating on appropriate regions of the face. It also assists in identifying spurious correlations or imbalances in the network's behaviour, which may occur due to data imbalances.

4.3.2 LIME for Transformer-based Models

Although Grad-CAM++ is effective at handling convolutional networks, its usability is limited in Vision Transformers (ViT) because they lack traditional convolutional layers. Alternatively, we used LIME (Local Interpretable Model-agnostic Explanations) to explain ViT's prediction. LIME modifies images, creates several variations with minimal differences, and then measures how the model's prediction is altered accordingly [22]. After analyzing these variations, LIME detects the influential patches in an image, which are responsible for determining the classification decision.

Several steps are involved in implementing LIME with ViT. It starts by segmenting an input image into superpixels—these are meaningful regions, not just tiny individual pixels. Then, we test the ViT model multiple times using slightly different versions of the image where specific superpixels are masked. By watching how the model's predictions change, LIME assigns an importance score to each region. This process ultimately generates a heatmap that clearly shows us the most critical areas of an image.

An example of how LIME interprets a deepfake image is is in Figure 8. That figure highlights the key regions in the image that genuinely influenced the model's classification.

The use of LIME increases the interpretability of the ViT model, allowing for the identification of the contribution of self-attention to classification, in addition to providing an alternative explanation to Grad-CAM++ for comparing localized (CNN-based) features with overall (transformer-based) features.

Figure 8. LIME visualization

4.3.3 Integration of XAI Techniques into the Detection Pipeline

To introduce explainability to the deepfake detection system, we employed Grad-CAM++ and LIME as post-hoc analysis methods, which do not disrupt the original classification process but are used to interpret model classification upon prediction. Upon processing an image through the detection pipeline, the system initially

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

classifies it as real or fake based on the ViT-InceptionResNetV1 fusion. After classification, the respective XAI technique is used:

- When the classification is based on InceptionResNetV1, Grad-CAM++ generates heatmaps of which parts of a face contributed to the prediction.

- When ViT makes the classification, LIME then highlights the most critical patches that led to the confidence of the model.

The inclusion of XAI techniques enhances transparency in the system, making it more interpretable and reliable. The visualization of how a decision is made adds another layer of verification and validation, confirming that what the model predicts matches human expectations. This is especially pertinent to scenarios in real-world applications, where knowing how and why an image is classified as fake by a model can help forensic investigators and security experts evaluate the authenticity of digital content.

4.3.4 MTCNN for Facial Landmark Detection

To further enhance the interpretability of the system, we implemented Multi-Task Cascaded Convolutional Networks (MTCNN) as a pre-processor for both ViT and InceptionResNetV1. The MTCNN is a specialized deep neural network used explicitly in face detection and landmarks' location in faces, which enables the detection of major face regions, such as the eyes, nose, and mouth [26]. We use the landmarks as reference points for explanation methods to guarantee that the results from both Grad-CAM++ and LIME are in accurate correspondence with meaningful face features. Its architecture consists of three cascaded neural networks: the Proposal Network (P-Net) for generating candidate face regions, the Refinement Network (R-Net) for eliminating false positives, and the Output Network (O-Net) for effectively detecting landmarks. Using MTCNN pre-processing, the system enhances robustness in deepfake detection by mitigating face alignment variations that could influence classification outcomes.

4.4 System Implementation and Experimental Setup

This section outlines the experimental setup, including the hardware specifications and software dependencies used for training and evaluation environments for both InceptionResNetV1 and Vision Transformer (ViT) models.

4.4.1 Hardware Specifications

Our workstation enabled efficient handling of deep learning training along with evaluation and explainability calculations. The Intel Core i9-9900K processor, featuring 8 cores and 16 threads, operates within the system at 3.6 GHz. Our deep learning computations utilized the NVIDIA RTX 2080 featuring 8GB GDDR6 memory. The system features 48GB DDR4 RAM, which enables simultaneous handling of large datasets and multiple model inferences without encountering performance limitations.

4.4.2 Software and Environment Setup

We used Windows 10 Pro OS with Python 3.8 to work with the deep learning packages. For explainability techniques, we employed Grad-CAM (version 1.5.2) for visualization of model decision-making in convolutional networks and LIME (version 0.2.0.1) for generating explainable interpretations for the Vision Transformer model. We also used Timm (1.0.9) for extracting the pre-trained ViT model and OpenCV (4.5.1) for image processing. For visualization, we installed Matplotlib (3.4.3). Additionally, we incorporated the Gradio (3.9) library to provide an interactive user interface for model prediction and visualization of explainability.

We coded and exercised the deepfake detection framework in the Anaconda environment using Jupyter Notebook, which provided a modular and interactive development environment for effective debugging, visualization, and analysis of the model. To tap into the acceleration offered by the GPU and maximize performance in training and inference tasks, we employed CUDA toolkit version 11.3.

4.5 Proposed Model

As demonstrated in Figure 9 below, the detection system follows a structured pipeline. First, the model resizes the input image at the preprocessing level to the ViT (224×224) and InceptionResNetV1 (160×160) input sizes before passing it through both models separately to generate independent classification probabilities. The model takes these two predictions and then averages them using a weighted mechanism. Finally, the model gives its final class decision—real or Fake—based on the calculated score.

This work emphasizes the merging of explainability techniques to ensure that model predictions achieve both precision and transparency. InceptionResNetV1 utilizes Grad-CAM++ to reveal critical facial regions used during classification, while LIME generates superpixel-based importance maps for the ViT model. The implementation

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

of these techniques enhances model transparency by enabling end-users to comprehend the reasoning behind predictions.

