IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116113

# Comparative study on Deepfake Detection Methods

**Darshan V Prasad[1], Harsha M[2], N Navneeth Krishna[3], Sanjay T.C[4], Dr. Kiran Y C[5]**

Student, Global Academy of Technology, Bengaluru, India[1 2 3 4]

Professor and HOD, Department of Information Science and Engineering, Global Academy of Technology, Bengaluru, India[5]

**Abstract:** Deep learning algorithms have recently expanded their applications beyond big data analytics to include intrusion detection systems. Artificial intelligence and image processing advances are changing and challenging how people interact with digital images and video, and because of their intrinsically contentious character and the reach of contemporary society, they are intended to propagate harmful content and disinformation to millions of people.

A picture may say a thousand words, but what if the photograph has been fabricated? The term "fake news" has recently gained popularity, yet with today's photo manipulation techniques, even the most vigilant eyes can be tricked. One of these areas is the use of several software such as faceapp and fakeapp to create modified media files known as deepfake. From massive data analysis to human biometric systems, deep learning algorithms are used.

Due to their user-friendly characteristics, these applications are growing more popular with the general public and are employed in a range of fields including digital fraud, cybercrime, politics, and even military actions. As a result, it's critical to build detection technologies that can detect and remove this form of forgery, as well as to take a new step forward in video and audio forensics.

## INTRODUCTION

Deepfakes are synthetic media created with software that portray humans talking or doing things they don't actually do. During the early stages of video and audio manipulation, several approaches in the field of image processing were developed. Anyone with even a single ounce of Photoshop skill can alter images to modify their contents, interpretation, and perhaps everything. However, this type of fabrication is also being widely researched in recent years, and commercial tools that can identify and explain it are also available. However, when compared to modern deep learning generation and detection systems involving auto encoders and deep convolution networks, the degree of accuracy has been quite poor.

In most cases, these learning phases necessitated a huge number of photos or videos in order to train the model to make photorealistic imitation duplicates. Typically, this dataset for free usage is created from publicly available photographs or videos of politicians and celebrities, which is why these people are the first targets of deepfake. The most well-known deepfake video was the one that resulted from Barack Obama's speech, which went viral on the internet and had an influence on political election in the United States. As a result, it has been classified as a national security danger since Deepfake is being used to create movies of various political figures for the purpose of political manipulation. This was even utilised to create a fictitious satellite picture of the Earth with an item that does not exist. This was done to deceive the military analysts and even lead a squad to cross a bridge in the middle of a conflict. Latest technological advancements have enabled the creation of deepfake films using only a single picture, posing a major threat to civilization. Along with the drawbacks, these deepfakes have several positives, particularly in sphere of media creation, where they may reproduce films of individuals who've lost their voices or update episodes without having to redo them. The groundwork for the creation of Deepfake in the latest scenario employing Deep Learning's Generative Adversarial Network (GAN). There are many GAN methods that is used to create deepfake images. Some of the GAN methods are StyleGAN, stackGAN, W-GAN, SR-GAN etc. The research in this field has lately released a number of deepfake datasets to aid other researchers in creating detection techniques for these deepfakes in an attempt to improve the quantity of data accessible.

