Advanced Deep Fake Image Detection with Vision Transformer and EfficientNet B07

A project report submitted

for partial fulfillment of the requirements for the award of the degree of Master of Science

by

Abhik De

Registration number: 213001818010026 of 2021-22

M.Sc. in Applied Statistics and Analytics

Under the supervision of

Anwesha Sengupta & Prasanta Narayan Dutta

Department of Applied Statistics

MAKAUT, WB

May, 2023

MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY, WEST BENGAL

Department of Applied Statistics

MAKAUT, WB

CERTIFICATE

This is to certify that the dissertation report entitled '***Advanced Deep Fake Image Detection with Vision Transformer and EfficientNet B07***', submitted by Abhik De (Reg. No: 213001818010026 of 2021-22, Roll No: 30018021026) to MAKAUT, WB, is a record of project work carried out by him under my supervision and guidance, and is worthy of consideration for the award of the degree of Master of Science in Applied Statistics and Analytics of the University.

Prasanta Narayan Dutta,
Co-Supervisor,
Dept. of Applied Statistics

Anwesha Sengupta,
Supervisor,
Dept. of Applied Statistics

Anwesha Sengupta,
Assistant Professor & Head of the Department (HOD),
Dept. of Applied Statistics

MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY, WEST BENGAL

Department of Applied Statistics

MAKAUT, WB

Declaration

I declare that, this project report has been composed by me and no part of this project report has formed the basis for the award of any Degree/Diploma or any other similar title to me.

Date: _____

Student's Name: **Abhik De**

Reg. No: 213001818010026
of 2021-2022,
Department of Applied Statistics,
MAKAUT, WB

Acknowledgment

I would like to express my profound and deep sense of gratitude to my guides Prof. Anwesha Sengupta & Prof. P.N. Dutta for their unending help, guidance, and suggestions without which this thesis would not have been a reality. I owe great indebtedness for their untiring efforts during the period of my project work. I am genuinely indebted to Prof. Sukhendu Samajdar, Director, School of Applied Science, Maulana Abul Kalam Azad University of Technology, Kalyani, West Bengal, India for encouraging this research work. I would like to thank all the faculty members. Without their inspiration, it was not possible for me to complete this project. Finally, my earnest thanks go to my friends who were always beside me when I needed them. I would always be grateful to them.

Abhik De
Roll No: 30018021026
Registration No: 213001818010026
of 2021-2022

Contents

Abstract:

This project focuses on the detection of deep fake images, which pose a significant threat to the objectivity of news reporting, legal forensics, and social security. Traditional image forgery detectors struggle to identify these manipulated images generated by generative adversarial networks (GANs). To address this issue, a transformer-based deep learning model is proposed. The vision transformer architecture, typically used for image classification, is employed to process images by dividing them into patches, embedding them, and utilizing a conventional transformer encoder. The study compares the performance of the vision transformer with an EfficientNet B07 architecture for classification tasks. By exploring these models, this project aims to enhance the detection accuracy of deep fake images and contribute to the field of digital forensics.

Keywords: Vision Transformer, Efficient Net B07, Generative Adversarial Networks, Deep fake, Convolutional Neural Network (CNN)

1. Introduction:

In recent years, generative models like variational autoencoders (VAEs) and generative adversarial networks (GANs) have gained significant attention in the field of deep learning. These models excel at synthesizing realistic images and videos [1]. Notably, advancements in GAN techniques, such as progressive growth of GANs (PGGAN) and BigGAN, have further pushed the boundaries of generating highly photorealistic visuals that are challenging for humans to distinguish as fake within a short timeframe [2].

However, the widespread use of generative models raises concerns when it comes to the manipulation and distribution of synthetic images on social media platforms. For instance, there are instances where bogus images are shared, such as in the context of pornographic films where fake facial pictures can be synthesized using techniques like cycleGAN [3]. Moreover, GANs have the capability to generate manipulated videos, including synthetic face features of well-known politicians, which pose serious ethical, social, political, and business challenges.

