Real and Fake Face detection in the field of data security Using Deep Learning

Oisharya Dhar
Reg: 200441728

Abstract

From the past few decades, deep learning has played an important role in solving various complex problems even in the field of computer vision and human-level control. A particular part of the deep neural network, the Convolutional Neural Network(CNN) has significantly achieved high accuracy in the field of classification. Using the leverage of deep learning a new technology emerges called deepfake which generates fake images. However, some attackers are using this deepfake to make fake identity to intrude in the data security system. For that reason, in this project, a deep learning algorithm is proposed using a pre-trained model named MobileNet(MNet) which determines if the face image is fake or not. Further to increase the efficiency, the MNet is modified using the convolutional neural network(CNN). Moreover, data augmentation is used for increasing the variation and the number of images of the dataset and this model achieved an accuracy of 79%.

KeyWords: DeepFake, Convolutional Neural Network(CNN), MNet, data augmentation

1 Introduction

Face recognition technology uses different facial attributes to detect and recognize a face such as facial attributes, facial landmarks etc. In the present world, most face recognition models are based on artificial intelligence and are used for different data security purposes such as for identification for accessing a particular dataset or retrieving important information from the dataset. As the importance of facial recognition is increasing the malicious technique for faking it is also increasing. And for this reason, fake facial images are being used. Fake images and videos mainly refer to the deform facial information that is mainly generated by any means of manipulation. This can be done both by handcrafting and digitally. Traditionally, with handcrafted fake images, the realistic facial expression is limited because of the lack of sophisticated editing tools. For example, initial work in facial expression manipulation [8] was only able to modify the lip motion of a person.

In the case of using AI for this purpose, there exists a category named DeepFakes, an artificial intelligence-synthesized content, which is mainly based on face swap. Facial manipulation can generally be divided into the below three categories[1],

- Face synthesis
- Face swap
- Facial attributes and expression

So for ease of mechanism, deep neural models are used, from examining facial expressions to synthesizing facial images and making artificial expressions and movements [2]. Moreover, these DeepFakes are also used for the videos modification. For example, Puppet-master[3] DeepFakes that include videos of a target person (who acted as a puppet) are animated using the facial, eye expression and head movements of another person (who is the master). As DeepFake uses AI to generate fake images and videos, it requires a large amount of data containing both image and video for training and creating realistic fake face images and videos. Thus public figures are the main target of DeepFakes, as they have a large number of images and videos available across the web which can play the role of the training dataset. The first DeepFake video for the malicious purpose was generated in 2017, where the face of a celebrity was swapped with the face of a porn actor. And it become an overnight threat to the world security system as this method can also be used to make a false provoking political speech as well[4]–[6].

As every aspect has both positive and negative sides this DeepFake also has some positive sides. Because of this DeepFake visual effects, digital avatars, Snapchat filters, creating voices have come up with a new form of entertainment[7]. However, the number of malicious uses of DeepFakes largely dominates that of the positive ones and that's why distinguishing fake faces from real ones has become an important area for research. Different kinds of detection methodologies are being tested by researchers to find the optimum one. At first, researchers mainly focused on characteristics of image compression information but these characteristics can easily be manipulated in such a way that it becomes impossible to detect the modified one. After that, the researchers also moved to a more sophisticated approach which is using deep neural networks.

In the case of data security, the impact of detecting a fake face from a real face is huge and some of them are,

- Hackers now attempt to modify data while leaving it in place[9].
- The cost of this scam was estimated to exceed $250 million in 2020[10].
- Fake identification is a technique that attackers use to copy identification or authentication to gain illegitimate access to vast data.
- In the case of biometric information.

2 Related Work

Detection of real and fake parts of images still is a challenging task because the attackers are also developing and using the latest image processing and machine learning techniques to improve their forgery for making fake images. One of the most popular methodologies that researchers use in the field of image analyzing is the frequency domain. This is very useful because in compressed JPEG the compressed history and other information remain saved. So by using this property of images as frequency the researchers can analyze the manipulated area on each image. One of the earlier attempts of using frequency domain is the JPEG Ghost[14]. Generally, when an image is manipulated of any kind, the manipulated part usually is copied from other images with different JPEG qualities. And using the different JPEG quality on the same image and then the JPEG Ghost extracts the difference between them as a feature. Then using the feature they detect the fake one from the real one. The new technologies for tackling this problem use a neural network. DeepFakes are not just creating fake images and videos but it heavily poking the privacy to privacy, security and democracy all together [11]. As soon as the attack using DeepFake begun researchers are also trying to tackle this problem, and proposing many algorithms and techniques to prevent this threat as soon as possible. As early fake images are handcrafted so as the detection methods. Handcrafted features were obtained by using the frequency domain but recently applied deep learning is taking its place to automatically extract features to detect DeepFakes[12], [13].

