Deep Fake Recognition using Vision Transformer Architecture

Zeerak Baig[1[0000-1111-2222-3333]]

[1] University of Johannesburg, Johannesburg JHB 524, RSA
217026768@student.uj.ac.za

Abstract. A photo-realistic image may be produced from low-dimension random noise using generative adversarial networks (GANs). Social media networks may employ a synthetic (false) picture with unsuitable material, which can have adverse effects. A robust and accurate picture forgery detector is required to detect false photos properly. However, because the bogus pictures produced by the GAN-based generator are created and altered from the original image, standard image forgery detectors cannot identify them. Deepfake digital photographs seriously undermine the objectivity of news reporting, legal forensics, and social security. To detect deep fakes, this project proposes a transformer-based deep learning model. Vision transformer employs a Transformer-like architecture for image classification. A sequence of vectors is created by dividing a picture into fixed-size patches, linearly embedding each one, adding position embeddings, and then feeding the assembled vectors to a conventional Transformer encoder. The study did not use a pre-trained transformer model. Instead, the paper looks to determine the performance of a primary vision transformer without any convolutional layers. For a comparative analysis, the study also makes use of an Efficient Net B07 architecture for classification purposes.

Keywords: Vision Transformer, Generative Adversarial Networks.

1 Introduction

Variational autoencoders and generative adversarial networks (GANs) are two recent examples of deep learning-based generative models that are frequently used to synthesis the photorealistic portion or entire content of an image or a video [1]. Additionally, more recent GAN modifications, such as progressive growth of GANs (PGGAN) and BigGAN, have been used to create very photorealistic images and videos that are hard for humans to detect as fakes in a short amount of time [2]. Most generative programs execute picture translation duties, which can result in significant issues if a bogus image is wrongly shared on social networking platforms. In a pornographic film, for example, the false facial picture may be synthesized using the cycleGAN [3]. Additionally, the GANs can produce speech videos, including any well-known politician's synthetic face features, which pose serious issues for society, politics, and business endeavors. Consequently, there is a desperate need for a reliable deep fake face picture identification method.

In recent years, deepfakes have multiplied in frequency and sophistication, to the point where forensic experts, decision-makers, and the general public are concerned about their potential to propagate misinformation. The development and detection of such forgeries have recently gained attention as a research area, which has resulted in a large increase in papers on deepfake creation, detection techniques, and datasets including the most recent deepfake generation techniques. Convolutional neural networks serve as the foundation of deep learning techniques that are particularly effective at spotting and avoiding deep learning fakes [4].

This project presents a transformer-based architecture for Deep Fake recognition. The Vision Transformer, often known as ViT, is an image categorization model that uses a Transformer-like design across selected areas of the picture. A conventional Transformer encoder is supplied with the resultant sequence of vectors after a picture has been divided into fixed-size patches, each of which has been linearly embedded and position embeddings added. The conventional method of adding an extra learnable "classification token" to the sequence is employed to conduct classification [5]. The study makes use of a traditional vision transformer to measure accuracy and efficiency in categorizing deep fake images and normal images. For a comparative study, the study also uses an Efficient Net B07 as a base model for our classifier. When compared to all existing CNN architectures, Efficient net B07 has achieved both higher accuracy and better efficiency [6].

2 Literature Review

As mentioned above, autoencoders and generative adversarial networks are two recent examples of deep learning-based generative models used to synthesize all or part of the photorealistic content of an image or a video. Therefore, before diving into our proposed methodology and implementation, it is essential to understand fundamental technologies that are currently applied to generate deep fake images.

2.1 Generative Adversarial Networks

Deep representations may be learned using generative adversarial networks (GANs) without needing much annotated training data. They do this using a competitive approach utilizing two networks to derive backpropagation signals. Several applications, including image synthesis, semantic image editing, style transfer, picture super resolution, and classification, can use the representations that GANs can learn [7].

To create samples from the learnt distribution, generative models learn to capture the statistical distribution of training data. We are interested in leveraging the representations that such models learn for tasks like classification and image retrieval in addition to synthesizing new data samples, which may be utilized for downstream tasks like semantic picture editing, data augmentation, and style transfer [7].