As a result, there is a pressing need for a reliable method to identify deep fake face pictures. This necessitates the development of advanced techniques and algorithms capable of accurately detecting manipulated images generated by generative models like VAEs and GANs.

In recent years, the proliferation of deepfakes has raised concerns among forensic experts, decision-makers, and the general public due to the potential spread of misinformation. Consequently, the research focus has shifted towards the development and detection of deepfake forgeries, leading to an influx of papers on deepfake creation, detection techniques, and datasets that encompass the latest generation methods [4]. Convolutional neural networks (CNNs) have emerged as a fundamental tool in deep learning for effectively identifying and mitigating deepfake instances.

This project proposes the utilization of a transformer-based architecture for deepfake recognition. The Vision Transformer (ViT) is employed, which adopts a Transformer-like design over selected regions of an image. By dividing the image into fixed-size patches and linearly embedding each patch while incorporating position

embeddings, a sequence of vectors is obtained. This sequence is then fed into a conventional Transformer encoder. To conduct classification, the traditional approach of including an additional learnable "classification token" in the sequence is employed [5]. The performance and efficiency of the proposed transformer-based architecture are evaluated for deepfake and normal image categorization, with a traditional vision transformer used as the primary model. For comparative analysis, an EfficientNet B07 model is also employed as a baseline classifier. Notably, EfficientNet B07 has demonstrated superior accuracy and efficiency compared to existing CNN architectures [6].

2. Literature review:

Before delving into our proposed methodology and implementation, it is crucial to gain a fundamental understanding of the key technologies employed in generating deep fake images. In recent years, two notable examples of deep learning-based generative models have emerged: autoencoders and generative adversarial networks (GANs). These models are utilized to synthesize realistic content, either in its entirety or specific parts, within an image or a video. By comprehending these foundational technologies, we can better appreciate the context and significance of our own approach.

Generative adversarial networks (GANs) have revolutionized the field of deep learning by enabling the learning of deep representations without requiring extensive annotated training data. This is achieved through a competitive approach involving two networks that generate backpropagation signals. GANs have found applications in various domains, such as image synthesis, semantic image editing, style transfer, picture super-resolution, and classification, by leveraging the representations they learn [7].

The primary goal of generative models is to capture the statistical distribution of the training data, allowing them to generate samples from this learned distribution. In addition to synthesizing new data samples, these learned representations can be valuable for tasks like classification and image retrieval. Moreover, they can be utilized for downstream tasks such as semantic picture editing, data augmentation, and style transfer [7].

The impact of GANs on the field of computer vision has been far-reaching, leading to the development of numerous innovative applications. For example, GANs have played a crucial role in algorithms that generate photorealistic images from human-editable

semantic representations like segmentation masks or drawings. Additionally, GANs have facilitated the development of image-to-image translation techniques, which enable the conversion of an image from one domain to a corresponding image in a different domain. These techniques have diverse applications ranging from image manipulation to domain adaptation [8].

Autoencoders are neural networks that offer a useful approach for simplifying the feature engineering process in machine learning research. They have the capability to automatically learn meaningful features and representations directly from the data. In addition to feature learning, autoencoders can also be employed for various tasks such as dimensionality reduction, data denoising, generative modeling, and even pretraining deep learning neural networks [9]. By leveraging autoencoders, researchers can streamline the process of extracting valuable features from the data, leading to more efficient and effective machine learning models.

Variational auto-encoders (VAEs) are deep generative models that have gained significant traction in various domains such as language modeling, protein design, mutation prediction, and image synthesis. These models operate by learning the underlying distribution of the data, enabling the generation of new and valuable data samples from the encoded distribution. The concept of VAEs has sparked extensive research and the development of different VAE architectures, leading to the emergence of unsupervised representation learning as a distinct field [10]. The success of VAEs lies in their ability to capture the intricate patterns and structure of the data, providing a powerful tool for generating novel and meaningful outputs in diverse applications.