Convolutional Neural Network which excels in image classification and recognition is being used to distinguish between manipulated images and real ones. Moreover, two specific pre-trained models, VGG16 and VGG19 [15] have paved the way for large-scale image recognition and they are also considered deep neural models. Though the success rate for large-scale images is high for these two deep models but training them is harder than any normal deep neural model.

Zhang et al.[16] used the bag of words method to extract a set of compact features and fed it into various classifiers such as SVM, random forest (RF) and multi-layer perceptron (MLP) for discriminating swapped face images from the genuine. Xuan et al.[17] used an image preprocessing step, e.g. to remove low-level

high-frequency they have used Gaussian blur and Gaussian noise. Hsu et al.[18] introduced a two-phase deep learning method for the detection of DeepFake images. For feature extraction, they have used the common fake feature network (CFFN) and as the backbone, the Siamese network architecture is used. These features are then fed to a small CNN which is concatenated with the last convolutional layer of CFFN. Dang et al.[19] proposed an attention-based mechanism to improve the feature maps for detecting and classification task. As this detecting problem can be classified as a binary classification problem, they also highlighted the informative regions (genuine face vs fake face)to improve the classification accuracy.
Chen et al.[20] proposed a novel detection method that combines both spatial domain and frequency domain which works as the input. They mainly used the Multilevel Facial Semantic Segmentation and Cascade Attention Mechanism for feature extraction and detection. They[20] further modify the structure of Xception[21], for the increment of accuracy.

2.1 Dataset

As the images are used for attacks are fake so they are non-existing and this manipulation is generated based on real images. Table I summarises the main publicly available datasets. Most of the effective datasets are based on two types of GAN architectures: ProGAN and StyleGAN [22]. In addition, there are some other kinds of dataset generators for fake faces using basic deep learning such as FaceForensics[23].

Table I: Publicly Available Datasets

Name	Number of fake images	Number of real images	Total Image
iFakeFaceDB	87000	----	87000
CelebA	5000	5000	10000
LFW face database	----	13332	13332
100k-Generated Images	100000	----	100000
real_and_fake_face	960	1081	2041

3 Methodology

The whole methodology can be divided into data preprocessing as in data augmentation and developing the model.

3.1 Dataset

The dataset used for this model is the real_and_fake_face dataset which contains both real and fake images. Summary of the dataset;

- Total images: 2041
- The real face: 1081, fake face: 960

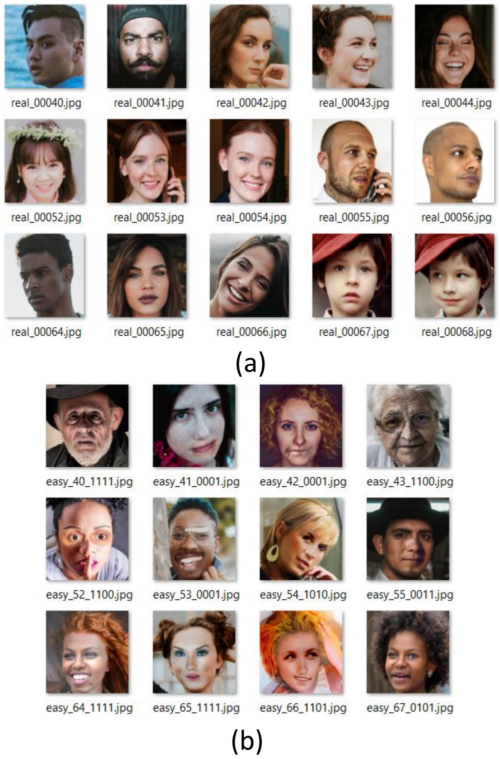


Figure 1:(a) images of real faces, (b) images of fake faces

Figure1 shows the images from the dataset. Here, figure 1(a) shows the dataset for real faces and 1(b) shows the fake generated faces.