DOI: 10.17148/IJARCCE.2022.116113

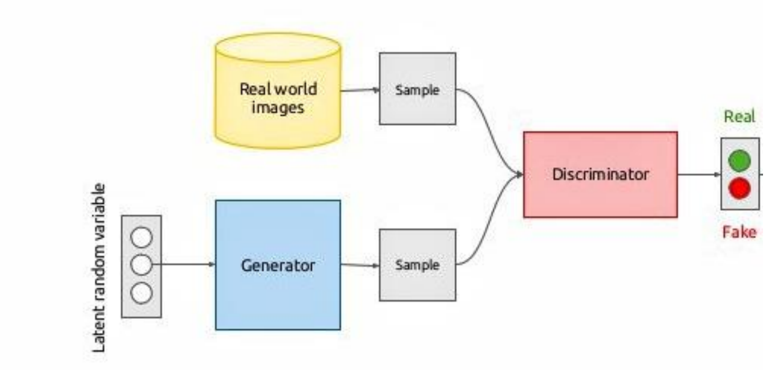

**Figure 1: GAN System Model**

As illustrated in fig 1, the deepfake video is generally made by employing two GAN networks which are based on the AI system. The first is known as the Generator, while the second is known as the Discriminator. Essentially, the generator is used to create deepfake videos, while the discriminator decides whether or not the video is fake. Each time the discriminator correctly recognises a video as fake, it provides a hint to the generator on how not to make the next Deepfake video. The generator and discriminator will combine to build a Generative Adversarial Network. The advancement of these deep learning techniques, which are employed in the execution of most of the online Deepfake generators, as well as their ease of use, has increased their popularity among both professionals and amateurs. Deepfake was created and generated using a variety of deep learning methods, including convolutional neural networks, recurrent neural networks, long short-term memory, and even a blend of these approaches. Finding the credibility of these digital evidences is a major issue for media forensics experts and investigators. To encourage greater research and development in the detection and prevention of Deepfake. Facebook and Microsoft have announced a Deepfake detection challenge. Google also sponsored a similar event by releasing a dataset (Google Net) for research purposes. More such benchmark datasets, on the other hand, aid in boosting performance and evolving methodologies. Nevertheless, it would be impossible to create a large dataset for each new deepfake creation method in order to train deep neural networks. As new sophisticated picture forging methodologies are released every day, distinguishing modified from authentic photographs is getting increasingly challenging. When trained on a specific counterfeiting method, naive classification systems based on Convolutional Neural Networks (CNNs) perform well at identifying changes in images and videos. On instances from unobserved manipulation techniques, however, their performance suffers noticeably.

## METHODOLOGY

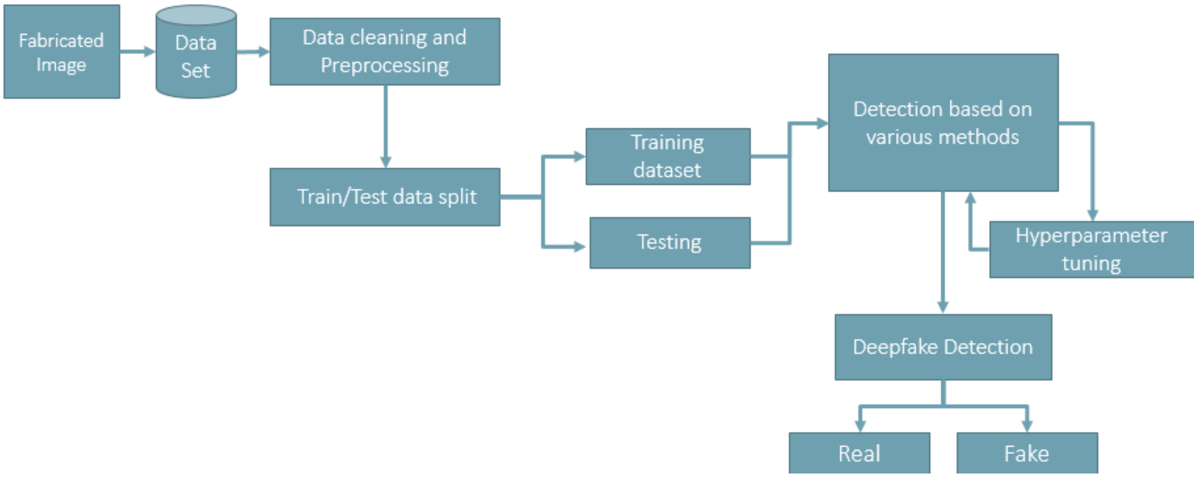

**Figure 2: Data Flow Diagram**

The dataset is available in Kaggle and is named as “140k Real and Fake faces”. This dataset consists of all 70k REAL faces from the Flickr dataset collected by Nvidia, as well as 70k fake faces sampled from the 1 million FAKE faces (generated by StyleGAN) that was provided by Bojan. In Pre-processing step, the images in the dataset was resized based on the input requirement of the training models. The dataset is split into training, validation and testing set.