Variational autoencoder (VAE) is a type of autoencoder that incorporates probabilistic modeling and variational inference techniques. It leverages generative adversarial networks (GANs) to compare the aggregated posterior of the autoencoder's hidden code vector with a prior distribution. This comparison ensures that samples generated from any region of the prior space are meaningful, as the aggregated posterior and the prior are aligned. The decoder component of the adversarial autoencoder builds a deep generative model to map the imposed prior to the data distribution [11]. By combining these techniques, VAEs enable the generation of diverse and high-quality samples while maintaining a probabilistic framework for inference.

3. Identification of Research Gap:

In the field of picture forgery detection, two main types of forensic systems are commonly used: active and passive schemes [12]. Active methods involve integrating a source image with an externally added signal, such as a watermark, that is visually imperceptible. The target image is then subjected to a watermark extraction process to recover the watermark and determine if the image has been altered. The presence of tampered areas can be identified by comparing the retrieved watermark image with the target image. On the other hand, passive picture forgery detectors rely on the statistical properties of the source image, which exhibit a high degree of consistency across different photos. By analyzing the inherent statistical information, these detectors can identify manipulated regions within an image [13].

However, when it comes to detecting fake images generated by generative adversarial networks (GANs), both active and passive methods face limitations. GAN-generated fake images are created from low-dimensional random vectors and do not exhibit statistical inconsistencies compared to their source images. As a result, passive image forgery detectors that rely on statistical properties are ineffective in detecting GAN-generated fake images. The images produced by GANs appear visually similar to real images and cannot be distinguished using standard image forgery detection techniques.

To address this challenge, deep neural networks have been proposed as a potential solution for detecting GAN-generated fake images. Deep learning approaches, particularly supervised learning-based methods, have been extensively studied for fraudulent picture identification. These approaches treat the problem of identifying fake images as a binary classification task, distinguishing between fake and real images. Deep neural networks have shown promising results in various recognition tasks and can be leveraged to identify GAN-generated fake images effectively [12].

4. Existing works:

Numerous studies have focused on developing deep fake image detectors using convolutional neural networks (CNNs). Jonathan et al. [14] conducted a comprehensive analysis of various fake image detectors based on deep neural networks. They evaluated the performance of these detectors using a dataset of 36,302 photos and found that both conventional and deep learning detectors can achieve detection accuracies of up to 95%. However, deep learning detectors demonstrated higher accuracy, maintaining up to 89% accuracy even on compressed data. This highlights the effectiveness of deep learning-based approaches in detecting deep fake images.

Mo, H. et al. [15] proposed a CNN-based approach for recognizing artificially created faces. Their experimental results supported their claims, showing that the system achieved an average accuracy of over 99.4%. This indicates the high potential of CNNs in accurately identifying deep fake faces.

In another study, a modified face identification method based on a hybrid ensemble learning approach was developed [16]. The experiment achieved an accuracy of 84.7%, while a pre-trained VG-Face model achieved an accuracy of 89%. These results demonstrate the effectiveness of ensemble learning methods in improving the detection accuracy of deep fake images.

5. Objective:

The primary objective of this study is to develop effective methods for identifying deep fake images using a diverse set of facial images. The first approach proposed in this research involves the utilization of a transformer-based architecture specifically designed for deep fake recognition. Known as the Vision Transformer (ViT), this methodology applies a Transformer-like design to specific regions of the image. The image is divided into fixed-size patches, and each patch is linearly embedded with the addition of position embeddings. The resulting sequence of vectors is then processed by a conventional Transformer encoder. To facilitate classification, a traditional approach is adopted, which involves incorporating an additional "classification token" in the sequence.

The second approach in this study explores a variation of the initial pipeline, as depicted in Figure 2. Instead of using the vision transformer as the classifier, an EfficientNet B07 architecture is employed for comparative analysis.

By investigating the performance of both the vision transformer and EfficientNet B07 architectures, this study aims to contribute to the development of robust deep fake image detection methods.

