INSTITUTE OF INFORMATION
AND COMMUNICATION TECHNOLOGIES
BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **26**, No 2
Sofia • 2026 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2026-0014

Trustworthy Deepfake Detection: Explainable LIME Method of ViT and CNN Architectures

Zoulikha Koudad[1,4], *Amina Bekkouche*[2,4], *Hamed Benahmed*[2,4], *Mourad Hadjila*[3,5], *Mohammed Merzoug*[2,4]

[1]*Department of First Cycle, École Supérieure en Sciences Appliquées de Tlemcen ESSAT, Tlemcen 13000, Algeria*
[2]*Department of Computer Science, Faculty of Science, University of Tlemcen, Tlemcen 13000, Algeria*
[3]*Department of Telecommunication, Faculty of Technology, University of Tlemcen, Tlemcen 13000, Algeria*
[4]*LRIT Laboratory, Faculty of Science, University of Tlemcen, Tlemcen 13000, Algeria*
[5]*STIC Laboratory, Faculty of Technology, University of Tlemcen, Tlemcen 13000, Algeria*
E-mails: zoulikha.Koudad@essa-tlemcen.dz amina.bekkouche@univ-tlemcen.dz hamed.benahmed@univ-tlemcen.dz mourad.hadjila@univ-tlemcen.dz mohammed.merzoug@univ-tlemcen.dz

***Abstract*:** *The rapid evolution of generative artificial intelligence has enabled the creation of highly realistic deepfake facial imagery, supporting innovative applications in filmmaking and digital media while simultaneously amplifying risks related to misinformation and public safety. As a result, deepfake detection has become a critical research priority that must combine strong predictive performance with transparent and trustworthy decision-making, since deep learning models remain largely opaque and difficult to interpret in sensitive judicial and information-critical contexts. In this work, we develop and evaluate five deepfake detection architectures, including a Vision Transformer and four Convolutional Neural Networks, trained on the 140K Real and Fake Faces dataset, with the best models achieving an accuracy of 95 percent. To address the fundamental challenge of explainability, we further integrate the LIME interpretability framework, which generates clear and visually intuitive explanations of model decisions, thereby enhancing transparency and strengthening user confidence in automated deepfake analysis.*

***Keywords*:** *Deepfakes face detection, Social security, Explainable Artificial Intelligence, Vision transformer, Convolutional Neural Networks.*

1. Introduction

The significant progress achieved in deep learning and generative AI methods has led to the creation of highly realistic synthetic images of people. While this can be useful in producing documentaries by recreating historical scenes or in reducing

production costs for films, it may also enable malicious actors to manipulate information. This phenomenon is known as deepfakes, a relatively recent concept. A deepfake refers to visual or audio content that is generated or altered using advanced artificial intelligence techniques, particularly deep learning, to make a person appear to say or do something they never actually said or did [1, 2].

With the remarkable progress of generative AI, deepfakes have become highly realistic and increasingly difficult to detect, which has facilitated the spread of propaganda for political purposes and poses a serious threat to social security. However, since deepfakes can also be used for beneficial purposes by charitable organizations, advertising companies, and other creative industries, several deepfake-generation platforms are publicly available on the web, including DeepFaceLab [3], FaceSwap [4], FaceFusion [5], and FaceChange Technologies [6]. Hence, there is a need to develop detection systems that are as intelligent and advanced as the generative systems themselves for effective DeepFake detection. The objective is not to design models capable of detecting clearly noticeable deepfakes, as these do not represent a significant threat; most viewers can identify them immediately. The real concern arises when deepfakes become subtle, harder to detect, and capable of misleading people. Many studies on DeepFake face detection have emerged, most of which rely on deep learning models. These models are often trained on datasets containing face images with clearly visible artifacts, such as obvious warping or lifting effects, or on datasets that do not include a sufficient number of high-quality fake images [7, 8]. Among the existing methods, some approaches detect specific facial features or artifacts within the images to determine whether they are manipulated [9-11]. Other approaches rely on Convolutional Neural Networks (CNNs) for deepfake detection [12-14].

Another issue arises from the fact that fake facial images are often indistinguishable to the human eye. On what basis does an effective DeepFake detection model make its decision? How does it manage to differentiate between real and fake content? There must be specific facial regions that lead the model to classify an image as fake. These regions are expected to vary from one image to another; otherwise, the manipulation would be easily detectable either by humans or by simpler models. To address this, many works have attempted to explain the decisions of their models using eXplainable Artificial Intelligence techniques (XAI) [15-17]. These are model-specific methods that interpret predictions based on the model's internal structure and the distribution of weights across its convolutional layers, such as Grad-CAM and Network Dissection [18, 19]. In contrast, T s i g o s et al. [20] compared several explainability methods and concluded that the perturbation-based method LIME [21-23] was the most effective for highlighting the regions that influence the model's decisions. This motivated the direct adoption of LIME as the chosen method for model explainability.

This paper presents five models for DeepFake face detection, including a customized small Vision Transformer (ViT-S/16) [24] and four CNN-based architectures built with transfer learning from VGG-16 [25], ResNet-50 [26], Inception-V3 [27], and MobileNet-V2 [28]. Transformers have already demonstrated strong capabilities in image classification and fine-grained structure

discovery, while chosen CNNs are pretrained on large-scale datasets “ImageNet” [29], and are known for their high classification power across thousands of image categories. Including both families of models, therefore, provides a robust and representative baseline for evaluating Deepfake face detection performance. All models are trained on the 140K Real and Fake Faces dataset. Importantly, the dataset is used independently, without combining it with other datasets where facial manipulations are more obvious, ensuring that the models learn to detect subtle and highly realistic deepfakes, a particularly challenging task since their differences are not easily discernible to the human eye. The performances of the five models are evaluated. To enhance transparency, the local perturbation method LIME is applied to explain the models’ decisions. For each face image, LIME highlights the specific regions that had the greatest influence on the model’s prediction, providing a visual and interpretable explanation of its reasoning.