Three CNN models were tested: VGG16, XceptionNet, ResNet50.

### VGG16

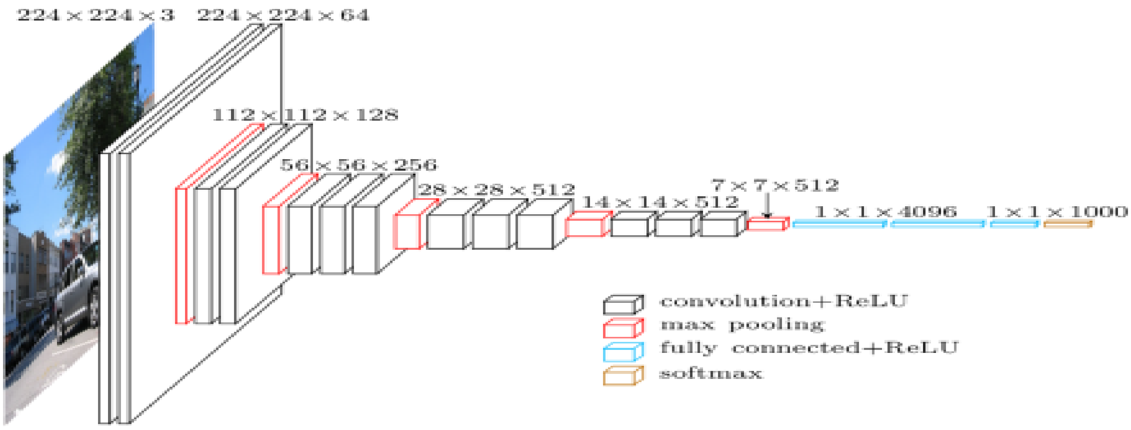

**Figure 3: VGG16 Architecture**

VGG16 is an object identification and classification algorithm that has a 92.7% [3] accuracy rating while classifying 1000 images into 1000 different categories. It is a well-liked technique for classifying images and is simple to employ with transfer learning.

The 16 in VGG16 stands for 16 weighted layers. Thirteen convolutional layers, five Max Pooling layers, three Dense layers, and a total of 21 layers make up VGG16; however, the learnable parameters layer only consists of sixteen weight layers. The input tensor size for VGG16 is 224, 244 and has three RGB channels. The most distinctive feature of VGG16 is that it placed a strong emphasis on having convolution layers of a 3x3 filter with stride 1 and always used the same padding and maxpool layer of 2x2 filter of stride 2. Throughout the whole architecture, the convolution and max pool layers are uniformly ordered. There are 64 filters in the Conv-1 Layer, 128 filters in Conv-2, 256 filters in Conv-3, and 512 filters in Conv-4 and Conv-5. A stack of convolutional layers is followed by three Fully-Connected (FC) layers, the third of which conducts 1000-way ILSVRC classification and has 1000 channels. The first two FC layers have 1024 channels each (one for each class). The soft-max layer is the last one.