Numerous innovative new applications have been developed because of GANs. For instance, it has been a critical component of algorithms for creating photorealistic pictures from human-editable semantic representations, such as segmentation masks or drawings. Numerous image-to-image translation techniques that convert an image from one domain to a matching picture in a different field have also been developed due to GANs. These techniques have many various applications, from picture manipulation to domain adaptation [8].

2.2 Autoencoders

An appropriate method for streamlining the feature engineering process in machine learning research is to employ autoencoders, which are neural networks that can automatically learn meaningful features and representations from the data. Additionally, dimensionality reduction, data denoising, generative modeling, and even pretraining deep learning neural networks may all be done with autoencoders [9].

2.2.1 Variational Autoencoders

Deep latent space generative models known as "variational auto-encoders" (VAEs) have found tremendous success in a variety of applications, including language modeling, protein design, mutation prediction, and image synthesis. The core concept behind VAEs is to learn the data distribution in a way that enables the generation of new, useful data from the encoded distribution. This idea has generated a lot of research and different VAE designs in recent years, spawning a new subject called unsupervised representation learning [10].

2.2.2 Adversarial Autoencoders

Variational autoencoder is a probabilistic autoencoder that does variational inference by comparing the aggregated posterior of the autoencoder's hidden code vector with any prior distribution. It does this using the newly developed generative adversarial networks (GAN). Generating from any region of the prior space produces relevant samples since the aggregated posterior and the prior are matched. To map the imposed prior to the data distribution, the adversarial autoencoder's decoder builds a deep generative model [11].

3 Problem Background

Two categories of forensics systems—active and passive schemes—are frequently utilized in classic picture forgery detection methods. In the active methods, the source picture is integrated without visual artefacts with an externally added signal (a watermark). The target picture is subjected to the watermark extraction method to recover the watermark from ascertaining if the image has been altered [12]. The target picture has tampered areas that may be found using the retrieved watermark image. On the other hand, passive picture forgery detectors rely on the source image's statistical data, which has a high degree of consistency across distinct photos. Therefore, the bogus portions in the image may be found using the inherent statistical information [13]. The GAN-generated fake pictures are created from the low-dimensional random vector. Therefore, the passive image forgery detectors cannot be utilized to detect them. In particular, the bogus images produced by the GANs remain unaltered from their source photos.

We may utilize a deep neural network to identify bogus pictures produced by the GANs, as they have been widely employed in numerous recognition tasks. Recently, supervised learning-based deep learning approaches for fraudulent picture identification have been researched. In other words, identifying fraudulent images has been approached as a binary classification issue (i.e., fake or real image).

3.1 Existing Works

Many approaches using convolutional neural networks have been proposed to develop deep fake image detectors. Jonathan et al [14] analyzed the performance of a number of fake image detectors based on very deep neural networks. Their study, which used a dataset of 36302 photos, demonstrates that both conventional and deep learning detectors can reach detection accuracies up to 95%, but only the latter maintain a high accuracy, up to 89%, on compressed data. Mo, H. et al. suggested a Convolutional Neural Network (CNN) based approach to recognize artificially created faces, and they provided experimental data to support their claims that the system can provide acceptable outcomes with an average accuracy of over 99.4% [15]. A hybrid ensemble learning approach-based modified face identification method was developed in [16]. This experiment was able to achieve an accuracy of 84.7%, but the pre-trained VG-Face achieved an accuracy of 89%.

4 Experiment Setup

The goal of this study is to identify deep fake images using a variety of facial images. Our first approach uses a transformer-based architecture for deep fake recognition. The Vision Transformer, sometimes referred to as ViT, is a methodology for categorizing images that applies a Transformer-like design to certain portions of the image. After an image has been partitioned into fixed-size patches, each of which has been linearly embedded and position embeddings added, the resultant sequence of vectors is provided to a traditional Transformer encoder. To do classification, the traditional approach of including an additional, learnable "classification token" in the sequence is used.

Our second approach is a variation of our initial pipeline as shown in Figure 2, where an Efficient net B07 architecture is used as a classifier instead of the vision transformer for a comparative study.

4.1 Datasets

The study made use of a single dataset for training and testing purposes. We used 140k Real and Fake faces dataset which is publicly available on Kaggle. This dataset includes all 70k REAL faces from the Flickr dataset gathered by Nvidia as well as a sample of 70k FAKE faces drawn from Bojan's 1 Million FAKE faces (created using StyleGAN). Dataset contains images divided into training, validation, and test folders, with each image resized to 256px [17].