2. Related work

2.1. Deepfake face detection

Although the field of deepfake face detection is quite recent, it has developed rapidly. Many studies have emerged using deep learning techniques as transfer learning from CNNs and Transformers to detect image features or perform binary classification between real and fake faces.

Among the studies based on CNNs, Jheelan and Pudaruth [14] used four CNNs with transfer learning and one simple original CNN, validated on a subset of the 140k Real and Fake dataset and the 20k_gan_8_1_1 dataset. However, since training, validation, and testing were conducted on only a small subset of the available dataset, the reported performance does not provide sufficient evidence of the models’ generalization ability, despite F1-scores ranging from 85% to 98%. Similarly, in the work of Sharma et al. [12], the authors employed a ResNet50 and a VGG16, along with a customized network composed of two convolutional layers. The final classification result was obtained through a voting mechanism among the three networks, which improved the overall accuracy. The combination of the results from three different models provided a higher likelihood of correctly classifying the faces. Naeem et al. [13] compared the performance of eight models, six CNNs, one DenseNet, and one Vision Transformer, for classifying real, fake, and synthetic images. The best results were achieved by the Vision Transformer, EfficientNet, and MobileNet models. This work addresses a three-class classification problem. A notable strength of the study is that it reports the performance of multiple models, thereby providing a valuable benchmark for future researchers when selecting suitable approaches. However, the fact that the study also focuses on detecting synthetic classes that can be easily identified by image processing techniques and even by the human eye does not constitute a real challenge, since the most critical cases are those in which people cannot distinguish between real and fake content.

Other methods perform image feature analysis, such as the work of Gong et al. [30], who conducted both global and local texture analysis to detect traces of

forgery and transition using specialized CNNs. The detected features are incorporated into a pyramidal module, followed by an attention-based module. This work is substantial, the model is thoroughly studied and composed of several sub-modules, but it was tested on small datasets, the largest of which is FaceForensics++, which, as shown in Fig. 2, contains many facial images that are visibly fake. Kim [11] employed a shallow but wide convolutional network to analyze texture and detect facial forgeries, focusing on high-frequency components such as redness and color variations, which are more difficult to reproduce. This approach achieved a low error rate on the CASIA-FASD, CelebA-Spoof, and NIA-ILD datasets. In this work, face forgery through manual attacks, such as replacing the face with a printed photo or a drawing of a cat, does not represent the challenge of high-level forgery that is undetectable by the human eye. Similarly, Liu et al. [10] proposed a variational autoencoder to learn residual feature maps for face forgery detection. A federated learning strategy was then applied to develop a decentralized detection model. The advantage of this federated learning approach lies in its ability to learn heterogeneous types of falsifications while preserving privacy; however, its detection performance remained moderate compared to other methods. Shi et al. [9] introduced the RFFR (Representation of Fundamental Facial Representations) approach, which trains the model on real images using Masked Image Modeling to generate residual blocks. The residual blocks of forged images exhibit different distributions, later distinguished by a Vision Transformer for classification. Despite the method's strong reported performance, the distortions highlighted in the examples appeared sufficiently pronounced to be easily detected without automated assistance.

Other approaches make use of transformers, such as Xiao et al. [31], who proposed a multi-branch adaptive attention mechanism by hybridizing an Efficient Vision Transformer with a multi-level wavelet transform for fake face detection. This work was validated on the FaceForensics++ and Celeb-DF (V2) datasets. Although the authors conducted a thorough and well-structured experimental study, the lifting effect visible in the dataset images was pronounced. And the same observation applies to the work of Man and Cho [32] that combined a Transformer with a GAN to learn fake image features, integrating this module with an Efficient network for frequency and noise detection, whose outputs are then passed to a classification module. Finally, Abirami [33] employed a Vision Transformer to detect falsified images shared on social media platforms. Although this approach is both effective and state-of-the-art, the study is constrained by the limited size of the datasets.

2.2. Deepfake face detection explainability

Intelligent systems have always been difficult to interpret or understand for users, but a specialist or developer could provide the necessary explanation to understand the decision-making process of such systems. However, with the development of deep learning, it has become impossible, even for the specialist designer, to explain on what basis a CNN or a Transformer made its decision. This is why the use of

XAI methods has become necessary. Similarly, in the field of deepfake detection, the use of explainability methods is increasingly demanded.

Mansoor and Iliev [19] applied the Network Dissection algorithm to explain the decision-making process in CNNs for fake face detection. The dissection method enables understanding the behavior of CNN units in real-versus-fake classification. This approach relies on an accessible concept dictionary. The model performs a forward pass of the dictionary images to extract neuron activation maps and measure overlaps between activated units and image regions. While the method helps explain the behavior of CNN neurons, it does not provide fine-grained detail about the specific small regions responsible for a fake decision, which remains a limitation, particularly from a human interpretability perspective. Tariq et al. [18] used Grad-CAM activation on a CNN model, which allows highlighting the regions of the image that most influenced the CNN's decision. The Grad-CAM explainability method derives its information from the activation maps of the last convolutional layer and projects this information onto the input image to highlight the regions that strongly influenced the CNN's decision. While this method provides a useful visual interpretation, it can sometimes be imprecise with respect to fine-grained details and is primarily applicable to CNN-based models. Cirillo, Gervasio and Amerini [34] used perturbation-based explainability methods on segmented images (these images are not necessarily faces), then applied gradient-based methods by masking the regions that influence the results the most or the least, to test and evaluate the robustness and vulnerability of existing models with high generalization capability. Perturbation-based methods involve masking the regions that have the greatest or least influence on the classification, serving as the employed explainability approach. The modified image is subsequently fed back into the real-versus-fake classifier to evaluate the model's robustness. Tsigos et al. [20] compared gradient-based explainability methods and perturbation-based methods on the EfficientNet model for deepfake detection. The evaluation was carried out on images from the FaceForensics++ dataset. The authors use Grad-CAM++, which extends the Grad-CAM method. It is also applied to the last convolutional layer of a CNN, providing greater accuracy. RISE and LIME are based on input image perturbations, while SHAP and SOBOL are attribution-based methods. The authors evaluated the explanations provided by each explainability method through adversarial image generation and evaluation. According to this quantitative evaluation, LIME proved to be the most effective method, a finding that was also confirmed in the qualitative assessment. LIME successfully highlighted the manipulated regions in fake images. This paper provides a highly useful comparative study for future research.