6. Dataset:

For this study, a single dataset was utilized to train and evaluate the proposed deep fake detection approaches. The dataset used is the "140k Real and Fake Faces" dataset, which is publicly available on Kaggle. This dataset comprises a collection of 70,000 real faces obtained from the Flickr dataset compiled by Nvidia, as well as a sample of 70,000 fake faces generated using StyleGAN by Bojan. The dataset is organized into separate folders for training, validation, and testing, with each image resized to a resolution of 256 pixels [17].

By utilizing this dataset, the study aims to train and evaluate the proposed deep fake detection models effectively.

7. Evaluation Matrix:

In this study, various evaluation metrics will be employed to assess the performance and accuracy of the classifiers. Two important metrics that will be reported are precision and recall. These metrics are particularly useful when dealing with imbalanced class distributions. Precision measures the proportion of correctly identified positive instances out of all instances classified as positive, while recall measures the proportion of correctly identified positive instances out of all actual positive instances [18].

To visualize the tradeoff between precision and recall at different classification thresholds, the study will utilize a precision-recall curve. The area under this curve is indicative of the overall performance of the classifier, with a larger area suggesting better recall and precision. A high accuracy score implies a low false-positive rate, meaning that the classifier provides accurate results.

Similarly, a high recall score indicates a low false-negative rate, indicating that most positive outcomes are correctly identified.

By considering these evaluation metrics, the study aims to comprehensively assess the performance and effectiveness of the classifiers in detecting deep fake images.

8. Methodology:

The methodology of this study consists of three main components: pre-processing, training, and validation, as depicted in Figure 1. The first step involves selecting a dataset, after which the images in the training set are used to train the vision transformer model. During the pre-processing stage, the image data is normalized and resized to ensure consistency and homogeneity.

The primary focus of this research is to develop a deep fake detection method using a conventional vision transformer architecture without utilizing any pre-trained models. Additionally, for comparative analysis, an EfficientNet B07 CNN architecture is employed. By comparing the performance of these two architectures, the study aims to provide insights into their effectiveness in detecting deep fake images.

Overall, the methodology encompasses dataset selection, pre-processing, training the vision transformer model, and conducting a comparative study with the EfficientNet B07 architecture.