DOI: 10.17148/IJARCCE.2022.116113

### XceptionNet

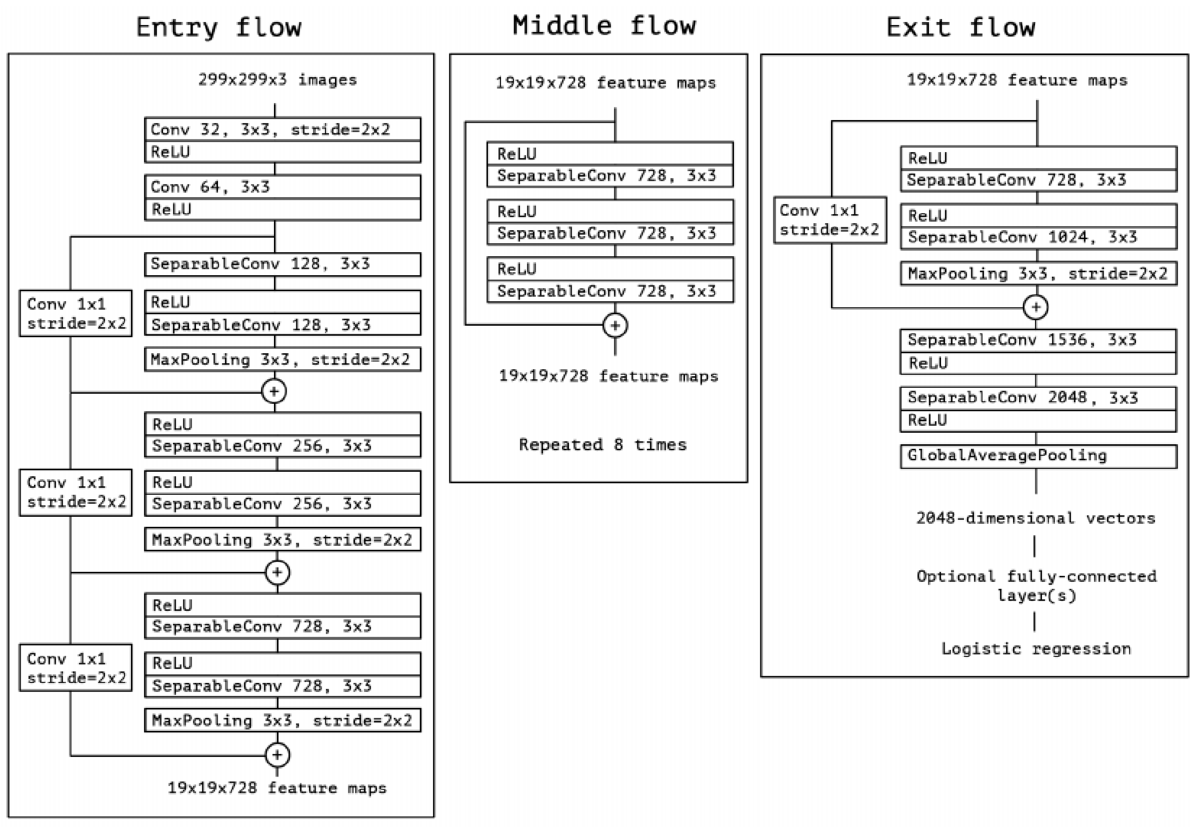

**Figure 4: XceptionNet Architecture**

Depthwise Separable Convolutions are used in the deep convolutional neural network architecture known as Xception. Xception, which means for "extreme inception," pushes the fundamental ideas of Inception to the limit. In Inception, the original input was compressed using 1x1 convolutions, and various filters were applied to each depth space based on the input spaces. Just the opposite occurs with Xception. Instead, it applies the filters to each depth map individually before using 1X1 convolution to compress the input space all at once. This process is nearly equivalent to a depthwise separable convolution, a technique that was first applied to the building of neural networks in 2014. Between Inception and Xception, there is yet another distinction. whether a non-linearity exists or not after the initial operation In the Inception model, a ReLU non-linearity follows both processes, whereas Xception doesn't add any non-linearity.