From the related work, it can be observed that some studies focused on deepfake detection without considering the trust that should be granted to users in understanding how these systems operate, given that the models are black boxes with no interpretability. Moreover, most of these methods rely on small datasets or datasets that partly contain highly noticeable lifting effects on facial images. A second group of studies concentrated on explainability, without ensuring the effectiveness of their models. From their experiments, it became evident that

explainability methods dependent on the model's architecture can overlook subtle facial details that may influence classification **and are only applicable to a specific type of model.**

Unlike previous works, we opted to develop a highly effective fake face detection system, tested on a comprehensive dataset where it is extremely difficult for humans to distinguish real from fake. To ensure the reliability of our approach, we selected an explainability method that is both model-agnostic and local, meaning that the features influencing the classification of each image are specific to that image alone. Consequently, our system is trustworthy, and every classification is justified.

3. Materials and methods

This section provides a detailed description of our work, starting with the dataset we used and the preprocessing steps applied to the data. Then, it presents the first part of our study, which concerns deepfake face detection, and concludes with our proposed explainability approach for the developed models. The steps of our work are visualized in Fig. 1.

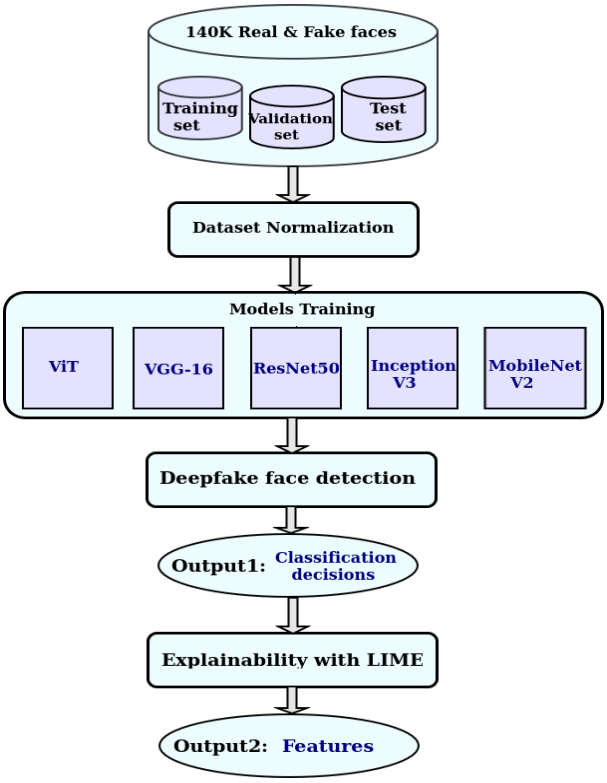

Fig. 1. Pipeline of the proposed deepfake face detection method, integrating five classifiers and a LIME-based explainability module

3.1. Dataset

We chose to work on the 140k Real and Fake Faces dataset [35], which has been attracting increasing attention in the field of deepfake detection. This dataset contains 140,000 face images, including 70,000 real faces collected from the Flickr dataset provided by NVIDIA. Flickr, [36] is a high-quality human face dataset created for Generative Adversarial Networks (GANs).

The 70,000 real face images included in the 140k Real and Fake Faces dataset are of high quality and highly diverse. They feature faces of different ages, ethnicities, and backgrounds, as well as individuals wearing glasses (vision or sunglasses), hats, or other accessories, and even well-made-up faces.

The dataset is well balanced, as it contains an equal number of fake faces (70,000) selected from the 1Million Fake Faces dataset [37] generated by StyleGAN and made available by T u n g u z [38] on Kaggle.

All images in the 140k Real and Fake Faces dataset are uniformly resized to 256×256 pixels, and the dataset is provided split into three subsets:

- Training set: 100k images (50k real, 50k fake),
- Validation set: 20k images,
- Test set: 20k images,

Fig. 2. Images (a)-(d) are clearly fakes, as the deformations are quite obvious; these samples are taken from the FaceForensics++ dataset. In contrast, images (e)-(l) come from the 140k Real and Fake Faces dataset, where it is not easy to distinguish which faces are real and which are fake. Specifically, the faces in (e)-(h) are real, while those in (i)-(l) are fake

The main advantage, or rather, the challenge posed by the 140k Real and Fake Faces dataset lies in the subtle differences between real and fake images, which are not easily distinguishable to the human eye. Although the one Million Fake Faces dataset contains many clearly fake or heavily distorted images, the fake faces selected for the 140k Real and Fake Faces dataset show no visible artifacts or deformations. This contrasts with other widely used datasets in the field, which often exhibit an obvious Photoshop effect in many fake images, as illustrated in Fig. 2.

The challenge of the 140k Real and Fake Faces dataset is that it is extremely difficult for humans to tell real and fake faces apart, making the task even more challenging for CNN or Transformer-based models. This naturally raises the issue of explainability: If humans cannot perceive the visual cues that distinguish fake from real images, what features or elements does the artificial model rely on to make its classification decisions?

Many current methods apply transformations to images to increase the size of the dataset. Data augmentation consists of generating multiple new images from the original dataset images by applying transformations such as shifting, rotation, and others. These transformations can be considered as alterations of the original image, while augmented real-face images are generally still classified as real. On the other hand, cropping may remove parts of the face that contain important clues for distinguishing fake images. Image resizing can also remove important information when the image is downscaled or introduce smoothing effects when it is upscaled. For these reasons, we do not apply any transformations to the images in our dataset, and we do not require data augmentation since we have the same number of real and fake images. The only transformation we apply is normalization, in order to adapt the images to the neural networks.