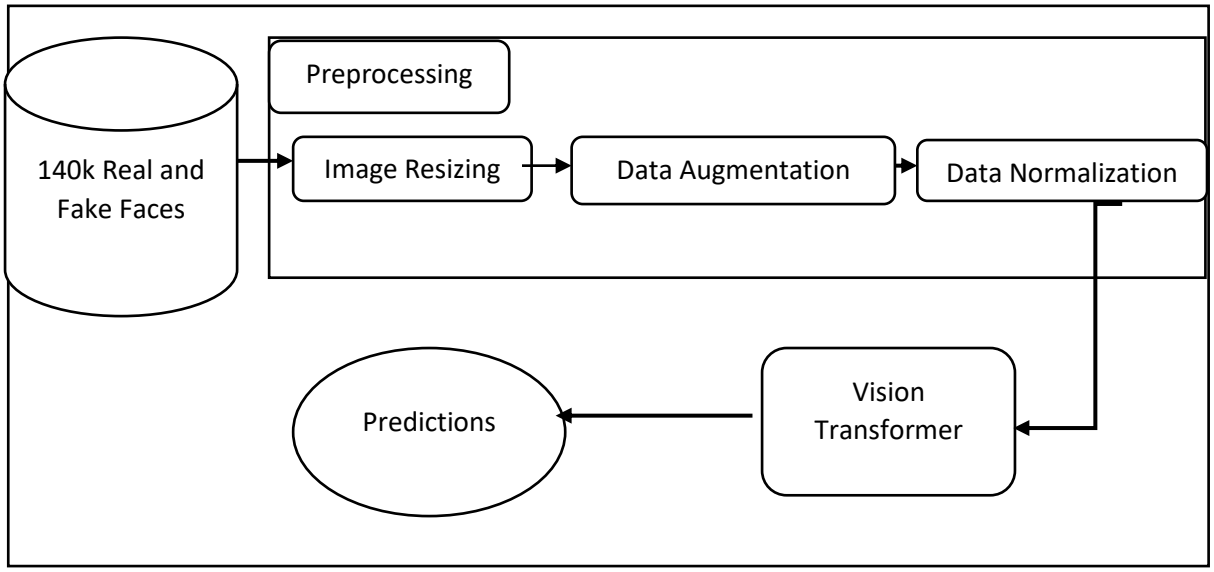

Figure 1: Proposed Method

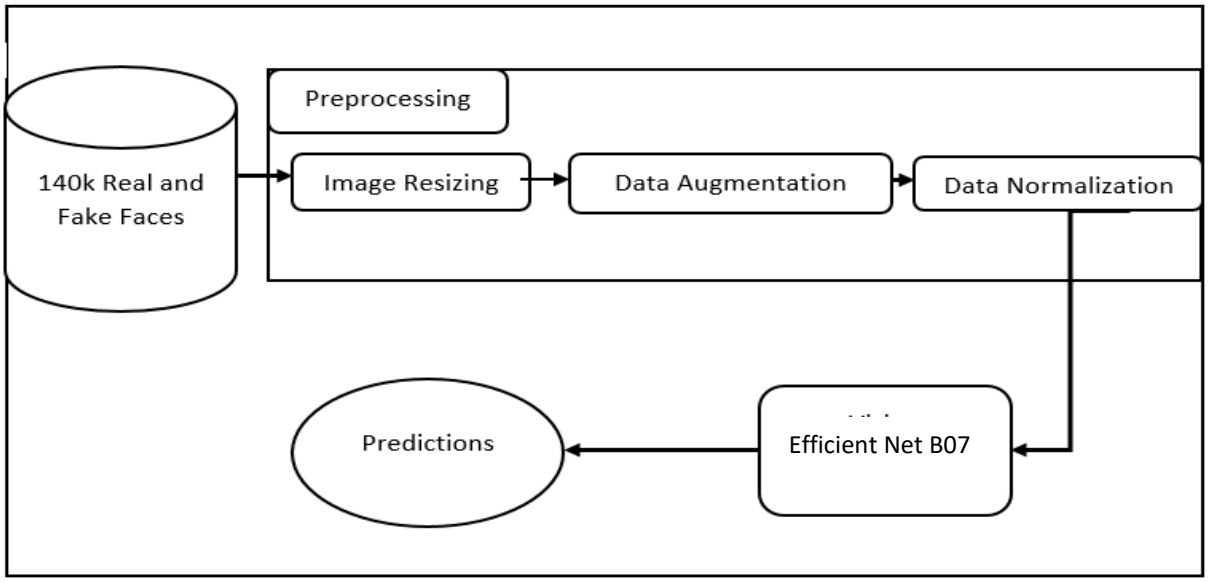

Figure 2: Pipeline variation with Efficient Net B07 architecture

8.1. Data Preprocessing:

To ensure balanced representation of the classes (real and fake), data preprocessing was performed on the 140k Real and Fake face dataset. From the original dataset of 140,000 images, a subset of 12,500 images was extracted for the experiment. This subset was divided equally into two classes: 5,000 images for training real face images and 5,000 images for training deep fake images. Similarly, the test set consisted of 2,500 images, equally distributed between the two classes, maintaining the same distribution as the training set.

The following table provides a detailed breakdown of the training and test sets:

Data Set	Real Face Images	Fake Face Images
Training Set	5000	5000
Test Set	1250	1250

Table 1 Real Face and Fake Face data division

By balancing the number of images for each class and ensuring an equal representation in the training and test sets, the study aims to

establish a fair and consistent evaluation of the deep fake detection models.

Prior to being fed into the classifiers, the images containing real and fake faces undergo several preparatory steps. Firstly, a normalization process is applied to adjust the size of the images to 224x224 pixels, ensuring uniformity across all images. This step helps in achieving consistency in image representation.