3.2 Data Augmentation

Data augmentation is used when the dataset is relatively small. The small dataset may cause underfitting and thus the model will fail to achieve the desired accuracy. In the case of data augmentation random flip, random cropping is mostly used. In this model, the augmentation that is used,

- Rotation: As faces are found always in the horizontally straight position for that reason rotation is kept at 0.
- Width and Height Shift: This determines the amount of horizontal and vertical shift respectively. Here, for our model the range is kept at 0.1.
- Shearing: As images can be found in different illumination and for that reason, the shearing range in data augmentation of our model is kept at 0.1.
- Zoom: Zoom augmentation zooms the image randomly adds new pixel values around the original one. Here for our model, it is also kept 0.1.
- Horizontal and Vertical Flip: this augmentation is kept off.

Figure 2 shows an augmented image from the dataset, which is slightly distorted than the original image.

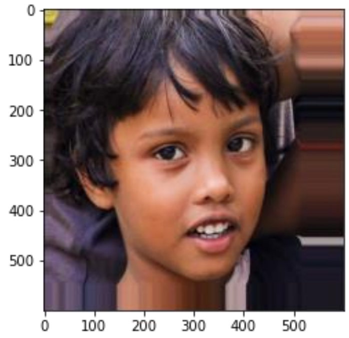


Figure 2: An augmented image

3.3 Model

The deep neural-based model contains different kinds of layers and functions such as a convolutional layer, fully connected layers, max-pooling and so on. A brief overview of these functions are given below,

- MNet: MNet or Mobile Net is a pre-trained convolutional neural network and is useful when the dimension of images of the dataset is relatively small. It uses depthwise separable convolutions and thus helped with smaller datasets. It reduces the number of parameters when compared to the network with regular convolutions with the same depth in the nets.
- Convolutional Layer works as a filter that strides through the whole input image and generates a feature map. The height and weight of the filters are smaller than the input image.
- Dense Layer or Fully connected layers are those neural layers where each neuron or node gets the input from all the neurons of the previous layer (flatten). Here, all the nodes are fully connected.
- Average pooling is the way of taking the average value from the patch of the feature map.
- Batch normalization is used for making the model faster and more stable. It normalizes the layers' inputs by re-scaling and re-centring.
- Adam optimizer is an adaptive learning method. It computes individual learning rates for different parameters.
- Cross-Entropy Loss: It measures the performance of a classification model whose output is a probability value between 0 and 1.
- Activation function:

  1. Rectified Linear Unit (ReLU) is used as the activation function for all the convolutional layers and the first two Dense layers in this model.
  The purpose of the ReLU is to give non-linearity to the model. It set all the negative values to zero and preserves only positive values. This function helps to reduce unnecessary noises from the data.
  2. Soft Max activation function is used for the last dense layer in this model. The function is used when the output value is needed to be normalized. It converts the inputs from weighted sum values into probabilities that range from zero to one.

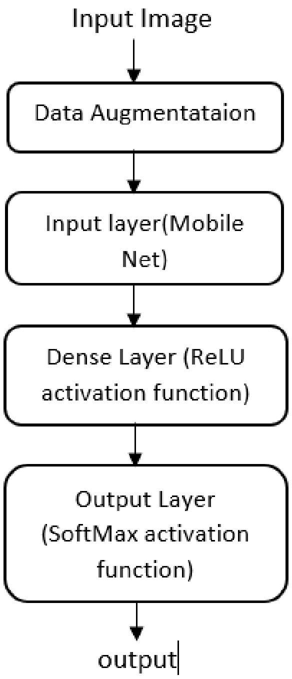

Input Image
Data Augmentataion
Input layer(Mobile Net)
Dense Layer (ReLU activation function)
Output Layer (SoftMax activation function)
output

Figure 3: Flow diagram of the model

Figure 3, shows the whole flow diagram of the model from data augmentation to output. From the flow chart, it is clear that the dense layer used the ReLU activation function and the output layer used the SoftMax activation function. Moreover, the Mnet is modified using this dense or fully-connected layer of CNN.

3.4 Evaluation Metrics

The proposed model is evaluated using Precession, Recall and F1 score.