3.2. Model architectures

Five deepfake face detection models were developed, all performing binary classification on facial images. The first model is a customized Vision Transformer. In addition, four CNN-based models were designed:

- VGG-16, composed of 16 convolutional layers, using small filters of 3×3.
- ResNet50, based on residual learning, a technique that enables very deep networks (50-layer) to train without vanishing or exploding gradients. ResNet won first place in the ILSVRC2015 classification task and first place in the ImageNet classification benchmark.
- InceptionV3, which reduces the number of parameters through convolution factorization, replacing large filters with smaller asymmetric ones. The architecture consists of three main modules (A, B, and C) and includes 48 convolutional layers.
- MobileNetV2 introduces inverted residual blocks and linear bottlenecks to create a deep yet computationally efficient CNN. Its architecture is composed of 17 inverted residual blocks, each structured with channel expansion, a 3×3 depthwise convolution, and a linear bottleneck.

These CNN architectures were originally pretrained on the ImageNet dataset for a 1000-class classification task. Consequently, the classifier component of each

CNN was reimplemented, and transfer learning was applied to adapt the models to the binary deepfake detection task. Thus, the modified model was trained on the 140K Real and Fake Faces dataset.

Vision transformer. The Small Vision Transformer model was customized by following the baseline ViT architecture [24], which is a direct adaptation of the original Transformer designed for NLP tasks [39].

In the ViT-S/16 implementation, each image is divided into 16×16 patches that are flattened and passed to the Transformer encoder, as shown in Fig. 3. Since the task is image classification, the decoder component of the original Transformer architecture is not required.

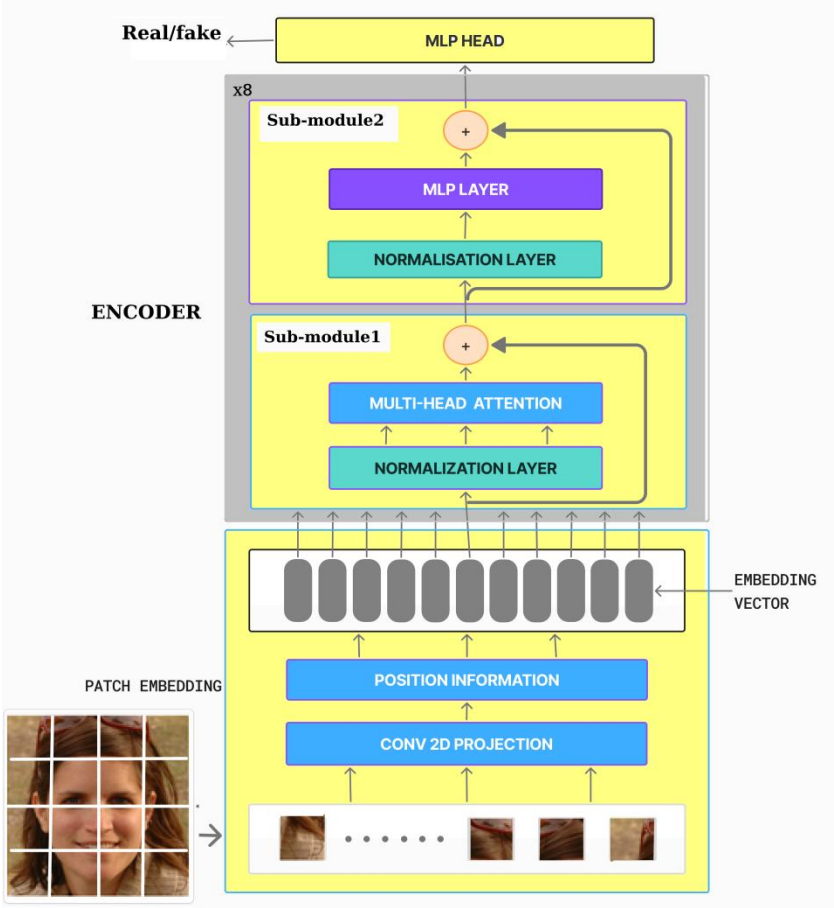

Fig. 3. Architecture of our vision transformer

A convolutional layer is used to project each patch into a 128-dimensional embedding vector. Positional encodings are then added to the patch embeddings, transforming the image into a sequence of embedding vectors.

These vectors are then passed to our Transformer encoder, which is constructed by stacking eight blocks. Each block consists of two sequential submodules. The first submodule contains a Layer Normalization layer followed by

a multi-head self-attention mechanism, with a residual connection between the submodule's input and output. The second submodule also begins with a Layer Normalization layer, followed by an MLP equipped with its own residual connection. The classification module of the proposed ViT-S/16 architecture is implemented as a single fully connected layer.

3.3. Explanation of the model decision using LIME

It is no longer sufficient to design a powerful deepfake face detection model; black-box models fail to garner trust among users, so the need to explain the basis of the model's decision has become crucial, both for the end-user and for the model developer. We need to understand which regions of the face the model relies on to determine that an image is fake, especially since it is often very difficult for humans to make this distinction.

Having implemented models with different architectures, including CNNs and Vision Transformers, we therefore chose to use an explainability method that is model-architecture agnostic. And because of the particularity of the problem of deep fake detection, a local interpretation method is suitable [20]. We adopted the LIME (Local Interpretable Model-agnostic Explanations) method, which identifies the areas of an image that most influence the model's decision. Since LIME is a model-agnostic approach, it can be applied to any type of classifier [21]. Unlike other interpretability techniques, LIME does not attempt to analyze the model's internal architecture (such as the weights of a CNN) but rather focuses on the relationship between the model's inputs and outputs. The method generates explanations for individual instances; therefore, it provides local interpretability.

For a given image x, LIME defines explainability in terms of a model $g \in G$, where G denotes a class of interpretable models and g admits a visual representation. The term $\Omega(g)$ quantifies the complexity of the explanation model.

Initially, an image segmentation algorithm partitions the image into homogeneous regions, referred to as superpixels, which share similar properties such as color, texture, and intensity.