The structure developed here for the model is designed to address some key weaknesses in existing methods for detecting deepfakes, including false positives, overfitting on specific datasets, and the non-interpretability of the models. By combining two architecturally disparate models and extending them with XAI approaches, the method offers a balanced trade-off between performance and interpretability.

We chose this approach because it is simple and effective. More complex fusion mechanisms, such as attention-based mechanisms, can potentially capture the dynamic importance of each model output. However, they usually require additional parameters and extensive training, which can increase the risk of overfitting, especially when the training data is limited. We believe our approach enables direct integration from both models, allowing you to control and adjust the weights manually. This method makes the decision balanced without the complexity when learning extra fusion layers. At the same time, it reduces the computational requirements, making it practical for real-world applications.

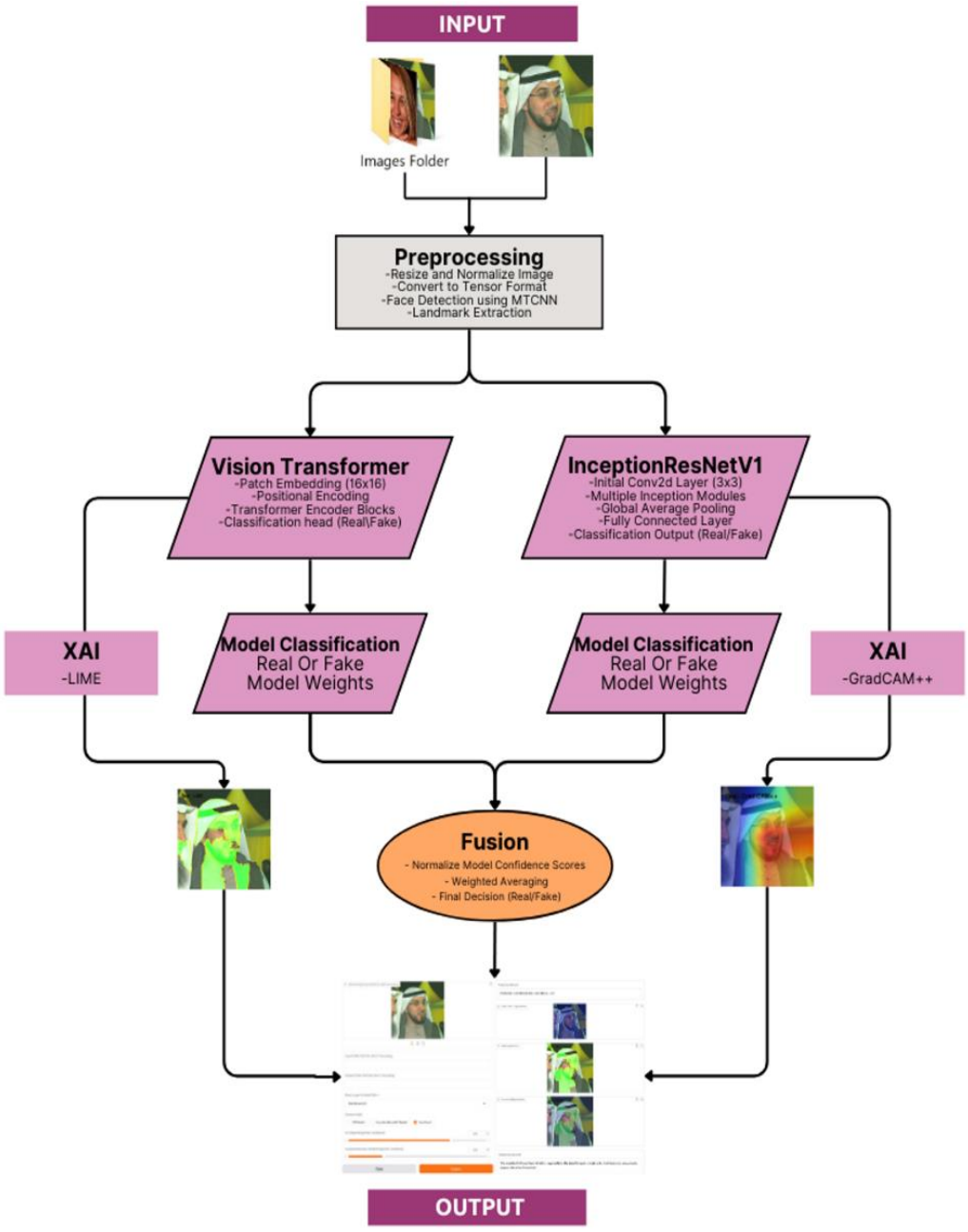

Figure 9. Proposed model pipeline

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

5. Results and discussions

We employed various evaluation measures, including accuracy, precision, recall, and F1-score, to assess the models underlying this paper. In this section, we present and discuss the significant findings of this experiment, which can be divided into four sections: classification outcomes of ViT and InceptionResNetV1 models, the effect of late fusion on detection performance, and the contribution of explainability techniques (Grad-CAM++ and LIME) to promoting the transparency of models.

5.1 Classification Results of ViT and InceptionResNetV1 Models

In this subsection, we present the results of deepfake detection using ViT and InceptionResNetV1 models. In earlier studies, deepfake detection has been explored using CNN-based architectures such as Xception and EfficientNet. Recently, Transformers-based models have garnered increased attention due to their capabilities in feature extraction. In this paper, we compare ViT and InceptionResNetV1 separately before applying late fusion, using experiments conducted with the deepfake detection dataset employed in this work. The results of the evaluation are presented in Table 4, including accuracy, precision, recall, and F1-score for both models.