Furthermore, a data augmentation process is performed on the images to increase their diversity and improve the model's ability to generalize to new and unseen images. Data augmentation techniques such as random rotations, flips, and zooms are applied to generate variations of the original images. These augmented images are then stored in a dataset that will be utilized for both training and evaluation of the vision transformer and EfficientNet B07 architecture.

By normalizing the image sizes and applying data augmentation, the study aims to enhance the model's performance by increasing the diversity of the training data and improving its ability to handle different variations of real and fake faces.

8.2. Vision Transformer Architecture:

The Vision Transformer (ViT) is an extension of the original Transformer architecture, which has demonstrated strong performance in natural language processing tasks such as machine translation. The Transformer's unique encoder-decoder structure enables parallel processing of sequential input without the need for recurrent connections. The self-attention mechanism in Transformers has been a key factor in their success, allowing them to capture long-range dependencies within a sequence.

The Vision Transformer aims to apply the Transformer architecture to the task of image classification, expanding its application beyond text-based tasks. Unlike traditional Convolutional Neural Network (CNN) designs that rely on filters with small receptive fields, the Vision Transformer utilizes the attention mechanism to attend to different parts of an image and integrate information across the entire image. This approach does not involve specific architectural modifications

tailored to image data, with the primary goal being to generalize the Transformer architecture to other modalities beyond text.

In the Vision Transformer, the encoder module of the Transformer is employed explicitly for classification purposes by mapping a sequence of image patches to their corresponding semantic labels. This enables the Vision Transformer to capture meaningful representations of the image content. The attention mechanism plays a crucial role in the Vision Transformer, allowing it to attend to various regions of the image and effectively integrate information from different parts. This stands in contrast to standard CNN designs that often use small filters with limited receptive fields [19].

The architecture of the Vision Transformer model can be visualized in Figure 3. It consists of a final head classifier, an encoder, and an embedding layer. To process an input image X from the training set, the first step involves dividing the image into non-overlapping patches. Each patch is treated as a separate token by the Transformer. For an image of size X, where c represents the number of channels, h is the height, and w is the width, patches of size c x p x p are extracted from each dimension.

By extracting patches from the image, a series of patches is obtained, denoted as (x1, x2,..., xn), where n is determined by the formula n = hw/p^2. The value of p typically ranges from 16 by 16 to 32 by 32, with a lower patch size resulting in a longer sequence and vice versa. This sequence of patches is then fed into the encoder module of the Vision Transformer, which processes and encodes the information from the patches. The attention mechanism in the encoder allows the model to attend to different parts of the image and capture the relationships between the patches.

It's worth noting that the citation number [20] refers to the source where the information about the typical patch size selection can be found.