4.2 Evaluation Metrics

The study will report several metrics to measure the accuracy of the classifiers. Precision and recall are helpful metrics of prediction success when the classes are severely unbalanced. Precision measures result from relevancy in information retrieval, whereas recall measures how many relevant results are returned [18]. The precision-recall curve depicts the tradeoff between precision and recall for various thresholds. High accuracy suggests a low false-positive rate, and high recall indicates a low false-negative rate, a significant area under the curve means good recall and precision. High scores imply that the classifier delivers accurate results and that most positive outcomes are positive.

5 Methodology

Pre-processing, training, and validation make up the method's three key components, as shown in Figure 1. After choosing the dataset, we take the images in the training set to train the vision transformer model. During the image preprocessing stage, the image data is normalized, and sizes are adjusted for homogenization purposes. This research aimed to detect deep fakes by training a conventional vision transformer without using any pre-trained models and compare it against an Efficient net B07 CNN architecture to provide a comparative study.

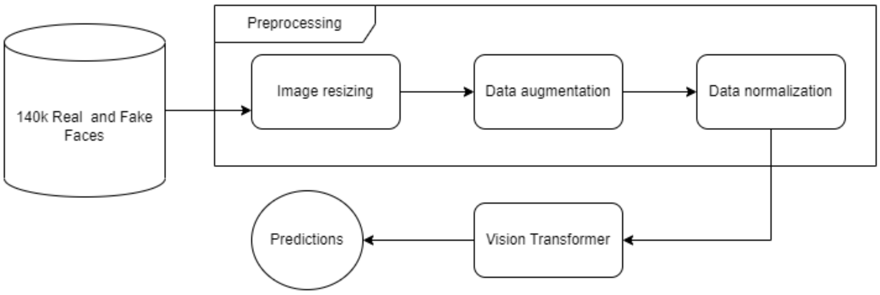

Figure 1 Proposed Method

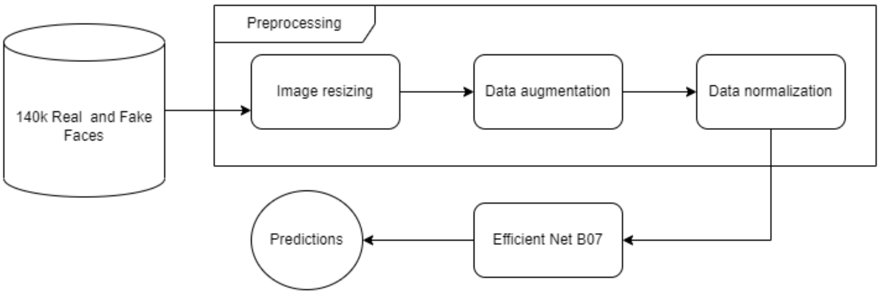

Figure 2 Pipeline variation with Efficient Net B07 architecture

5.1 Data Preprocessing

To carry out training with the 140k Real and Fake face dataset, it was first necessary to balance the number of images of the two classes (real and fake image). The dataset used had a total of 140000 images. Our experiment extracted a total of 10000 images from the original dataset. The extracted images were equally divided into the respective classes in such a way that we had 4000 training images for real face image and 4000 images for deep fake images. The test set consist of 2000 images equally divided among the two classes, like the training set as shown in the table below.

Table 1 Real Face and Fake Face data division

Data Set	Real Face Images	Fake Face Images
Training Set	4000	4000
Test Set	1000	1000

The images containing real and fake faces go through a normalization process, resizing by 224x224 pixels, and data augmentation process before being saved in a dataset that will be used by both the vision transformer, and the Efficient Net B07 Architecture.

5.2 Vision Transformer Architecture

Vision Transformer is an architecture that is built on the original Transformer. The original Transformer is a prevalent architecture because it performs well in NLP applications like machine translation. Without a recurrent network, the Transformer's architecture of encoders and decoders allows it to process sequential input in parallel. The performance of Transformer models has been dramatically influenced by the self-attention mechanism, which is hypothesized to capture long-range connections between the sequence's parts.

