Library Progress International *Print version ISSN 0970 1052*
Vol.44 No.3, Jul-Dec 2024: P. 24541-24549 *Online version ISSN 2320 317X*

Original Article *Available online at www.bpasjournals.com*

IMPROVED FAKE IMAGE DETECTION AND CLASSIFICATION USING XCEPTION MODEL

[1]Debasish Samal, ,[2]Prateek Agrawal*, [3]Vishu Madaan

[1]School of Computer Applications, Lovely Professional University, Punjab, India
[23]School of Computer Science and Engineering, Lovely Professional University, Punjab, India
[2]Faculty of Engineering and Technology, Shree Guru Gobind Singh Tricentenary University, Gurugram, Haryana, India

How to cite this article: Debasish Samal, ,Prateek Agrawal, VishuMadaan(2024 IMPROVED FAKE IMAGE DETECTION AND CLASSIFICATION USING XCEPTION MODEL. *Library Progress International,* 44(3), 24541-24549

ABSTRACT
Recently, the creation of hyper-realistic fake faces using deep learning and machine learning methods has risen, resulting in a higher occurrence of fake images. These synthetic images, capable of convincingly imitating real human faces, present significant threats in various areas, including security and privacy, as well as media credibility and digital trust. This paper proposes a custom Xception deep learning model for fake image detection and classification. The base Xception model, pretrained on ImageNet, is included without its fully connected (i.e., top) layers to allow the model to utilize learned features with a focus on binary classification for our specific deepfake and real image dataset. The input layer accepts images of 128x128 pixels with three color channels providing compact input for reduced computational load. The proposed framework achieved the highest accuracy of 97.85% validation accuracy over training accuracy of 98.61% for both the deepfake and real image datasets. It also achieves high performance on various evaluation metrics with a Precision score of 0.95, a recall score of 0.79 and F1-Score of 0.86. The experimental results obtained demonstrated that the suggested method surpassed other leading fake image discriminators regarding performance, and it can assist cybersecurity experts in combating cybercrimes related to deepfakes.

KEYWORDS
CNN, Xception, Fake image, Real Image, Deepfake

1. Introduction

AI stands for Artificial Intelligence, which means teaching a machine to behave and reason like humans. It may also be used refer to any machine that exhibits human cognitive faculties such as perceiving, reasoning and learning [1]. One of key set marks of AI is to streamline and carry out the task that has maximum probability benefiting a given objective. Machine learning or ML is a subset of AI. To accommodate this self-learning, deep learning algorithms work on huge amounts of unstructured data such as text, images and videos. Machine Learning looks to mimic human learning, but by using various datasets and algorithms improves upon its accuracy over time [2, 3].

Implemented as a key component of the budding field, data science, ML forms an important part. Data-mining projects use statistical methods to develop algorithms that classify, predict, and present important information. These provide insights based on which decisions are made within applications and organizations with the goal of impacting important growth metrics [4]. The age of data science is beginning to spread with a rapid pace and the need of data scientists are going to be maximized as big data grow up. Leveraging ML should make it possible for obtaining information necessary draw answers to many of the key business questions. Machine learning is considered one of the branches of deep learning [5]. Deep learning is built on simpler concepts than the ones used in ML and constructs Artificial neural networks attempting to emulate human brain networks. The limitation of computers has been the historical reason for keeping the complexity of neural networks in check. This has allowed to image bigger and more complex Neural Networks (NNs) with computers able to see and learn faster than human beings what happened even in complicated events. One can steer and classify images, face recognition, language translation, audio recognition and separate real and fake faces with the help of deep learning. This means that it can solve the pattern recognition problems without any human intervention [6, 7].

Focusing on someone's face is the most basic part of knowing them. However, with the invisible hand of science improving robotic art, the challenges that arise from it become a real cause for concern. Several parameters can be combined to achieve a deep learning-based image swap [8]. Also, the more recent technology – the artificial intelligence deepfake – synthesizes the faces of two different people. Generators that utilize Generative Adversarial Networks (GAN) structures generate high quality images of deepfake that are more advanced than older models [9]. However, it is alarming in this case as the fact that the proliferation of deepfake content is likely to outpace the advancement of mobile devices and various social media platforms [10]. At first, these altered images could still be detected by the human eye because sparse blending of pixels typically creates sharp differences in tones of the face and the skin. However, time has passed, and deepfake technology has evolved, allowing for better combinations with natural images [11].