- Precision: Precision can be defined as the percentage of the detection of true positive among all the predictions (false positive+ true positive). The equation [24] can be given as,

$$\frac{True\ Positive}{True\ Positive + False\ Positive}$$

- Recall: Also known as sensitivity. A model can find all the relevant cases. The equation [24],

$$\frac{True\ Positive}{True\ Positive + False\ Negative}$$

- F1 Score: The F1 score is the harmonic mean of the previous two (precision and recall) and can be shown as [25],

$$2 * \frac{Precesion * Recall}{Precesion + Recall}$$

3.5 Training Process

The training process can be divided into two major steps, data pre-processing and training the model. The whole process takes approximately 24 hours of which the maximum time is taken for the training process. Using this image generator function, train image and test image generators were generated. This train image generator and test image generator will be fed to the model. The average image size is used as the target size for both these generators The next step is to feed the model with the generated train image generator for training purposes. The test image generator works as validation data for our model. For this model, the iteration value is set to 250, which means this training dataset will train the model in up to 250 epochs.

4 Experiment

For the experimenting purpose, the model is altered to see which specification gives the best accuracy for predicting fake faces.

Case1: For case 1 experiment are done with a model without early stopping with 100 epochs.

Case2: For case 2 experiments are done with a model with early stopping with 100 epochs.

Case 3: For case 3 experiments are done with a model without early stopping with 250 epochs.

Table II: Comparison of different Cases

Classes	Case 1			Case 2			Case 3		
	Precision	recall	F1 score	precision	recall	F1 score	precision	recall	F1 score
Real Face	0.84	0.20	0.33	0.53	1.00	0.69	0.99	0.61	0.75
Fake Face	0.52	0.96	0.67	0.33	0.00	0.69	0.69	0.99	0.81
Accuracy	-------	------	0.56	------	------	0.53	-------	------	0.79

Table 2 is the summary of all the experiments that are conducted by altering the model. It is clear from the table that Case 3 works the best with the highest accuracy and Case 2 works the worst among the three. In the case of real faces, case 3 produced the best result in all the sections (precision and F1 score). For case 2 all three matrices produce quite low scores.

Finally, when accuracy is generated from F1-score Case 3 gives more accuracy than cases 2 and 3. And it achieves an accuracy of 79%.

5. Results

From section 4, it is clear that Case 1 works the best for the given dataset. In figure 4, the loss for training and validation data is shown. It is clear from figure 4 that loss for both training and validation data drops as the number of epochs increases. Although the loss of the training set is less than the loss of validation still it is pretty low for the test set. The exact opposite result can be seen for the accuracy curve, the accuracy increased linearly with the number of epochs. Figure 5 represents the accuracy of training and validation data. For both figures 4 and 5, the blue line represents the scores for training data and the orange line represents the scores for validation data. The final score is given in table 2.