The suggested Vision Transformer is an effort to expand the application of the conventional Transformer to picture categorization. With no integration of a data-specific architecture, the key objective is to generalize them to other modalities outside the text. The encoder module of the Transformer is used explicitly by Vision Transformer to carry out classification by mapping a series of picture patches to the semantic label. The attention mechanism used by the Vision Transformer allows it to attend across various parts of the picture and integrate information throughout the whole image, in contrast to standard CNN designs that often use filters with a small receptive field [19].

Figure 3 displays the whole end-to-end architecture of the model. A final head classifier, an encoder, and an embedding layer generally make up this system. In the first phase, non-overlapping patches are created from a picture X from the training set (we omit the image index I for simplicity). The Transformer sees each patch as a distinct token. So, with an image of size X, where c is the number of channels, h is the height, and w is the width, we extract patches from each dimension c x p x p. This creates a series of patches with lengths (x1, x2,..., xn) where n=hw/p2. A 16 by 16 or 32 by 32 patch size is typically selected, with a lower patch size resulting in a longer sequence and vice versa [20].

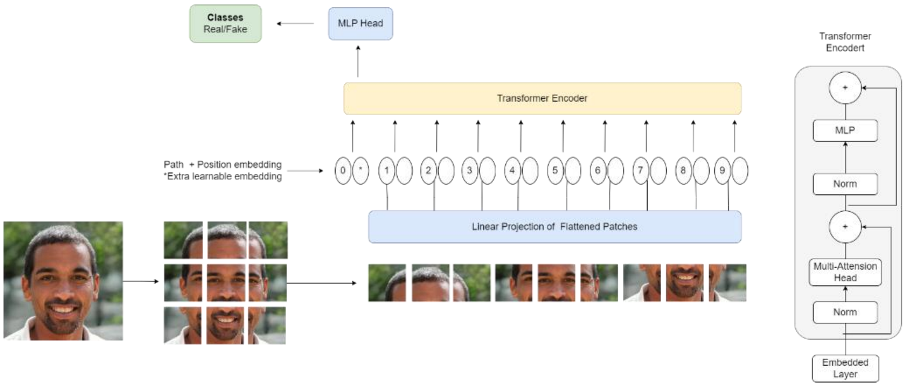

Figure 3 Vision Transformer Architecture

5.3 Efficient Net B07 Architecture

Before feeding our convolutional neural network with train and test samples, image samples must be preprocessed. The images are firstly resized to 224 by 224, and we convert them from greyscale to RGB space by repeating the intensity values across all three channels. The process then reads the image in RGB format and applies pixel normalization.

Once the image pre-processing has been completed, our convolutional neural network is ready to accept the input data. Before feeding data to the CNN, the training data goes through data augmentation stage, which increases the diversity of dataset without the need to collect more data.

The proposed study makes use of Keras's sequential model. Efficient Net B07 forms the first layer of our model. EfficientNet is a convolutional neural network design and scaling technique using a compound coefficient to scale all depth, breadth, and resolution parameters consistently. The EfficientNet scaling approach evenly increases network width, depth, and resolution with a set of preset scaling coefficients, in contrast to standard practice, which scales these elements arbitrarily [20]. Our architecture is then followed by a two-dimensional Global Average Pooling layer. Global average pooling layer is intended to take the role of fully connected layers in a conventional CNN. We average each feature map, and resultant vector is sent straight to a softmax layer rather than constructing fully linked layers on top of the feature maps. We then add a dropout layer with a 20% dropout rate. A specific number of neurons in the network are ignored or dropped out randomly using the technique. Finally, we have fully connected layer with a sigmoid activation function for a bonary classification task. Efficient Net B07 architecture can be seen in Figure 4 below.