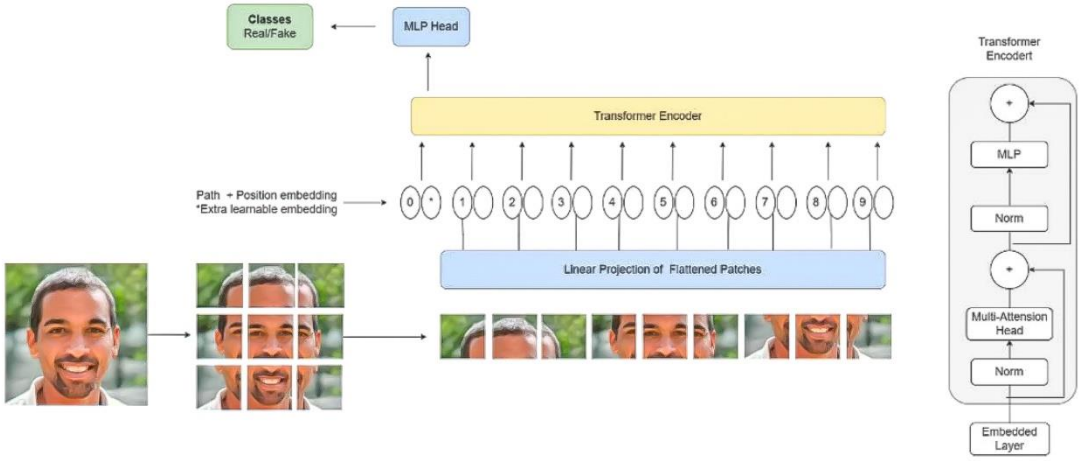

Figure 3 Vision Transformer Architecture

8.3. Efficient Net B07 Architecture:

To prepare the image samples for our convolutional neural network (CNN), we perform preprocessing steps. Initially, the images are resized to a standard size of 224 by 224 pixels. If the images are in grayscale, we convert them to RGB format by replicating the intensity values across all three color channels. Following this, the RGB images undergo pixel normalization.

After completing the image preprocessing, our CNN is ready to receive the input data. However, before feeding the data to the CNN, the training data undergoes a data augmentation stage. This process enhances the diversity of the dataset without requiring additional data collection.

In our study, we employ Keras's sequential model, with EfficientNet B07 serving as the first layer of our model. EfficientNet is a convolutional neural network design that incorporates a scaling technique using a compound coefficient. This approach ensures consistent scaling of depth, breadth, and resolution parameters by utilizing a predefined set of scaling coefficients. This is in contrast to the conventional practice of arbitrarily scaling these elements [20].

Subsequently, our architecture includes a two-dimensional Global Average Pooling layer. This layer is intended to fulfill the role of fully connected layers typically found in traditional CNNs. We compute the average value for each feature map, and the resulting vector is directly fed into a softmax layer, eliminating the need for constructing fully connected layers on top of the feature maps. Additionally, we introduce a dropout layer with a dropout rate of 20%. This technique randomly disregards or drops out a specific number of neurons in the network during training.

Finally, we include a fully connected layer with a sigmoid activation function, which is suitable for our binary classification task. This layer produces the final output probabilities. Figure 4 depicts the architecture of EfficientNet B07.

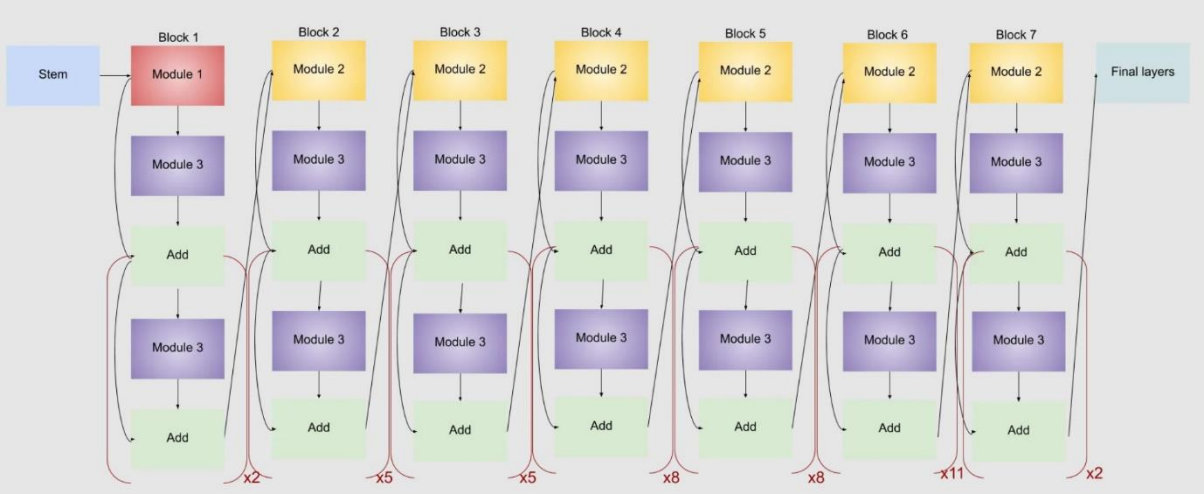

Figure 4 Efficient Net B07 Architecture

9. Results and Discussion:

9.1. Vision Transformer Architecture Accuracy:

The summary of accuracy scores is displayed in table 2 below. According to results, the visions transformer architecture achieved an overall accuracy of 83.56 %. The vision transformer classifier achieved a precision, recall and f-1 score as following-