Let the classification model be represented as a function f, where $f(x)$ denotes the predicted class label (real or fake) for the image x.

Let $Z = \{z_1, z_2, \ldots, z_N\}$ be the set of perturbed samples derived from x by masking subsets of its superpixels. A similarity measure $\pi_x(z_i)$, i=1, …, N, is then computed between each perturbed instance z_i and the original image x. Each perturbed image is then evaluated by the classifiers (ViT-S/16, VGG16, ResNet50, InceptionV3, MobileNetV2). The perturbed samples, together with their corresponding predictions, form a cloud of points around x. The distance of each point from x is important and is used to compute a proximity weight relative to x.

Let $\mathcal{L}(f, g, \pi_x)$ represent the degree to which g fails to approximate f at the point defined by π_x. The explanation is then obtained by solving the equation

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}}\, \mathcal{L}(f, g, \pi_x) + \Omega(g). \quad (1)$$

The resulting explanation corresponds to a set of weights assigned to each superpixel such that the interpretable model g locally approximates f in the neighborhood of x.

The points closest to x have a greater influence in generating a simple function g that approximates our deepfake detection model in the neighborhood of x. The parameters of g. Thus, it represents the local importance of each feature (superpixel) in the classifier's decision.

4. Results and interpretation

This section discusses the deepfake face detection results obtained with our Vision Transformer (ViT) model as well as with the four CNN-based models. And then presents the interpretability results.

4.1. Deepfake face detection

In the experiments, images from the 140k Real and Fake Faces dataset were employed, with the original resolution of 256×256 preserved throughout, as specified in the previous section. Regarding the parameters of our models, the ViT contains eight Transformer blocks. In each block, the first sub-module includes a Layer Normalization followed by a Multi-Head Self-Attention mechanism with eight attention heads. The second sub-module contains a Layer Normalization followed by an MLP composed of two linear layers separated by a GELU activation function. At the output of the Transformer encoder, we added a linear MLP layer that acts as a classifier.

Before entering the Transformer blocks, each image is divided into 16×16 patches, which are then converted into embedding vectors of size 128. Our ViT was trained for 30 epochs with a batch size of 64, using the AdamW optimizer and the binary cross-entropy loss function.

For the CNN-based models (VGG16, ResNet50, InceptionV3, and MobileNetV2), we replaced the original classifier with an MLP composed of three dense layers with ReLU activations, followed by a final layer with a sigmoid activation. These models were trained for 30 epochs with a batch size of 128, using the Stochastic Gradient Descent (SGD) optimizer and the binary cross-entropy loss function. Table 1 summarizes the parameters of all the models.

Table 1. Training configuration of the five deepfake detection models

Training	Epochs	Batch size	Optimizer	Loss function
ViT-S/16	30	64	AdamW	Binary cross entropy
VGG16	30	128	SGD	Binary cross entropy
ResNet50	30	128	SGD	Binary cross entropy
InceptionV3	30	128	SGD	Binary cross entropy
MobileNetV2	30	128	SGD	Binary cross entropy

4.2. Performance evaluation

The performance results of the five deepfake face detection models on the 140k dataset are summarized in Table 2. Training and validation accuracy, as well as training and validation loss, are reported for each architecture. The results clearly demonstrate the superiority of the ViT model, which attains a training accuracy of 0.9848 and a validation accuracy of 0.9452, along with a minimal training loss of 0.0410 and a validation loss of 0.2064. MobileNetV2 ranks as the second most effective model, followed by InceptionV3, VGG-16, and finally ResNet50.

Table 2. Training and validation accuracy and loss for the five deepfake face detection models

Model	Train accuracy	Validation accuracy	Train loss	Validation loss
ViT-S/16	**0.9848**	**0.9452**	**0.0410**	**0.2064**
VGG-16	0.8516	0.8379	0.3771	0.3600
ResNet50	0.5859	0.6126	0.6653	0.6126
InceptionV3	0.8828	0.8625	0.2832	0.3199
MobileNetV2	0.9295	0.8932	0.1773	0.2638

The training process of our Vision Transformer is summarized by the loss evolution curve and the accuracy evolution curve for both the training and validation sets, as shown in Fig. 4. The curves reflect the successful training process. Naturally, at some point, overfitting begins to appear, which in our case occurs around epoch 18, where the validation curves start to diverge from the training curves. The best model was therefore selected as the one saved just before the onset of overfitting.

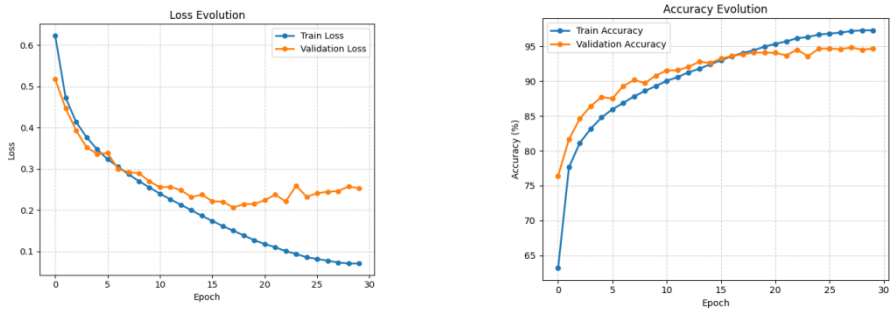

Fig. 4. Loss evolution for ViT Training (a); Accuracy evolution for ViT training (b)

Subsequently, the five models are evaluated on the test dataset, which consists of 20,000 facial images, evenly balanced between 10,000 fake and 10,000 real samples. The test results of the ViT-S/16 model are illustrated by the confusion matrix, as shown in Fig. 5. The model correctly identifies 95.34% of real faces and 93.73% of fake faces. However, 6.27% of fake faces are misclassified as real, and 4.66% of real faces are misclassified as fake. These results highlight not only the robustness of the Transformer architecture but also the effectiveness of the training process, especially given the challenging nature and high variability of the dataset.