Table 4: Evaluation of ViT and InceptionResNetV1 Models

	InceptionResNetV1		**Vision Transformer (ViT)**	
	Real	*Fake*	*Real*	*Fake*
Accuracy	97.5%		97.0%	
Precision	98.3%	96.7%	97.0%	98.0%
Recall	96.6%	98.3%	98.0%	96.0%
F1 Score	97.5%		97%	

5.2 InceptionResNetV1 Model

As indicated in Table 4, both models achieved high classification accuracy in distinguishing between deepfakes and genuine content. The InceptionResNetV1 shows slightly better accuracy than ViT with an accuracy of 97.5% compared to ViT's 97.0%. InceptionResNetV1's performance is because of its deeper convolutional layers, which are more adept at capturing local spatial features and hence more attuned to fine-grained artifacts contained in deepfakes.

Figure 9 displays the training loss and accuracy curves for InceptionResNetV1 over 50 epochs, whereas Figure 10 shows its confusion matrix, which emphasizes classification accuracy.

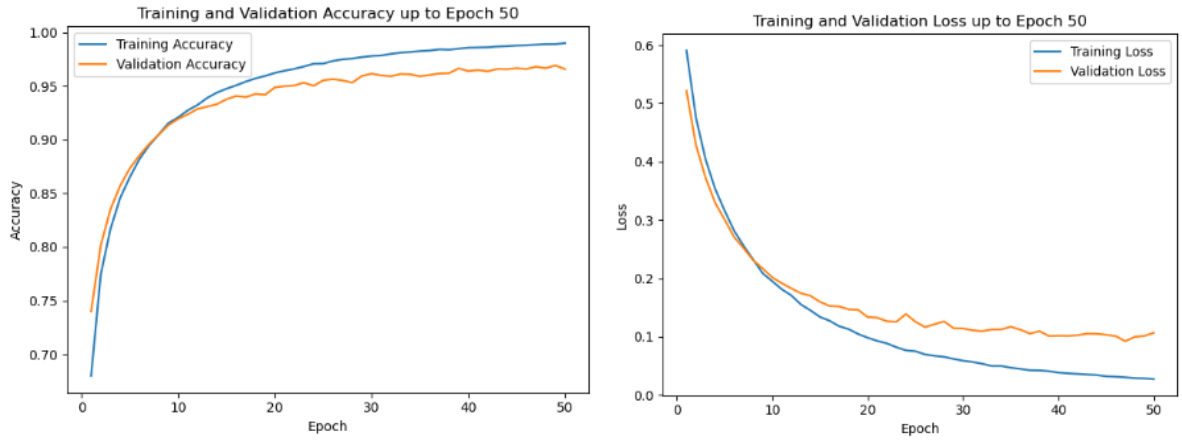

Figure 9. Accuracy and Train vs Test Loss of InceptionResNetV1 Model

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

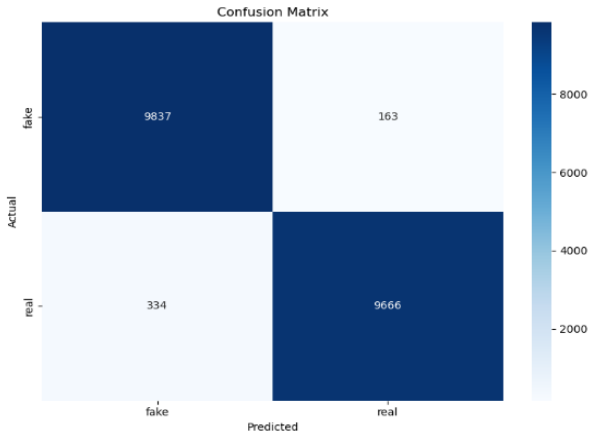

Figure 10. Confusion Matrix of InceptionResNetV1 Model

Results, as shown in Figure 9, demonstrate that training accuracy consistently improves, reaching high values throughout the epochs. At the same time, the loss function demonstrated decreasing values, indicating that the model successfully learned deepfake patterns. Figure 10 also validates, through performance in the area of the confusion matrix, that the model has good classification capacity, as precision, along with recall values, exceeds 96% in classifying images as real or fake.

5.3 Vision Transformer (ViT) Model

We tested and evaluated the pretrained ViT model before starting the fine-tuning process, so at this stage, its performance reflects its original feature extraction capability. Although it scored lower in accuracy compared to InceptionResNetV1, ViT still demonstrated good detection ability, especially in recognizing long-range relationships among facial features. The Transformer-based architecture of ViT is highly resistant to occlusions and illumination variations, which are central to deepfake detection.

We then fine-tuned the Vision Transformer (ViT) for fusion purposes on the dataset used to train the InceptionResNetV1, as described in Section 4.1.1. The model began with pre-trained parameters and was then fine-tuned to assist with binary classification, identifying real and fake images.

During the training process, we set the batch size to 16, the learning rate to 0.0001, and used an Adam optimizer to ensure stable convergence. We trained it for 10 epochs using the cross-entropy loss as the objective function to optimize classification. In the fine-tuning process, we monitored accuracy and loss performance metrics using the validation set. Figure 11 depicts a confusion matrix of the fine-tuned ViT, which attained similar, but slightly reduced, accuracy compared to InceptionResNetV1.