	Precision	Recall	F1-Score
Real Face	0.86	0.80	0.83
Fake Face	0.81	0.87	0.84

9.2. 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 tech-niques. Table 3 below shows the overall accuracy scores of our CNN architecture. Efficient Net classifier was able to achieve an overall accuracy of 84.88%. This entails that the Efficient net architecture outperformed the vision transformer architecture by 1.32 % in terms of accuracy. Efficient Net architecture achieved an overall precision score recall and f1 score as following:

	Precision	Recall	F1-Score
Real Face	0.85	0.85	0.85
Fake Face	0.85	0.85	0.85

10. Conclusion:

In this comparative study on deep fake recognition, we investigated two different classifiers: a vision transformer and an EfficientNet B07 architecture. Both deep learning approaches showcased several advantages and achieved satisfactory accuracy scores. The vision transformer model achieved an accuracy rate of 83.56%, while the CNN-based architecture achieved an accuracy rate of 84.88%. As the need for detecting image and video forgery and tampering continues to grow in the field of digital forensics, the development of effective systems for deep fake photo detection becomes crucial. The results obtained from this study demonstrate promise, and further improvements can be made by expanding the dataset and fine-tuning the hyperparameters. These findings serve as a valuable example for the creation of standardized tools aimed at detecting and classifying deep fake images, as well as identifying instances of forgery.

11. Limitations:

1. Dataset Bias: The performance of the deep fake image detection model heavily relies on the quality and diversity of the dataset used for training. If the dataset is biased or lacks certain types of deep fake variations, the model may not generalize well to real-world scenarios.

2. Generalization to New Deep Fake Techniques: As deep fake techniques evolve and new methods are developed, the trained model may struggle to detect previously unseen or sophisticated deep fake images. Continuous monitoring and updating of the model with new data and techniques is necessary to enhance its detection capabilities.

3. Computational Resources: Training deep learning models, especially with large-scale datasets and complex architectures like Vision Transformers and EfficientNet, requires significant computational resources. The limitation of computational power may restrict the scalability and efficiency of the model.

12. Future Scope:

1. Adapting to Video-based Deep Fakes: The current project primarily focuses on detecting deep fake images. Extending the model's capabilities to detect deep fake videos would be a valuable future direction. Video-based deep fake detection involves analyzing temporal dependencies and spatio-temporal inconsistencies, which can be challenging but essential for combating video-based misinformation.

2. Countermeasures Against Advanced Deep Fakes: As deep fake techniques continue to advance, exploring countermeasures to identify and mitigate the impact of advanced deep fakes becomes crucial. This may involve developing additional features or algorithms that can detect subtle artifacts or inconsistencies specific to advanced deep fake methods.

3. Real-Time Deployment: Integrating the deep fake detection model into real-time systems and platforms, such as social media platforms or video streaming services, would enable proactive identification and prevention of the spread of deep fake content. Real-time deployment would require optimizing the model for efficient inference and addressing the challenges of processing high volumes of data in real-time.

4. Robustness Against Adversarial Attacks: Investigating the model's robustness against adversarial attacks specifically designed to deceive the deep fake detection system would be essential. Adversarial attacks aim to manipulate or evade the detection mechanisms, and developing defenses against such attacks would enhance the reliability and effectiveness of the model.

5. Collaborative Efforts and Benchmarking: Collaborating with researchers and organizations working on deep fake detection to share datasets, evaluation metrics, and benchmarking protocols would facilitate the development of more robust and standardized deep fake detection systems. This would help in comparing and advancing the performance of different models, promoting the field's progress as a whole.

13. 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 seam-less 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 Fo-rensics Analysis: An explainable hierarchical ensemble of weakly supervised models. Fo-rensic 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 embed-ding 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 Pro-cessing 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. Pro-ceedings 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).