### ResNet50

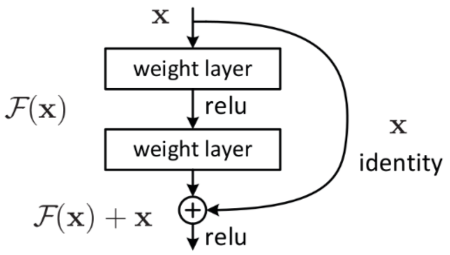
x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

**Figure 5: ResNet Skip Connection**

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116113

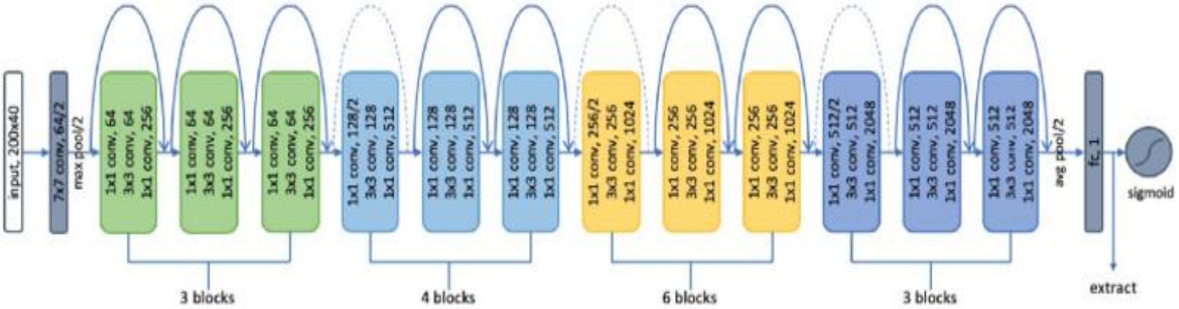

**Figure 6: ResNet50 Architecture**

A common neural network that serves as the foundation for many computer vision applications is called ResNet, short for Residual Networks. This model was the victor of the 2015 ImageNet challenge. ResNet represented a significant advancement in that it successfully enabled us to train incredibly deep neural networks with more than 150 layers. Due to the issue of vanishing gradients, very deep neural network training was challenging before ResNet. Skip connection was first introduced by ResNet. There are 5 stages in the ResNet-50 model, each with a convolution and an identity block. Each identity block and each convolution block each have three convolution layers. There are around 23 million trainable parameters in the ResNet-50.

In hyperparameter tuning, the number of epochs and batch were changed and the result was recorded.

## RESULT

The metrics used is the ROC AUC score. ROC-AUC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. The ROC-AUC score of three models is given below in the form of a bar graph.

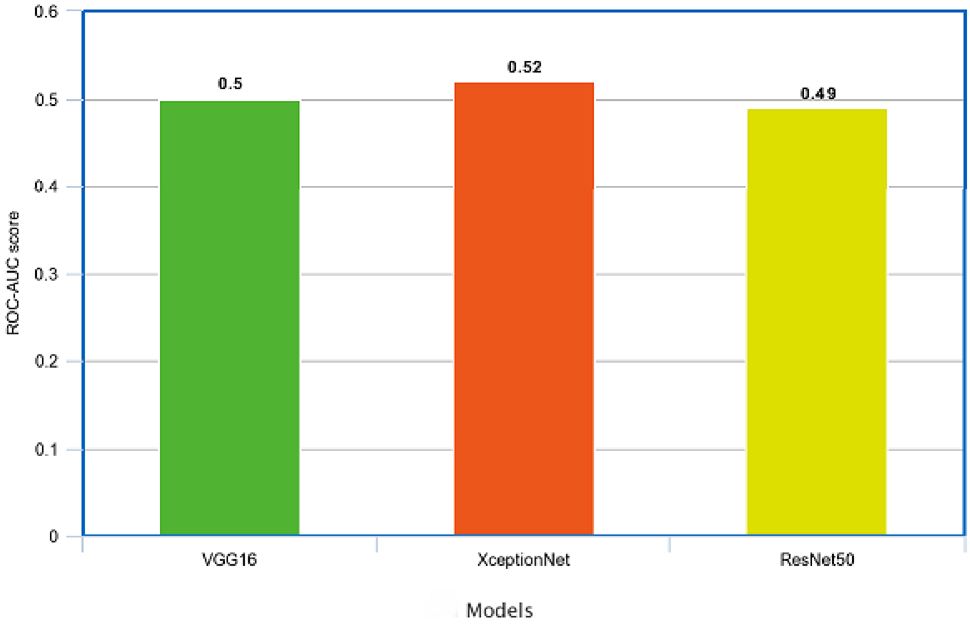

**Figure 7: ROC-AUC score**

DOI: 10.17148/IJARCCE.2022.116113

The VGG16 model classified 50%, XceptionNet classified 74% and ResNet50 classified 99% of test data correctly.