Such performance demonstrates the model's strong generalization ability and its suitability for realistic deepfake face detection scenarios.

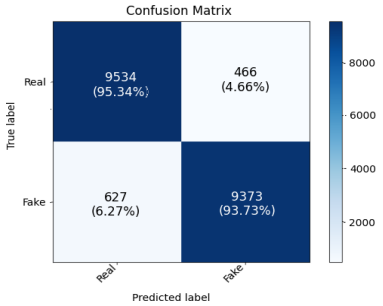

Fig. 5. ViT confusion matrix

The test results for the five models are also expressed in terms of precision, which indicates the rate of correctly classified elements among a predicted class, and recall, which represents the rate of correctly classified elements within a given actual class. The most important metric is the recall of the fake class, as it measures the rate of fake samples correctly detected among all fake samples in the dataset. The results are presented in Table 3, showing that the highest recalls are achieved by the ViT-S/16 model. With an overall recall of 0.95, a recall of 0.95 for the real class and 0.94 for the fake class, and similarly for precision, ViT-S/16 reaches 0.95 for the fake class and 0.94 for the real class. MobileNetV2 ranks second, with both recall and test accuracy around 0.90, followed by InceptionV3 and VGG-16, whose recall and test accuracy values are approximately 0.86. Finally, ResNet50 shows the lowest performance, with a recall of 0.69 and a test accuracy of 0.63.

Table 3. Test accuracy and recall of the five deepfake face detection models

Model	Class	Test accuracy	Recall
ViT-S/16	Fake	0.95	**0.94**
	Real	0.94	**0.95**
	Overall	**0.95**	**0.95**
VGG-16	Fake	0.85	0.82
	Real	0.83	0.86
	Overall	0.84	0.84
ResNet50	Fake	0.63	0.52
	Real	0.59	0.69
	Overall	0.61	0.61
InceptionV3	Fake	0.86	0.86
	Real	0.86	0.86
	Overall	0.86	0.86
MobileNetV2	Fake	0.90	0.89
	Real	0.89	0.90
	Overall	0.90	0.90

The results clearly confirm the superiority of the ViT-S/16 model, which achieved the best performance overall. It achieves balanced precision and recall across both real and fake classes, demonstrating its ability to correctly identify each category. Its robustness is further highlighted by strong performance on the large and complex 140k-image dataset, indicating effective learning even on realistic and challenging deepfake faces. MobileNetV2 also produced good results, although slightly lower than those of ViT, indicating that it remains a competitive lightweight architecture for deepfake detection despite its reduced complexity.

To further validate the effectiveness of our ViT-S/16 model, we compared its accuracy results with those of other methods cited in the related work. The results are presented in Table 4.

Table 4. Test accuracy comparison between our Vit and the other methods in deep fake face detection

Authors	Accuracy
Jheelan and Pudaruth [14]	0.68
Sharma et al. [12]	0.90
Gong et al. [30]	0.95
Kim [11]	0.93
Shi et al. [9]	0.87
Xiao et al. [31]	0.92
Man and Cho [32]	0.92
Our VIT	**0.95**

The deepfake face detection accuracy of the proposed model surpassed the performance benchmarks established in the extant literature and demonstrated parity with the findings reported by Gong et al. [30]. This outcome suggests that the transformer architecture, by virtue of its inherent attention mechanism, effectively isolates the salient features necessary for robust image authentication.

4.3. Explainability with LIME

This section presents the experiments conducted to interpret the models' decision-making processes. The objective is to identify the image regions that most strongly influence the predictions. To address the black-box nature of the model and enhance its trustworthiness, we employed the LIME explainability method, which provides transparent, human-interpretable insights into how predictions are generated.

For the explainability tests, the five models previously trained on the 140k Real and Fake Faces dataset were used. A test image from the fake class was selected, correctly classified by all models; however, its class is not obvious to a human observer, as illustrated in the test image of Fig. 6.

The image was loaded at its original resolution, and the parameters of Lime Image Explainer were configured as follows: top_labels = 1 to obtain an explanation for the predicted class only, num_samples = 1000, the number of perturbed images generated around the original image, and positive_only = False to display both positively and negatively contributing regions.

The 1000 generated images were then passed to each of our models independently to compute their predictions. LIME subsequently extracted the most influential superpixels, with num_features = 8 to retain the eight most relevant

regions, while keeping the rest of the image visible. The color of each highlighted region indicates whether it contributes positively or negatively to the final decision.

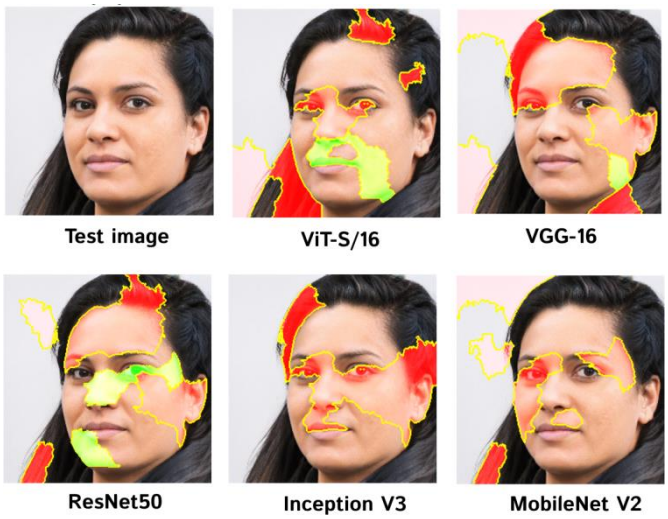

Fig. 6. Explanation with LIME on a test image for ViT-S/16 and the four CNN models

Fig. 6 presents a comparative interpretation of the discriminative regions extracted by several deep neural network architectures using LIME, providing trustworthy insights into their internal decision processes. In all visualizations, red superpixels indicate regions that positively support the model's prediction, whereas green areas correspond to features that counteract or negatively influence the decision. The convolution-based models (VGG-16, ResNet50, InceptionV3, and MobileNetV2) consistently highlight semantically relevant facial components such as the eyes, nose, mouth, and hair contours. VGG-16 and MobileNetV2 show localized and well-defined activated regions, suggesting a dependence on fine-grained texture patterns. ResNet50 and InceptionV3 display broader activation distributions, reflecting a combination of local feature extraction and higher-level contextual cues. In contrast, the ViT-S/16 Vision Transformer produces more spatially diffuse and globally distributed highlighted zones, consistent with its patch-based tokenization and its ability to model long-range dependencies rather than focusing on pixel-level details. Overall, these results reveal heterogeneous interpretability behaviors across architectures: CNNs predominantly rely on local discriminative structures, whereas the Transformer exhibits a more holistic representation. This architectural divergence underscores that different models develop distinct yet complementary internal reasoning when processing the same input image, thereby reinforcing confidence in their predictions through transparent and trustworthy explanations.