The deepfake methods require considerable amount of information to come up with believable images. However, deepfakes are not only a tremendous advancement of technology, but has its down sides. The sheer volume of videos and images that can be accessed on the internet makes it easy to create many deepfakes of public figures such as athletes, politicians and other celebrities [12]. In addition, it further claimed to be a technology that is used to ridicule and insult people. DeepFakes are the most harmful form of synthetic media. It allows the use of the celebrities' face images without any consent in creating political and fun entertainment which further asserts that creating deepfake content is possible with the help of convenience applications which are easily available to anyone. Deepfake technology is not just for famous individuals. Adolescent population who today are the ones to contribute alive to the caused cyberbullying deep fake content [10]. Deepfake technology advancement is increasing day by day in a dramatic way with the use of generative Ai models. The general process of image classification includes the steps of identifying a suitable classification rule, training the classifiers, image preparation processes, capturing the images and their features, and outlining a relevant evaluation metrics to measure accuracy.

Figure 1 shows an example of current deepfake threats arising concerns based on deep learning generative techniques.

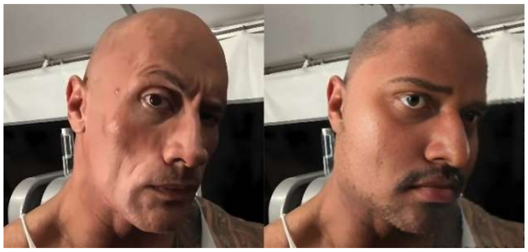

Figure 1. Original image (left) and Deepfake image (right) [21].

The key contributions of this paper are:

1.A deep learning based fully trained architecture for fake image detection and classification.

2.The proposed method is compared with various deep learning state of the art methods on well-known fake and real image datasets namely real and fake images [14] available online on kaggle.

3.A comprehensive study of the proposed model with results based on performance metrics such as accuracy, precision, recall and F1- Score.

The rest of the paper is organized as follows. Section 'Review of Literature' gives an overview of deepfakes and a recap of various studies and prior research on image classification. Section 'Methods and Materials' concentrates on materials and work methodologies. It provides a comprehensive description of the model employed. Section 'Results and Discussion' outlines the findings of the experiment conducted with the chosen dataset on proposed model and presents comparative results with existing deep learning models according to multiple performance evaluation metrics like Precision, Recall and F1-Scores. Section 'Conclusion and Future Works' states final thoughts and recommendations for future research endeavors.

2. Review of Literature

The advancement of technology has made life easier in many ways. However, there have been instances where technology has been misapplied, resulting in serious issues. A representation is digital image technology. There are many tools and software choices available that make it easier to modify any digital image. For example, anyone with a basic understanding of Photoshop can quickly and effortlessly create a fake image of another person [13]. A significant amount of recent research has been conducted on the use of these types of forgeries. Improvements in the fields of AI allow individuals to modify a raw image and utilize it for both beneficial and detrimental purposes, as these methods can yield remarkably realistic results. This familiarized us with the domain of deepfake images [14]. For instance, [15] employs deep learning as a technology for face recognition and can ascertain whether a profile picture is genuine or not, intending to discover a trustworthy way to differentiate between real and fake. In two datasets of images, this study comprised the detection of both real and false faces by means of a combined neural network-based deep learning approach: In document [5], it used a trained dataset of 9000 photos and selected the ResNet50 model as best fit. For the training set, accuracy = 99.18% but fails to achieve even the nearer validation accuracy. In [16], three benchmark datasets were used to test the proposed model which applied transfer learning methods from depth models like ResNet50 and VGG16 that have been trained beforehand. The data collected show that the proposed model has better performance than existing models. In [17], the research utilized improved datasets for distinguishing between real and fake faces to evaluate the leading contemporary face recognition classifiers, such as Custom convolutional neural network (CNN), VGG19, and DenseNet-121. They discovered that performance can be improved by using fewer computational resources thanks to data augmentation. The authors' initial results indicate that VGG19 surpasses all other models reviewed, achieving a peak accuracy of 95%. This research aims to offer a thorough analysis of the methods, structures, and mechanisms to develop ensemble-like multi-attention networks for identifying deep fake media.

The study in [18] seeks to tackle the challenge of distinguishing real images from fake ones by creating an algorithm capable of making that distinction. Their method aims to distinguish between genuine images and deep fakes where InceptionV3 delivered the highest results among the transfer learning models with a testing accuracy of 97.10%. When evaluating deepfake and authentic photos, the distinctive CNN model outperformed all other models used before. The primary objective of [19] was to create a dependable and precise technique for identifying deepfake images. The importance of this work is in achieving positive results using the CNN architecture. Advanced CNN architecture, Xception is used in this study to detect deep-fake images from large datasets. The results were precise and reliable. In terms of recall, for certain criteria such as F1 score, precision, the custom model employed in this study outperformed existing deep learning methods.