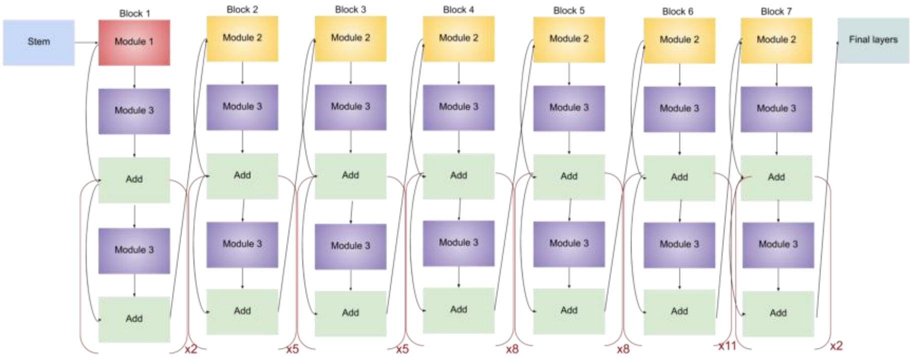

Figure 4 Efficient Net B07 Architecture

6 Results

6.1 Vision Transformer Architecture Accuracy

The summary of accuracy scores is displayed in table 2 below. According to our results, the visions transformer architecture achieved an overall accuracy of 80.7 %. The vision transformer classifier achieved a precision, recall and f-1 score of 80.5 %. Figure 5 below demonstrates the training and validation curves for our classifier.

Table 2 Classification report of vision transformer-based architecture

	Precision	Recall	F1-Score
Real Face	80%	83%	81%
Fake face	82%	79%	80%

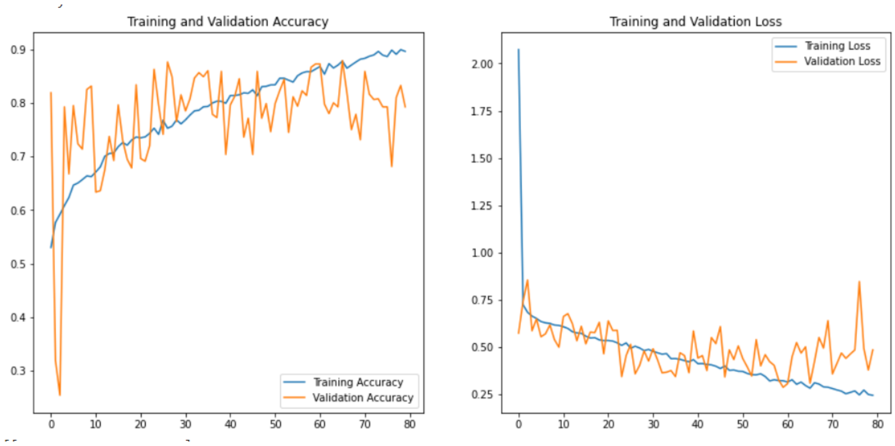

Figure 5 Training and validation learning curves for Vision Transformer Architecture

6.2 Efficient Net B07 Architecture Accuracy

Variations in our initial pipeline included a different architecture for classification purposes. We feed the images to a convolutional neural network based on Efficient Net B07 architecture using the same image pre-processing and normalisation techniques. Table 3 below shows the overall accuracy scores of our CNN architecture. Efficient Net classifier was able to achieve an overall accuracy of 81.8 %. This entails that the Efficient net architecture outperformed the vision transformer architecture by 1.1 % in terms of accuracy. Efficient Net architecture achieved an overall precision score of 83%, a recall of 82%, and an f1 score of 81.5 %. Figure 6 below shows us the training and validation curves for the Efficient Net B07 architecture.