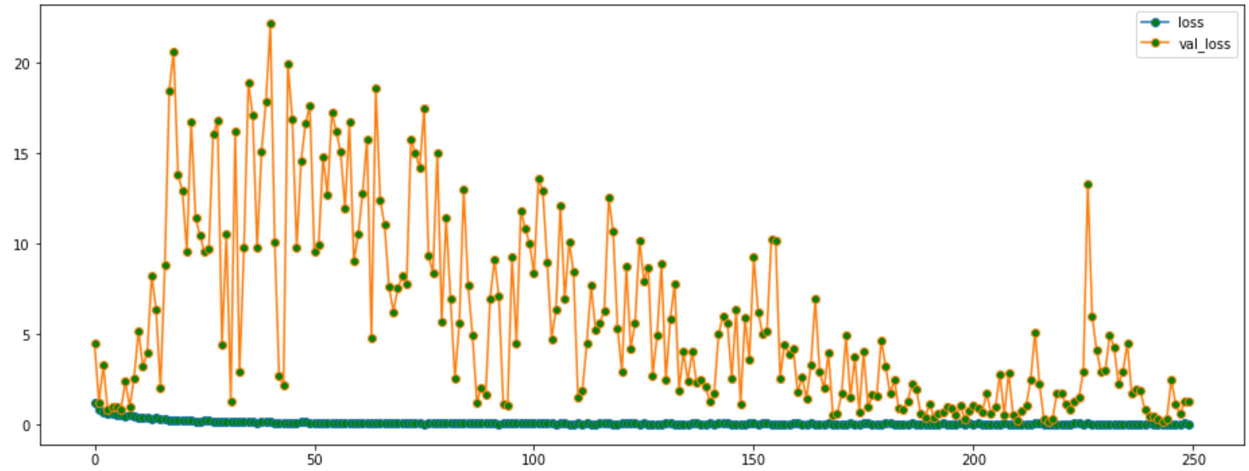


Figure 4: Loss score for train and test dataset

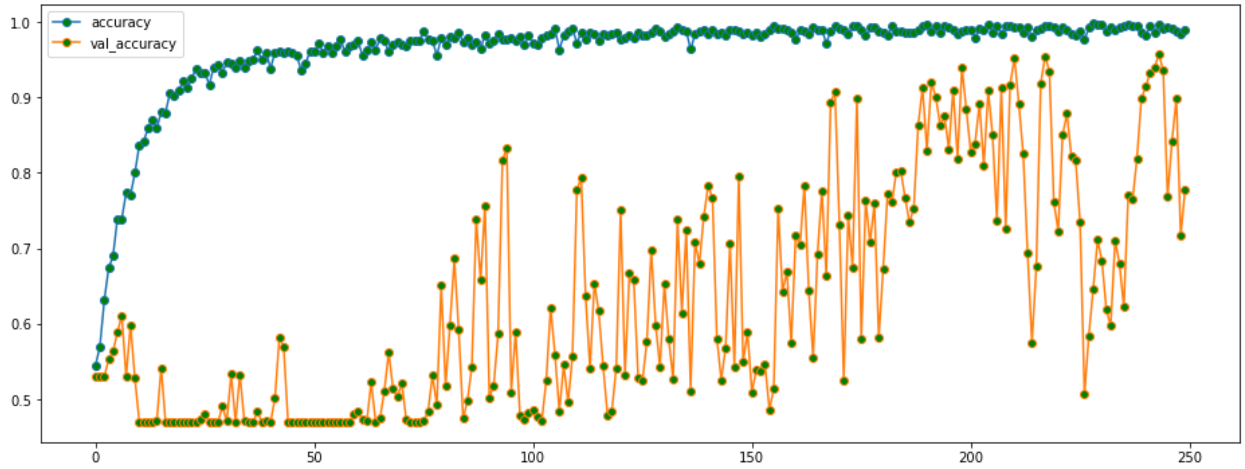
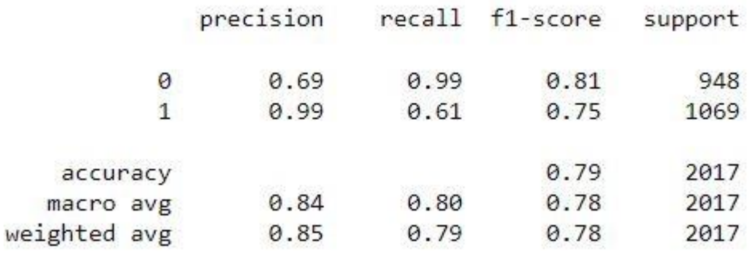


Figure 5: Accuracy score for both test and train dataset

All the three evaluation metrics scores are presented in figure 5. Here, Fake Faces and Real Faces classes are denoted as class 0 and class 1.



```
              precision    recall  f1-score   support

           0       0.69      0.99      0.81       948
           1       0.99      0.61      0.75      1069

    accuracy                           0.79      2017
   macro avg       0.84      0.80      0.78      2017
weighted avg       0.85      0.79      0.78      2017
```

Figure 6: Evaluation Matric Score for Final Model

6. Conclusion and Future Work

As malicious attacks using manipulated face images are increasing daily, to date many security aspects are facing a shortage of resources to identify all the attacks properly. For that reason, this paper proposed a deep learning approach to detect fake faces from real faces within the shortest amount of time by using pre-trained MNet. Most of the traditional methods trained the models from the scratch and for that reason, the training time is way too much. Using this pre-trained model, the training time can be reduced significantly. In future, the overall accuracy can be increased by fine-tuning the parameters. Different specifications of the neural model can be applied to determine for which specification the overall precision increases. Moreover, this model can be trained into a bigger dataset for making a more accurate prediction.