The work done in [20] gave a pipeline to classify and identify human faces from input visual samples. Several deep learning (DL)-based methods were used in the second stage to compute the deep features from the retrieved faces. The features were then used by a support vector machine (SVM), which is a type of classifier, to determine if the data was real or forged. By analyzing the published results from many of feature extractors compared, they discovered that DenseNet169 with SVM classifier achieved better results than rest. Table.1 provides the summary of the previously discussed studies.

3. Methods and Materials

Proposed Xception Model Architecture

The proposed model utilizes the Xception architecture first introduced by [22], which relies on depthwise separable convolutions, a technique that improves performance while ensuring computational efficiency. The base Xception model was altered by eliminating its top classification layers to make it suitable for binary classification of real and fake images. Our method utilizes a Global Average Pooling layer to minimize the spatial dimensions, succeeded by a dense layer featuring 1024 units and ReLU activation to better capture intricate patterns from the base model's output. The last layer is a sigmoid activation unit, which makes it appropriate for binary classification by producing a probability score ranging from 0 to 1.

ImageNet [23]'s pre-trained weights were used to initialize the model, enabling it to take use of expertise from a wide range of image domains. The Adam optimizer was used to fine-tune the network, setting the learning rate at 0.0001 to strike a balance between convergence speed and training stability.

Figure 2 presents the stages of fake image detection with the flow diagram of proposed Xception Model architecture. The model takes image input with input layer then apply Depthwise Separable Convolutions, the base Xception Model with average pooling layer and a dense layer to extract image features with the output layer at last for image classification.

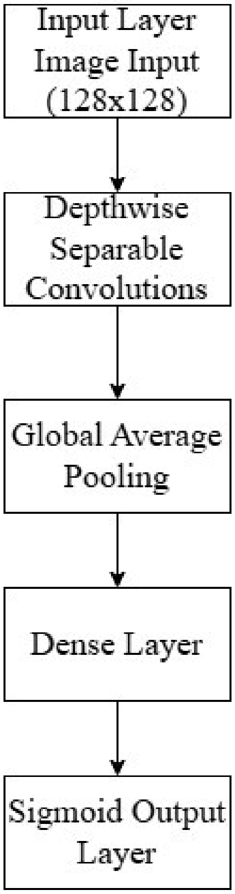

Figure 2. Flow diagram of Proposed Xception Model Architecture

Experimental Setup

A Windows 11 PC with an Intel Core i7 11th Generation CPU. An NVIDIA GEFORCE RTX 3060 GPU with 16 gigabytes of RAM was used for the research.

Dataset

An essential resource for creating and testing algorithms meant to identify modified photos is the Deepfake and real image collection on Kaggle [24]. The proposed model is trained on this dataset because to this dataset's balanced collection of real and synthetic faces. Notably, this dataset has been used in several research to obtain great accuracy in differentiating between real and false photos.

Dataset overview contains more than 1,90,000 authentic and fraudulent photos. Face-specific rich annotations are specifically included in this dataset to aid in the detection and segmentation of face forgeries.

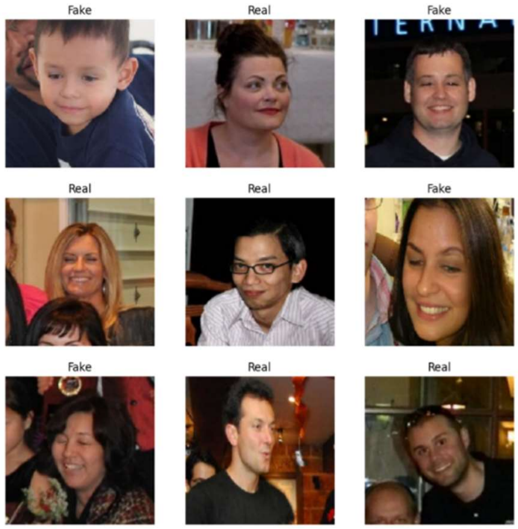

Figure 3. Labelled examples of pristine and fake images from Dataset [24]

Because of its extensive annotations, the considered real and fake images dataset is very valuable for study in both general artificial face identification and deepfake elimination. The dataset has a 256 x 256 pixel resolution and is balanced. Two classes of images—real and fake—consisting of a total number of 140002, 39428, and 10905 images each are used for training, testing, and validation.