Table 3 Classification report of Efficient Net Architecture

	Precision	Recall	F1-Score
Real Face	89%	73%	80%
Fake face	77%	91%	83%

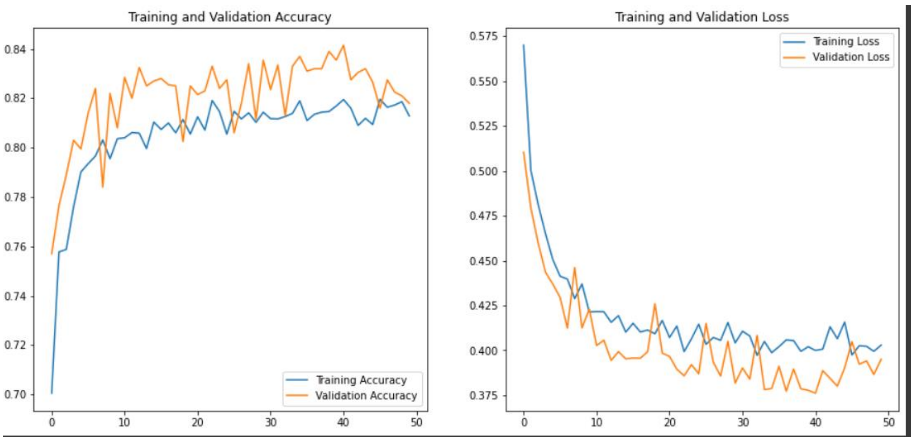

Figure 6 Training and validation curves for Efficient Net B07 Architecture

7 Conclusion

Two different classifiers were presented for deep fake recognition for a comparative study. Our initial approach made use of a vision transformer as the classifier. In contrast, a variation of the same pipeline used an Efficient net B07 architecture for deep fake image recognition. The study presented two deep learning approaches with various advantages and a decent accuracy score. The vision transformer model achieved an accuracy of 80.7 %, whereas the CNN-based architecture acquired an accuracy of 81,8%. Detecting forgery and tampering with images and videos is gaining importance in digital forensics. Therefore, methods for developing systems to detect deep fake photos can also serve as the basis for forming standardized tools for detecting deep fake images, classification, and forgery detection. The results obtained in this study can be considered adequate, however, with more data and hyperparameter tuning, we can achieve better results.

References

1. Hsu, C.-C., Zhuang, Y.-X., Lee, C.-Y.: Deep fake image detection based on pairwise learning. Applied Sciences. 10, 370 (2020).
2. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative Adversarial Networks: An overview. IEEE Signal Processing Magazine. 35, 53–65 (2018).
3. Simoes, G.S., Wehrmann, J., Barros, R.C.: Attention-based adversarial training for seamless nudity censorship. 2019 International Joint Conference on Neural Networks (IJCNN). (2019).

4. Silva, S.H., Bethany, M., Votto, A.M., Scarff, I.H., Beebe, N., Najafirad, P.: Deepfake Forensics Analysis: An explainable hierarchical ensemble of weakly supervised models. Forensic Science International: Synergy. 4, 100217 (2022).
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. (2020). https://doi.org/10.48550/ARXIV.2010.11929.
6. Tang, M.: https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html, (2019).
7. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative Adversarial Networks: An overview. IEEE Signal Processing Magazine. 35, 53–65 (2018).
8. Liu, M.-Y., Huang, X., Yu, J., Wang, T.-C., Mallya, A.: Generative adversarial networks for image and video synthesis: Algorithms and applications. Proceedings of the IEEE. 109, 839–862 (2021).
9. Lopez Pinaya, W.H., Vieira, S., Garcia-Dias, R., Mechelli, A.: Autoencoders. Machine Learning. 193–208 (2020).
10. Wei, R., Garcia, C., El-Sayed, A., Peterson, V., Mahmood, A.: Variations in variational autoencoders - a comparative evaluation. IEEE Access. 8, 153651–153670 (2020).
11. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial Autoencoders. (2015).
12. Hsu, C.-C.: Image authentication with tampering localization based on watermark embedding in Wavelet domain. Optical Engineering. 48, 057002 (2009).
13. Sitara, K., Mehtre, B.M.: Digital Video Tampering Detection: An overview of passive techniques. Digital Investigation. 18, 8–22 (2016).
14. Marra, F., Gragnaniello, D., Cozzolino, D., Verdoliva, L.: Detection of gan-generated fake images over social networks. 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). (2018).
15. Mo, H., Chen, B., Luo, W.: Fake faces identification via Convolutional Neural Network. Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. (2018).
16. Dang, L.M., Hassan, S.I., Im, S., Moon, H.: Face image manipulation detection based on a convolutional neural network. Expert Systems with Applications. 129, 156–168 (2019).
17. Sharma, J., Sharma, S., Kumar, V., Hussein, H.S., Alshazly, H.: Deepfakes classification of faces using Convolutional Neural Networks. Traitement du Signal. 39, 1027–1037 (2022).
18. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. Proceedings of the 23rd international conference on Machine learning - ICML '06. (2006).
19. Bazi, Y., Bashmal, L., Rahhal, M.M., Dayil, R.A., Ajlan, N.A.: Vision Transformers for Remote Sensing Image Classification. Remote Sensing. 13, 516 (2021).
20. Koonce, B.: EfficientNet. Convolutional Neural Networks with Swift for Tensorflow. 109–123 (2021).