### Challenges

Open source face swapping software and applications result in a large number of Deepfake video clips that have a larger impact on social network. Identifying and screening such video clips content has become a challenging topic. The absence of quality Deepfake and actual video datasets which can be used for training datasets for research purposes is the major difficulty in the creation of a Deepfake detection algorithm and the datasets available will include difficult to train on normal machines. High-end machines with excellent computation power is required.

## CONCLUSION

Deepfakes are digitally edited clips of individuals doing or saying things they don't actually do. Visual inspection is insufficient to draw a conclusion on the validity, and available technologies to detect if the film has been tampered with are likewise insufficient. Because the visual appearance of Deepfakes will ultimately be so enormous that judging truthfulness only on the basis of visual validation would be challenging. The approach to solution here is to use of ingenious technologies to develop a methodology that can detect the deception in Deepfakes. As many deepFake technique can create images with finite resolutions and sizes, which must then be pixilated and tailored to suit the image that must be bartered with the original. The subsequent picture on the ROI creates unique artefacts in the deepfake image that may be managed to capture by analysing for variations the ROI and the remaining portion of the picture. The technique based on image quality metrics (IQM) and support vector machine (SVM) in CNN network may be employed for the classification of the video as real or false statistical correlations between the feature space for the extraction of the feature vectors related to the input. Future research might lead to the development of a system that might automatically detect deepfakes that employ an audio-visual method for detecting inconsistencies in facial movements and verbal in speech. Best method for deepfake detection of images and video will be identified.

## REFERENCES

[1] Wodajo, Deressa and Solomon Atnafu. “Deepfake Video Detection Using Convolutional Vision Transformer.” 2021.
[2] M. A. Younus and T. M. Hasan, "Effective and Fast DeepFake Detection Method Based on Haar Wavelet Transform," 2020 International Conference on Computer Science and Software Engineering (CSASE), 2020.
[3] L. Guarnera, O. Giudice and S. Battiato, "Fighting Deepfake by Exposing the Convolutional Traces on Images," in IEEE Access, vol. 8.
[4] Y. S. Malik, N. Sabahat and M. O. Moazzam, "Image Animations on Driving Videos with DeepFakes and Detecting DeepFakes Generated Animations," IEEE 23rd International Multitopic Conference (INMIC), 2020.
[5] A. Khodabakhsh and C. Busch, "A Generalizable Deepfake Detector based on Neural Conditional Distribution Modelling," International Conference of the Biometrics Special Interest Group (BIOSIG) 2020.
[6] L. Guarnera, O. Giudice and S. Battiato, "DeepFake Detection by Analyzing Convolutional Traces," IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020.
[7] Tariq, Shahroz & Lee, Sangyup & Woo, Simon. A Convolutional LSTM basedResidual Network for Deepfake Video Detection 2020.
[8] Durall López, Ricard & Keuper, Margret & Pfreundt, Franz-Josef & Keuper, Janis. 2019.
[9] Unmasking DeepFakes with simple Features Salpekar, Omkar. “DeepFake Image Detection.” 2020.
[10] Cozzolino, Davide, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. "Forensictransfer: Weakly-supervised domain adaptation for forgery detection." 2019.
[11] T. Karras, S. Laine and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019.
[12] Songsri-in, Kritaphat & Zafeiriou, Stefanos. Complement Face Forensic Detectionand Localization with Facial Landmarks 2019.
[13] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
[14] H. Zhang et al., "StackGAN++: Realistic Image Synthesis with Stacked GenerativeAdversarial Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019.