6. Conclusion

The rise of deepfakes has significantly undermined public trust in information sources. To address this challenge, we developed a robust and trustworthy deepfake

detection model. In addition to achieving high accuracy, our approach emphasizes transparency: we employ an explainability method that provides clear, instance-specific justifications for each classification, thereby ensuring that the system's decisions are interpretable and reliable. Deepfakes have become more realistic than ever, to the extent that humans can no longer reliably detect them. They pose a serious threat to societal security, making the development of reliable deepfake detection methods an urgent priority. Five deep learning models for deepfake face detection were developed and evaluated, including a Vision Transformer (ViT-S/16) and four CNN architectures, trained on the high-quality 140K Real and Fake Faces dataset. Among these models, the Vision Transformer achieved the highest accuracy (0.95) and a high recall of 0.94 for the fake class, which is particularly important because misclassifying a real face as fake leads to immediate rejection, but failing to detect a fake face represents a far more serious threat, particularly as it can lead to significant breaches of social security. The remaining models ranked as follows: MobileNet-V2 (accuracy 0.90), Inception-V3 (0.86), VGG-16 (0.84), and ResNet-50 (0.61). These results demonstrate that the models can effectively distinguish real faces from highly realistic fake faces, even when manipulations are subtle and not easily detectable by human observers. Beyond detection, model interpretability was emphasized. By applying the perturbation-based LIME method, the facial regions most influencing each model's predictions were identified, yielding clear, visually meaningful explanations. As future work, exploring hybrid architectures that combine CNNs and Transformers could be beneficial. Such models may improve detection accuracy by leveraging both fine-grained local features and global contextual dependencies, enhancing robustness against highly realistic deepfakes.

References

1. Altuncu, E., V. N. L. Franqueira, S. Li. Deepfake: Definitions, Performance Metrics and Standards, Datasets and Benchmarks, and a Meta-Review. – arXiv Preprint arXiv:2208.10913, 2022.
2. Mirsky, Y., W. Lee. The Creation and Detection of Deepfakes. – ACM Computing Surveys, Vol. **54**, 2021, No 1, pp. 1-41. DOI: 10.1145/3425780.
3. Perov, I., D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, Dpfks, C. S. Facenheim, R. P. Luis, J. Jiang, S. Zhang, P. Wu, B. Zhou, W. Zhang. DeepFaceLab: Integrated, Flexible, and Extensible Face-Swapping Framework. – arXiv Preprint arXiv:2005.05535, 2020.
4. Kowalski, M. FaceSwap. – GitHub Repository, 2024. **https://github.com/MarekKowalski/FaceSwap/**
5. Contributors, FaceFusion. – In: GitHub Repository, 2024. **https://github.com/facefusion/facefusion**
6. Rombach, R., A. Blattmann, D. Lorenz, P. Esser, B. Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. – In: Proc of IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR'22), New Orleans, USA, 2022, pp. 10684-10695.
7. Li, Y., X. Yang, P. Sun, H. Qi, S. Lyu. Celeb-DF: A Large-Scale Challenging Dataset for Deepfake Forensics. – In: Proc of IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR'20), 2020, pp. 3207-3216.