References

[1]. Theaisummer.com. 2021. [online] Available at: <https://theaisummer.com/DeepFakes/> [Accessed 5 December 2021].

[2] Lyu, S., 2018. Detecting DeepFake videos in the blink of an eye. *The Conversation*, *29*.

[3] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K. and Li, H., 2019, June. Protecting World Leaders Against Deep Fakes. In *CVPR workshops* (Vol. 1).

[4] Nguyen, T.T., Nguyen, C.M., Nguyen, D.T., Nguyen, D.T. and Nahavandi, S., 2019. Deep learning for DeepFakes creation and detection: A survey. *arXiv preprint arXiv:1909.11573*.

[5] Chesney, R. and Citron, D., 2019. DeepFakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Aff.*, *98*, p.147.

[6] Hwang, T., 2020. DeepFakes: A Grounded Threat Assessment. *Centre for Security and Emerging Technologies, Georgetown University*.

[7] Marr, B., 2019. The best (and scariest) examples of AI-enabled DeepFakes. *Re-trieved from https://www. forbes. com/sites/bernardmarr/2019/07/22/the-best-and-scariest-examples- of-aienabled-DeepFakes*.

[8] Bregler, C., Covell, M. and Slaney, M., 1997, August. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques* (pp. 353-360).

[9] KanngiesserCo-founder, D., CEO, CryptowerkDecember 27 and 2018 (2018). *Toxic Data: How "DeepFakes" Threaten Cybersecurity*. [online] Dark Reading. Available at: https://www.darkreading.com/application-security/toxic-data-how-DeepFakes-threaten-cybersecurity [Accessed 5 Dec. 2021].

[10] Panda Security Mediacenter. (2021). *DeepFake Fraud: Security Threats Behind Artificial Faces*. [online] Available at: https://www.pandasecurity.com/en/mediacenter/technology/DeepFake-fraud/ [Accessed 5 Dec. 2021].

[11] Chesney, R. and Citron, D.K., 2018. Deep fakes: A looming challenge for privacy, democracy, and national security. 107 california law review (2019, forthcoming); u of texas law. *Public Law Research Paper*, (692), pp.2018-21.

[12] de Lima, O., Franklin, S., Basu, S., Karwoski, B. and George, A., 2020. DeepFake detection using spatiotemporal convolutional networks. *arXiv preprint arXiv:2006.14749*.

[13] Amerini, I. and Caldelli, R., 2020, June. Exploiting prediction error inconsistencies through LSTM-based classifiers to detect DeepFake videos. In *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security* (pp. 97-102).

[14] Farid, H., 2009. Exposing digital forgeries from JPEG ghosts. *IEEE transactions on information forensics and security*, *4*(1), pp.154-160.

[15] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[16] Zhang, Y., Zheng, L. and Thing, V.L., 2017, August. Automated face swapping and its detection. In *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)* (pp. 15-19). IEEE.

[17] Xuan, X., Peng, B., Wang, W. and Dong, J., 2019, October. On the generalization of GAN image forensics. In *Chinese conference on biometric recognition* (pp. 134-141). Springer, Cham.

[18] Hsu, C.C., Zhuang, Y.X. and Lee, C.Y., 2020. Deep fake image detection based on pairwise learning. *Applied Sciences*, *10*(1), p.370.

[19] Dang, H., Liu, F., Stehouwer, J., Liu, X. and Jain, A.K., 2020. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition* (pp. 5781-5790).
[20] Chen, Z. and Yang, H., 2021, June. Attentive semantic exploring for manipulated face detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1985-1989). IEEE.
[21] Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258).
[22] Karras, T., Laine, S. and Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4401-4410).
[23] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Nießner, M., 2018. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*.
[24] Hansell, David M., Alexander A. Bankier, Heber MacMahon, Theresa C. McLoud, Nestor L. Muller, and Jacques Remy. "Fleischner Society: glossary of terms for thoracic imaging." *Radiology* 246, no. 3 (2008): 697-722.
[25] Apostolopoulos, Ioannis D., and Tzani A. Mpesiana. "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks." *Physical and Engineering Sciences in Medicine* 43, no. 2 (2020): 635-640.