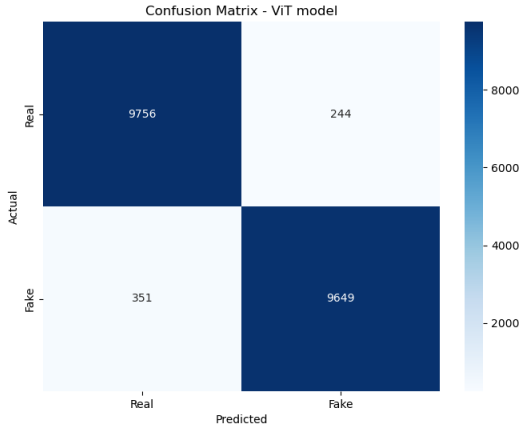

Figure 11. Confusion Matrix of ViT Model

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

After completing the fine-tuning, we test the model on the test set to examine its performance in classification. We calculate standard evaluation metrics, such as accuracy, precision, recall, and F1 score, to measure the model's performance in distinguishing between real and fake images. The analysis results provided insight into how effectively the model performed in detecting deepfakes, leveraging its transformer-based architecture for feature extraction and classification.

5.4 Comparative Analysis of Pretrained and Fine-Tuned ViT Models

This section compares the performance differences between pre-trained and fine-tuned ViT models. The analysis investigates how generalizable the pre-trained model is to novel datasets, as well as the improvements that can be acquired through fine-tuning. Through a comparison of classification results, this work demonstrates how adapting affects deepfake detection performance.

5.4.1 Model Performance Evaluation

We tested the ViT model in three configurations to examine how fine-tuning affects classification performance, as well as how the pre-trained model performs on unseen datasets. The results show that there is a noticeable improvement in classification accuracy with fine-tuning; however, in zero-shot evaluation, inherent biases in the model are revealed when it is applied to new datasets.

5.4.2 Classification Metrics Comparison

The comparative analysis across different configurations of ViT models in Table 5 shows the improvements achieved through fine-tuning and evaluates the pretrained model's ability to classify images from different datasets.

When we assessed the pretrained ViT model on the Flickr dataset, it showed poor classification ability with only 52% accuracy. Interestingly enough, recall for real images was particularly low at 24% with a corresponding level of significant misclassification rate, with many real images misclassified as fake. Our analysis is that a pretrained model, when we applied it to an unseen data set, failed to generalize, and that led to poor deepfake classifying ability. But when we fine-tuned the pretrained ViT model on the OpenForensics dataset, the model performed better with 93% accuracy. But a closer examination of recall measures demonstrates a bias for fake images, with 98% for fake images compared to a paltry 88% for real images. This signals toward a result that even though a pretrained model adapted much better for OpenForensics than for Flickr, it still demonstrated a level of bias toward mislabeling a real image as a fake image due to the intrinsic distribution of data or the original training bias level of a model. Fine-tuning a ViT model on Flickr data allowed for appreciable gains, with accuracy rising to a level of 97%. Recall measures for both fake images and real images are balanced equally, with 98% for fake images and and 96% for real images, indicating a balanced calibrated level with minimal bias level. This indicates a strong role for fine-tuning while moving a pretrained model to a fresh set of data with notable error reduction in the classification aspect as well as more stable behavior with different types of images.

Table 5: Performance Metrics Comparison of ViT Models

Model	**Dataset**	**Accuracy**	**Precision**		**Recall**		**F1 Score**	
			Real	Fake	Real	Fake	Real	Fake
PreTrained ViT	Kaggle (Flicker)	52%	55%	51%	24%	62%	34%	62%
PreTrained ViT	Kaggle (Openforensics)	93%	98%	90%	88%	98%	98%	93%
FineTuned ViT	Kaggle (Flicker)	97%	97%	98%	98%	97%	96%	97%

When we tested the pretrained ViT with the dataset used for Inception, it showed better classification ability than its original zero-shot result. Unlike its weaker performance on Flickr, the model showed strong classification accuracy in differentiating between real and fake images. This means that the dataset used to train Inception provided a structured training setting, allowing the ViT model to exploit more stable features for classifying deepfakes. However, with unseen datasets, the pre-trained ViT continued to show classification biases, further confirming that fine-tuning is necessary for enhancing flexibility.

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

Overall, these comparisons between pre-trained and fine-tuned ViT models illustrate that fine-tuning significantly improves classification accuracy while reducing dataset-specific bias. Although pre-trained models show some degree of generalizability and rely extensively on the testing set to which they are applied, fine-tuning provides more accurate deepfake detection when fine-tuned on a different set from that on which they were pre-trained.

5.5 Fusion-Based Classification Approach

The fusion-based classification technique combines the outputs of both models for advanced deepfake identification. Both models process the input separately and according to their corresponding architectural strengths before combining their outputs using an average weighing mechanism. This method makes the decision process more trustworthy based on information from both models combined.

5.5.1 Individual Model Contributions

InceptionResNetV1 is a class of convolutional neural network (CNN) architecture that focuses on extracting local features. It can detect fine-grained details in an image, which is crucial for distinguishing between fake and real images. Not to mention its ability to include spatial hierarchies, which makes it highlight disparities in pixel distributions alongside textures that you can find in fabricated images.

Unlike InceptionResNetV1, the Vision Transformer Network (ViT) captures long-range relationships and global interactions within each image. While CNNs are inherently biased towards spatial locality, ViT considers the entire image, which makes it highly effective in detecting structurally inconsistent and fake feature alignments in deepfakes.

5.5.2 Fusion Mechanism and Weighted Averaging

The fusion approach used in this study is based on the average weighting applied to both models' scores for predictions. Instead of relying on one model's class choice decision, the final decision of InceptionResNetV1 and ViT is determined as the mean of each respective confidence score from both. We believe that this strategy enables both models to contribute in a balanced manner, taking advantage of their complementary strengths.

In this setup, each model processes an input image and outputs a score for the real or fake class individually. The model then combines the results from each model and calculates their average, with a final decision based on the highest value. This approach is flexible, as the user can manually adjust the weight if needed, with more emphasis placed on a single model's output. By combining both models, the system leverages InceptionResNetV1's hierarchical feature extraction and ViT's global feature learning capabilities to develop a more robust deepfake detector structure. To illustrate the benefit of this fusion approach, we provide three visualization samples. Figure 12 presents the InceptionResNetV1 model; Figure 13 presents the ViT model, while Figure 14 presents the fusion model.