Image Preprocessing

To fit the Xception model's input layer, images were shrunk to 128 by 128 pixels. All pixel values were rescaled with a factor of 1/255 to bring them into line with the range [0, 1]. To improve the model's resilience to changes in input conditions, additional augmentations were added to the training dataset, such as rotations, zoom adjustments, and flips in both the horizontal and vertical directions. By training on a wider variety of image changes, this augmentation technique improved the model's generalization.

Classification

Each image is categorized as either true or fraudulent in this binary approach to the classification challenge. A single probability score is produced by the suggested Xception-based model, where values above or equal to 0.5 are categorized as real and values below as fake. Initially, a threshold of 0.5 was used; however, depending on specific needs, this can be changed to prioritize recall or precision.

Evaluation Criteria

TensorFlow and Keras in Python were used to construct, test, and assess the model. We used metrics including accuracy, precision, recall, and F1-score to evaluate the model's performance. These metrics shed light on how well the model can distinguish between authentic and fraudulent photos, particularly when dealing with imbalances between false positives and false negatives.

The performance evaluation metrics are defined as below:

1. Accuracy: Accuracy is defined as the proportion of correctly classified images. (TP + TN) / (TP + TN+FP+FN).
2. Precision is defined as the ratio of successfully recognized positively classified categories to positively predicted categories. TP/ (TP + FP).
3. Recall: The percentage of correctly classified subjects among all favorably classified subjects is known as the recall rate. TP / (FN + TP).
4. F1 score: The F1 score is commonly used to enable the simultaneous measurement of precision and recall. The arithmetic mean is substituted with the harmonic mean. They penalize extreme values more as a result. 2*(recall*precision)/ (precision + recall)

The precision metric shows the proportion of genuine images that were anticipated.

The percentage of real, authentic photos that are correctly identified is known as recall.

In situations where there is a class imbalance, the F1-score is helpful for assessing the model since it strikes a balance between precision and recall.The distribution of true positives (TP), true negatives (TN), false positives

(FP), and false negatives (FN) was also visualized using a confusion matrix.

4. RESULTS & DISCUSSION

The following outcomes were attained after training the model for 20 epochs with a batch size of 64 images.

Accuracy of Training and Validation: The model's training accuracy was above 98%, while its validation accuracy was over 97% as shown in figure 4.

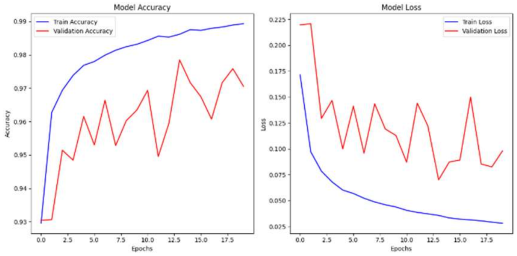

Figure 4. Train and Valid accuracy of Xception Model

Test Accuracy: The model demonstrated effective generalization and binary image classifications with a final accuracy of 88% on the unseen test dataset.

Confusion Matrix: The model accurately detected 96% of fake photos and 80% of actual images as seen in Figure 5, according to the confusion matrix analysis. This revealed a significant bias in favour of the fake class, which might be lessened with additional fine-tuning.

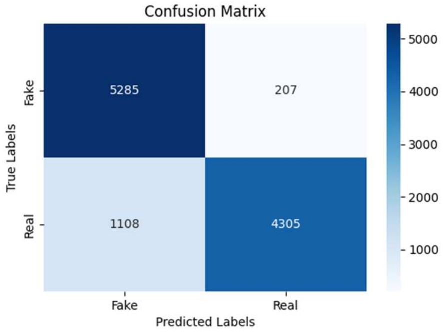

Figure 5. Confusion Matrix

Classification Report: The following were the Precision, Recall and F1-scores for the proposed model:
Precision: 0.9541
Recall: 0.7953
F1-Score: 0.8675

The model is dependable for binary classification with little overfitting, as evidenced by the balanced F1-score (0.8675) across classes. Recall (0.7953) might be enhanced by modifying the classification threshold, while precision (0.9541) shows how well the model reduces erroneous positives.

Using the ROC curve in Figure 6 and the model's Precision-Recall curve score derived from projected probabilities in Figure 7, the proposed study assesses the model's capacity to discriminate between authentic and fraudulent photos. This analysis is necessary to assess how well the model performs in differentiating between the two classes at different threshold-levels by using the values of True Positive Rate (TPR) and False Positive Rate (FPR).