8. Rossler, A., D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner. FaceForensics++: Learning to Detect Manipulated Facial Images. – In: Proc of IEEE/CVF Int. Conf. on Computer Vision (ICCV'19), 2019, pp. 1-11.
9. Shi, L., J. Zhang, Z. Ji, J. Bai, S. Shan. Real Face Foundation Representation Learning for Generalized Deepfake Detection. – Pattern Recognition, Vol. **161**, 2025, 111299. DOI: 10.1016/j.patcog.2024.111299.
10. Liu, D., Z. Dang, C. Peng, Y. Zheng, S. Li, N. Wang, X. Gao. Fedforgery: Generalized Face Forgery Detection with Residual Federated Learning. – IEEE Transactions on Information Forensics and Security, Vol. **18**, 2023, pp. 4272-4284. DOI: 10.1109/TIFS.2023.3293951.
11. Kim, H. Novel Deep Learning-Based Facial Forgery Detection for Effective Biometric Recognition. – Applied Sciences, Vol. **15**, 2025, No 7, 3613. DOI: 10.3390/app15073613.
12. Sharma, J., S. Sharma, V. Kumar, H. S. Hussein, H. Alshazly. Deepfakes Classification of Faces Using Convolutional Neural Networks. – Traitement du Signal, Vol. **39**, 2022, No 3. DOI: 10.18280/ts.390330.
13. Naeem, S., R. Al-Sharawi, M. R. Khan, U. Tariq, A. Dhall, H. Al-Nashash. Real, Fake, and Synthetic Faces – Does the Coin Have Three Sides? – In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG), Istanbul, Turkey, 2024. DOI: 10.1109/FG59268.2024.10581973.
14. Jheelan, J., S. Pudaruth. Using Deep Learning to Identify Deepfakes Created Using Generative Adversarial Networks. – Computers, Vol. **14**, 2025, No 2. DOI: 10.3390/computers14020060.
15. Yang, W., Y. Wei, H. Wei, Y. Chen, G. Huang, X. Li, R. Li, N. Yao, X. Wang, X. Gu, M. B. Amin, B. Kang. Survey on Explainable AI: From Approaches, Limitations, and Applications Aspects. – Human-Centric Intelligent Systems, Vol. **3**, 2023, pp. 161-188. DOI: 10.1007/s44230-023-00038-y.
16. Salih, A., Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. Petersen, G. Menegaz, K. Lekadir. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. – Advanced Intelligent Systems, 2023, arXiv:2305.02012.
17. Mersha, M., K. Lam, J. Wood, A. AlShami, J. Kalita. Explainable Artificial Intelligence: A Survey of Needs, Techniques, Applications, and Future Direction. – Neurocomputing,
Vol. **599**, 2024, 128111. DOI: 10.1016/j. neucom.2024.128111.
18. Tariq, S., S. S. Woo, P. Singh, I. Irmalasari, S. Gupta, D. Gupta. From Prediction to Explanation: Multimodal, Explainable, and Interactive Deepfake Detection Framework for Non-Expert Users. – In: Proc of 33rd ACM Int. Conf. on Multimedia (MM'25), 2025, pp. 11716-11725. DOI: 10.1145/3746027.3755786.
19. Mansoor, N., A. I. Iliev. Explainable AI for Deepfake Detection. – Applied Sciences, Vol. **15**, 2025, No 2, Article 725. DOI: 10.3390/app15020725.
20. Tsigos, K., E. Apostolidis, S. Baxevanakis, S. Papadopoulos, V. Mezaris. Towards Quantitative Evaluation of Explainable AI Methods for Deepfake Detection. – In: ACM, New York, USA, 2024, pp. 37-45. DOI: 10.1145/3643491.3660292.
21. Ribeiro, M. T., S. Singh, C. Guestrin. Why Should I Trust You? Explaining the Predictions of Any Classifier. – In: Proc of ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Francisco, USA, 2016, pp. 1135-1144. DOI: 10.1145/2939672.2939778.
22. Schallner, L., J. Rabold, O. Scholz, U. Schmid. Effect of Superpixel Aggregation on Explanations in LIME – A Case Study with Biological Data. – arXiv Preprint arXiv:1910.07856, 2019.
23. Hase, P., M. Bansal. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? – In: Proc of Annual Meeting of the Association for Computational Linguistics (ACL'20), 2020, pp. 5540-5552. DOI: 10.18653/v1/2020. acl-main.491.
24. Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. – In: Proc of Conf. on Learning Representations (ICLR'21), 2021.

25. S i m o n y a n, K., A. Z i s s e r m a n. Very Deep Convolutional Networks for Large-Scale Image Recognition. – arXiv preprint arXiv:1409.1556, 2014.
26. H e, K., X. Z h a n g, S. R e n, J. S u n. Deep Residual Learning for Image Recognition. – In: Proc of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'16), 2016, pp. 770-778.
27. S z e g e d y, C., V. V a n h o u c k e, S. I o f f e, J. S h l e n s, Z. W o j n a. Rethinking the Inception Architecture for Computer Vision. – In: Proc of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'16), 2016, pp. 2818-2826.
28. S a n d l e r, M., A. H o w a r d, M. Z h u, A. Z h m o g i n o v, L.-C. C h e n. MobileNetV2: Inverted Residuals and Linear Bottlenecks. – In: Proc of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'18), 2018, pp. 4510-4520.
29. D e n g, J., W. D o n g, R. S o c h e r, L.-J. L i, K. L i, L. F e i-F e i. ImageNet: A Large-Scale Hierarchical Image Database. – In: Proc of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'09), 2009, pp. 248-255.
30. G o n g, R., R. H e, D. Z h a n g, A. K. S a n g a i a h, M. J. F. A l e n a z i. Robust Face Forgery Detection Integrating Local and Global Texture Information. – EURASIP Journal on Information Security, Vol. **2025**, 2025, 3. DOI: 10.1186/s13635-025-00189-4.
31. X i a o, Y., Y. Z h o u, P. C h e n g, L. N i, X. W u, T. Z h e n g. An Attention-Based Framework for Detecting Face Forgeries: Integrating Efficient-ViT and Wavelet Transform. – Mathematics, Vol. **13**, 2025, No 16, Article 2576. DOI: 10.3390/math13162576.
32. M a n, Q., Y.-I. C h o. Exposing Face Manipulation Based on GAN-Transformer and Fake Frequency Noise Traces. – Sensors, Vol. **25**, 2025, No 5, 1435. DOI: 10.3390/s25051435.
33. A b i r a m i, P. Enhanced Fake Image Detection in Social Media Using Vision Transformer. – Int. J. for Research in Applied Science and Engineering Technology, Vol. **13**, 2025, pp. 4570-4574. DOI: 10.22214/ijraset.2025.69302.
34. C i r i l l o, L., A. G e r v a s i o, I. A m e r i n i. Explainability-Driven Adversarial Robustness Assessment for Generalized Deepfake Detectors. – EURASIP Journal on Information Security, Vol. **2025**, 2025, 23. DOI: 10.1186/s13635-025-00211-9.
35. Xhlulu. 140k Real and Fake Faces Dataset. – Kaggle Dataset, 2020.
36. NVIDIA Research. Flickr-Faces-HQ (FFHQ) Dataset. – GitHub Repository, 2019.
37. K a r r a s, T., S. L a i n e, T. A i l a. A Style-Based Generator Architecture for Generative Adversarial Networks. – In: Proc of IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR'19), 2019. DOI: 10.1109/CVPR.2019.00453.
38. T u n g u z, B. 1 Million Fake Faces Generated by StyleGAN. – In: Kaggle Dataset, 2020.
39. V a s w a n i, A., N. S h a z e e r, N. P a r m a r, J. U s z k o r e i t, L. J o n e s, A. N. G o m e z, L. K a i s e r, I. P o l o s u k h i n. Attention Is All You Need. – In: Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998-6008.

Received: 16.12.2025, Revised version: 28.02.2026, Accepted: 05.03.2026