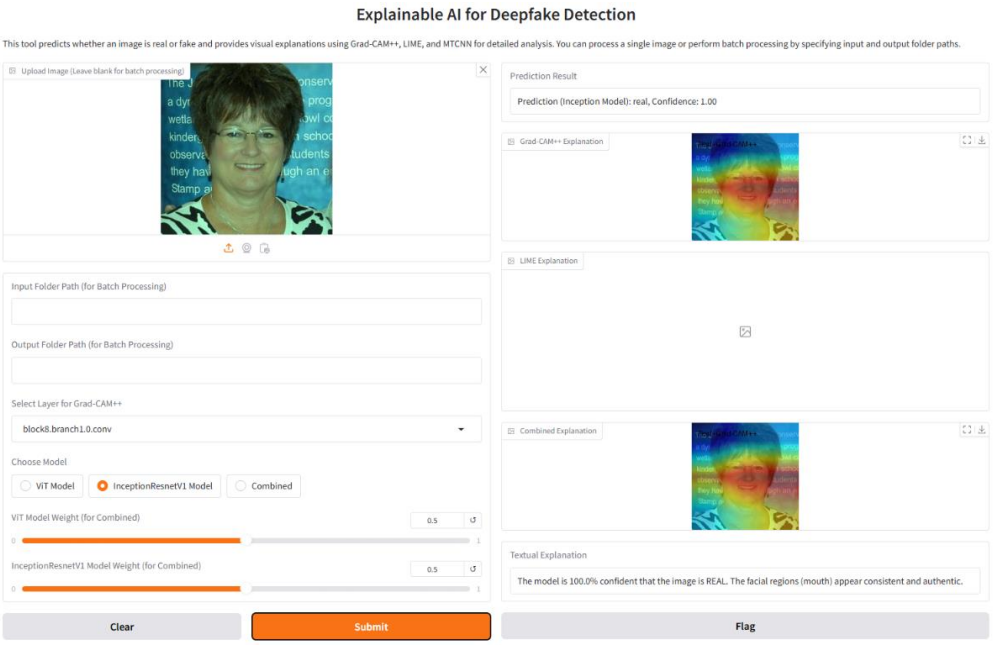

Figure 12. Visualization examples of the InceptionResNetV1 model with Grad-CAM++

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

As indicated in Figure 12, the InceptionResNetV1 model predicts the image to be real with a confidence value of 1.00. The exact figure also displays the respective Grad-CAM++ explanation, which highlights critical areas that contributed to the decision. The textual description is also present at the bottom of the figure, providing the detected anomalies in brief: "The model is 99.8% confident that the image is REAL. The facial features, including forehead, eyes, nose, and mouth, appear natural".

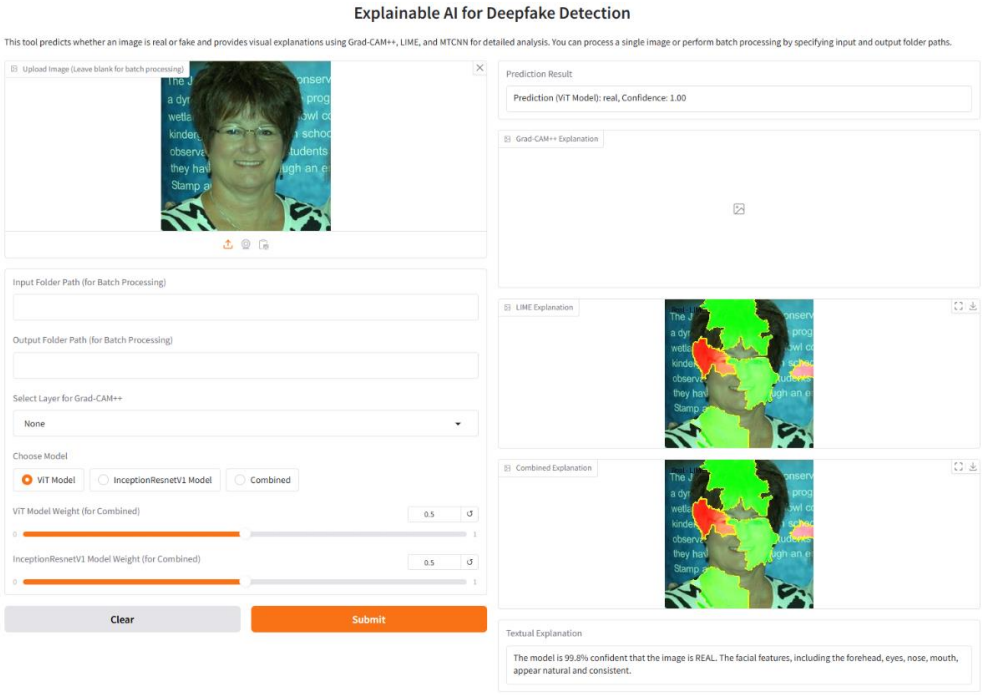

Figure 13. Visualization examples of ViT model with LIME

As indicated by Figure 13, which is for the ViT model, it also classifies the photo as counterfeit, with a confidence level of 1.00. Figure 13 presents an explanation based on LIME, illustrating regions of importance that contribute to classification. There is also a textual explanation at the bottom, elaborating on the inconsistencies that have been detected.

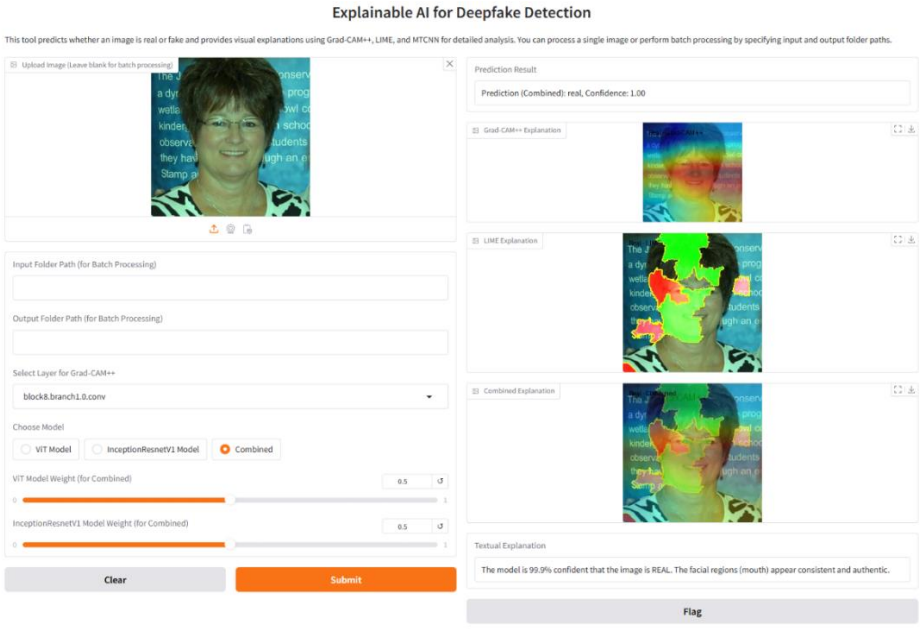

Figure 14. Visualization examples of the fusion model with Grad-CAM++ and LIME

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

As indicated in Figure 14, the average fusion method is. The model calculates the prediction based on the average prediction scores of both models. The figure also illustrates the Grad-CAM++ and LIME explanations of both models separately, as well as a combined visualization that combines both interpretability methods in a single result picture. The justification for the fused model's conclusion is presented in textual form at the bottom of the figure.

5.5.3 Evaluation of the Fusion-Based Approach

We evaluated this fusion-based deepfake detection framework to gain insight into its capabilities compared to ViT and InceptionResNetV1 models when they run individually. Using confidence-weighted decision-making, the fusion combines the complementary strengths of both models to improve classification accuracy.

The fusion model outperformed both ViT (96.2%) and InceptionResNetV1 (97.5%) in terms of performance based on key evaluation metrics, with an accuracy of 99.19%. The precision, recall, and F1-score all improved in the case of fusion. Table 7 below gives a comparative performance analysis of ViT, InceptionResNetV1, and the fusion models.

Table 7: Comparison of Model Performance

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Vision Transformer (*Real\Fake*)	96.2	0.98 / 0.94	0.94 / 0.98	0.962
InceptionResNetV1 (*Real\Fake*)	97.5	0.98 / 0.96	0.96 / 0.98	0.975
Fusion Approach (*Real\Fake*)	99.19	0.98 / 0.99	0.99 / 0.988	0.991

Moreover, Figure 15 illustrates the confusion matrix of the fusion-based approach, confirming its effectiveness in reducing misclassification errors. Compared to the individual models, the false negative rate (i.e., real images misclassified as fake) has been significantly reduced, further demonstrating the robustness of the fusion strategy.

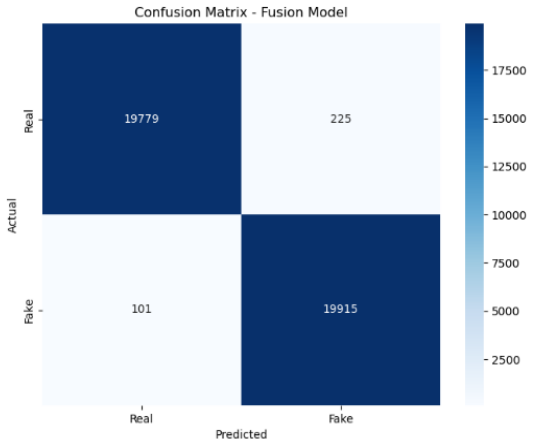

Figure 15. Confusion Matrix of the Fusion-Based Model

5.5.4 The Role of XAI in Deepfake Detection

In this section, we present the application of Grad-CAM++ on InceptionResNetV1 and LIME on ViT-based predictions, aiming to understand how both methods perceive deepfake objects. We compare both methods to evaluate how both perform in different scenarios.

5.5.4.1 Grad-CAM++ for Explaining InceptionResNetV1 Decisions

Grad-CAM++ calculates the gradient of the target class from the feature maps of a convolutional layer and weights each neuron's input for the final prediction. This allows for a more accurate visualization of the actual regions that contributed to the class decision. In deepfake detection, this approach is effective, especially in highlighting the presence of the artifacts caused by manipulation methods involving StyleGAN-generated faces or face swaps with autoencoders.

5.5.4.2 Visualizing Deepfake Detection Using Grad-CAM++

Figure 16 illustrates how Grad-CAM++ highlights the critical regions in both real and fake images classified by InceptionResNetV1.

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

Figure 16. Grad-CAM++ Explanation for Fake and Real Images

From the above heatmaps, we see that Grad-CAM++ concentrates on facial areas, especially the areas around the eyes, mouth, and forehead. On fake images, the model identifies regions of interest based on the artifacts caused by the deepfake generation methods, i.e., the inconsistencies in the textures and the abnormal color changes. On real images, the areas of activation tend to be more uniformly distributed, suggesting confidence in genuine facial characteristics.

5.5.4.3 LIME for Explaining ViT-based Decisions

LIME works by perturbing regions of an input image and observing the model prediction variations in order to understand which regions are most accountable for classifying the image. The technique is particularly well adapted for ViTs, which lack the luxury of convolutional feature maps but do apply attention mechanism-driven extraction of image global dependencies.

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

Figure 17. LIME Explanation for Fake and real Images

5.5.4.4 Visualizing Deepfake Detection Using LIME

As shown in Figure 17 above, LIME identifies specific areas of interest using Grad-CAM++. In contrast to direct facial feature detection, LIME identifies distortions in specific patches of skin texture, lighting, and abnormal transitions at boundaries. The model is more fascinated with regions that have distortions highlighted, which is parallel to the inclination for distortion in deepfakes, specifically in selective areas rather than the entire face.

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

5.2.2.4 Comparing Grad-CAM++ and LIME for Deepfake Detection

To dig deeper into the comparison between the two methods, Grad-CAM++ and LIME, as shown in Figures 18 and 19, are implemented on the same image and compared to understand how both methods interpret deepfake content.

Figure 18. Comparison of Grad-CAM++ and LIME on a Fake Image

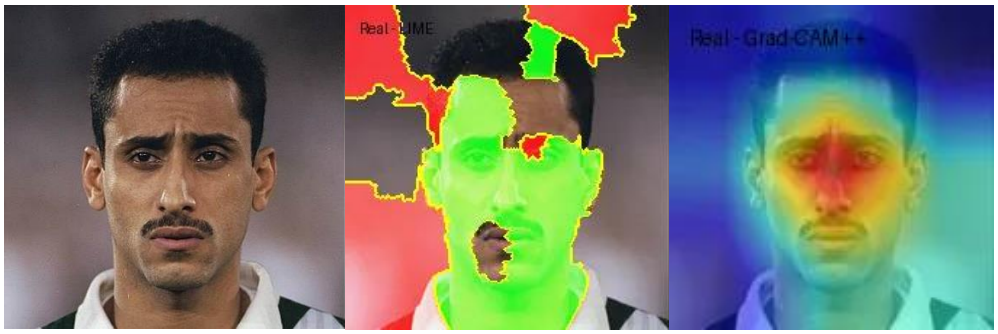

Figure 19. Comparison of Grad-CAM++ and LIME on a Real Image

The heatmap results show that Grad-CAM++ produces more localized and compact heatmaps than LIME, with larger areas of influence. It suggests that Grad-CAM++ performs better in indicating clear areas that impact a decision of a CNN, while LIME tends to show a broader scope of an image's impact on the Transformer model.

The incorporation of XAI methodologies, including Grad-CAM++ and LIME, does not affect the performance of detection in a direct way but greatly facilitates increased interpretable deepfake detection models and decision reasoning. Grad-CAM++ particularly identifies salient face areas used by InceptionResNetV1, and LIME identifies patch-level feature influence in ViT classification. With both methodologies' incorporation, more explainable deepfake tools result with increased confidence in AI-based tools for media tampering identification. The future work potential area might include the incorporation of additional XAI methodologies for additional interpretability for deepfake detection models.

6. Conclusion

This paper proposes a deepfake detection system that combines Vision Transformer (ViT) and InceptionResNetV1 models through weighted averaging fusion. The synergistic collaboration of ViT's ability to learn global dependencies with the local feature extraction strengths of InceptionResNetV1 results in high classification accuracy with interpretability. The use of Explainable AI (XAI) techniques such as Grad-CAM++ and LIME increases transparency through the provision of visual and analytical explanations of model choices. Experimental results prove the fusion-based technique outperforms standalone models with an accuracy of 99.19%. This validates the fusion of convolutional and transformer-based models for enhancing deepfake detection performance. The use of an explainability framework ensures interpretability of predictions by the model, hence suitability of the system for use in security and forensic analysis.

7. Limitations and Future work

The proposed technique is promising in accuracy and interpretability but can be enhanced in subsequent work by increasing training as well as evaluation datasets to cover various forms of deepfakes to improve robustness to evolving forms of manipulations. Computationally efficient optimization will also be important in enabling

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

deepfake detection in real-time, especially in use in video streaming and social media tracking. Addressing these will make deepfake detection more reliable and efficient, as well as suitable for use in real-world applications.

Another limitation of our study is that it accepts only static images. While our models perform well on individual frames, we recognize the importance of analysing video data for real-world applications. Video-based deepfake detection presents additional challenges, such as capturing temporal dependencies and managing varying frame quality, which our current approach does not address.

Regarding the scalability of our model to video data, the weighted average fusion method can be applied either frame-by-frame or on aggregated frame-level features; this allows direct extension to temporal data. This simplicity avoids the computational overhead and complexity found in attention-based or other learned fusion methods that include temporal modelling components. We plan to explore temporal fusion techniques in future work to enhance video deepfake detection performance.

Acknowledgement: "The authors gratefully acknowledge Qassim University, represented by the Deanship of Graduate Studies and Scientific Research, on the financial support for this research under the number (QU-J-PG-2-2025-53141) during the academic year 1446 AH/2024 AD"

Conflicts of Interest: "The authors declare that there are no conflicts of interest regarding the publication of this research. No financial, personal, or professional relationships have influenced the study's design, implementation, or findings."

References

[1] G. Vecchietti, G. Liyanaarachchi, and G. Viglia, "Managing deepfakes with artificial intelligence: Introducing the business privacy calculus," *J. Bus. Res.*, vol. 186, p. 115010, Jan. 2025, doi: 10.1016/j.jbusres.2024.115010.

[2] F. Abbas and A. Taeihagh, "Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence," *Expert Syst. Appl.*, vol. 252, p. 124260, Oct. 2024, doi: 10.1016/j.eswa.2024.124260.

[3] Amerini *et al.*, "Deepfake Media Forensics: Status and Future Challenges," *J. Imaging*, vol. 11, no. 3, p. 73, Feb. 2025, doi: 10.3390/jimaging11030073.

[4] R. Babaei, S. Cheng, R. Duan, and S. Zhao, "Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis," *J. Sensor Actuator Netw.*, vol. 14, no. 1, p. 17, Feb. 2025, doi: 10.3390/jsan14010017.

[5] M. Nagm *et al.*, "Detecting image manipulation with ELACNN integration: a powerful framework for authenticity verification," *PeerJ Comput. Sci.*, vol. 10, 2024, doi: 10.7717/peerj-cs.2205.

[6] G. A. Pereira and M. Hussain, "A review of transformer-based models for computer vision tasks: Capturing global context and spatial relationships," *arXiv*, 2024.

[7] J. Maurício, I. Domingues, and J. Bernardino, "Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review," *Appl. Sci.*, vol. 13, no. 9, p. 5521, Apr. 2023, doi: 10.3390/app13095521.

[8] V. Bengani, "Hybrid Learning Systems: Integrating Traditional Machine Learning with Deep learning Techniques," *Tech. Rep.*, 2024, doi: 10.13140/RG.2.2.10461.22244/1.

[9] A. Adeniran, A. P. Onebunne, and P. William, "Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical decision-making," *World J. Adv. Res. Rev.*, vol. 23, no. 3, pp. 2447–2658, Sep. 2024, doi: 10.30574/wjarr.2024.23.3.2936.

[10] S. Raghuvanshi, "Machine Explainability: A Guide to LIME, SHAP, and Gradcam," *Medium*. Accessed: Mar. 16, 2025. [Online]. Available: https://suryansh-raghuvanshi.medium.com/machine-explainability-a-guide-to-lime-shap-and-gradcam-60f6265f365f

[11] V. L. L. Thing, "Deepfake Detection with Deep Learning: Convolutional Neural Networks versus Transformers," in *Proc. IEEE Int. Conf. Cyber Secur. Resilience (CSR)*, Jul. 2023, pp. 246–253, doi: 10.1109/CSR57506.2023.10225004.

[12] B. Kaddar, S. A. Fezza, W. Hamidouche, Z. Akhtar, and A. Hadid, "HCiT: Deepfake Video Detection Using a Hybrid Model of CNN features and Vision Transformer," in *Proc. Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2021, pp. 1–5, doi: 10.1109/VCIP53242.2021.9675402.

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025

[13] S. A. Khan and D.-T. Dang-Nguyen, "Hybrid Transformer Network for Deepfake Detection," in *Proc. Int. Conf. Content-based Multimedia Indexing*, New York, NY, USA: ACM, Sep. 2022, pp. 8–14, doi: 10.1145/3549555.3549588.

[14] Koçak, M. Alkan, and S. M. Arıkan, "Deepfake Video Detection Using Convolutional Neural Network Based Hybrid Approach," *J. Polytechnic*, Sep. 2024, doi: 10.2339/politeknik.1523983.

[15] W. H. Abir *et al.*, "Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods," *Intell. Autom. Soft Comput.*, vol. 35, no. 2, pp. 2151–2169, 2023, doi: 10.32604/iasc.2023.029653.

[16] H. Soudy, O. Sayed, H. Tag-Elser, *et al.*, "Deepfake detection using convolutional vision transformers and convolutional neural networks," *Neural Comput. Applic*, vol. 36, pp. 19759–19775, 2024, doi: 10.1007/s00521-024-10181-7.

[17] Cavia, L. Huang, and R. Smith, "Real-Time Deepfake Detection in the Real-World," *arXiv: 2406.09398*, Jun. 2024. [Online]. Available: https://arxiv.org/abs/2406.09398

[18] F. Wodajo, T. Mekonnen, and G. Alemu, "Deepfake Video Detection Using Generative Convolutional Vision Transformer," *arXiv: 2307.07036*, Jul. 2023. [Online]. Available: https://arxiv.org/abs/2307.07036

[19] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[20] R. Mubarak *et al.*, "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats," *IEEE Access*, vol. 11, pp. 144497–144529, 2023, doi: 10.1109/ACCESS.2023.3344653.

[21] xhlulu, "140k Real and Fake Faces," *Kaggle*. Accessed: Mar. 21, 2025. [Online]. Available: https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces/data

[22] K. Manjil, "deepfake and real images," *Kaggle*. Accessed: Mar. 21, 2025. [Online]. Available: https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images

[23] Hugging Face, "ViT-Base-Patch16-224 Model Card," *Hugging Face*. Accessed: Mar. 21, 2025. [Online]. Available: https://huggingface.co/google/vit-base-patch16-224/tree/main

[24] R. R. Selvaraju *et al.*, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[25] J. An, Y. Zhang, and I. Joe, "Specific-Input LIME Explanations for Tabular Data Based on Deep Learning Models," *Appl. Sci.*, vol. 13, no. 15, p. 8782, Jul. 2023, doi: 10.3390/app13158782.

[26] B. Smith, J. Doe, and R. Johnson, "Deepfake Detection Using Machine Learning Techniques: A Survey," *J. Mach. Learn. Res.*, vol. 25, no. 1, pp. 1-30, Jan. 2024.

DOI: https://doi.org/10.54216/FPA.210205
Received: March 09, 2025 Revised: June 03, 2025 Accepted: July 01, 2025