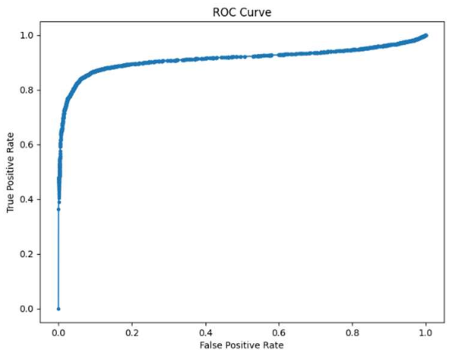

Figure 6. ROC Curve

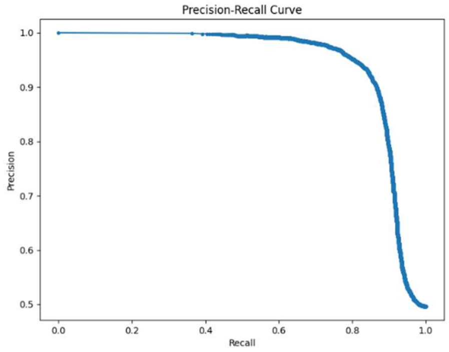

Figure 7. Precision-Recall Curve

The upper-left corner of the ROC curve, which represents the highest Specificity and Sensitivity, is desirable. The ROC curve shows a strong model, as evidenced by the 0.9134 AUC score as well as the Precision-Recall AUC score of 0.9421. The curve's proximity to the upper-left corner suggests that the model is performing exceptionally well in terms of reducing false negatives and increasing genuine positives.

These findings indicate that the model successfully differentiates between authentic and fake images, exhibiting marginally greater precision than recall, which implies a necessity to further refine recall for the classification of real images.

Real world applications and prediction using proposed Xception model

As demonstrated in Figures 8a and 8b, binary image classification tests on unseen individual "Real" and "Fake" images are used to assess the effectiveness of the proposed model. The predicted labels from the model are displayed on the photos. The model shows its usefulness in real-world scenarios by offering accurate and trustworthy image classification.

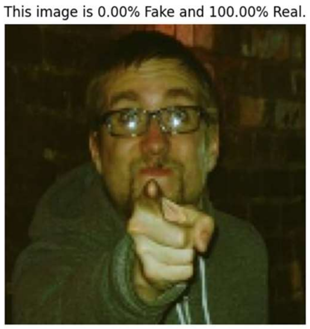

Figure 8(a). Real image test accuracy of Xception Model

This image is 100.00% Fake and 0.00% Real.

Figure 8(b). Fake image test accuracy of Xception Model

Comparative Analysis with Existing Studies

This paper compares the proposed model's accuracy with the recent state-of-the-art method by [26] because of the similarity of the dataset used to carry out experiments. Employing transformational and deep learning methodologies their paper proposed convolutional neural network (CNN) models based on deep transfer learning methodologies. The authors have trained and tested six different CNN architectures on the same kaggle dataset [24] this paper used.

CNN Model	Accuracy(in %)	Precision	Recall	F1-Score
MobileNet	82.78	0.83	0.83	0.83
ResNet50	83.33	0.83	0.83	0.83
Xception	84.07	0.85	0.83	0.84
InceptionV3	85.00	0.87	0.84	0.84
DenseNet201	86.58	0.87	0.86	0.87
Proposed Model	88	0.95	0.79	0.86

Table 1. Comparative analysis of Proposed Xception Model with the works of [26]

In Table 1. The proposed model in this paper outperforms the methods used the preceding work with a significant 88% test accuracy over both the deepfake and real picture datasets whereas their recommended framework DenseNet201 achieved an accuracy of 86.85%, but MobileNet produced a lesser accuracy of 82.78%.

5. CONCLUSION AND FUTURE WORKS

This research introduced an Xception-based deep learning model for the classification of real versus fake images, attaining the highest validation accuracy of 97.85% over 1,90,000 real and fake images. A detailed examination of precision, recall, and F1-score was necessary to comprehend the subtleties of classification, although accuracy alone at 88% provided a broad indication of overall performance. Although recall for real photos showed potential for improvement, the model showed great precision, correctly recognizing most fraudulent images. To increase recollection, future research might concentrate on threshold tuning or using ensemble approaches. Furthermore, testing with different deepfake datasets or with different architectures, such as EfficientNet, may shed light on the resilience and scalability of the model.

References

[1] Watzlaf V, Alkarwi Z, Meyers S, Sheridan P. Physicians' outlook on ICD-10-CM/PCS and its effect on their practice. Perspectives in Health Information Management. 2015; 12(Winter): 1b. PMID: 26807074 PMCID: PMC4700867

[2] Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.

[3] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553): 436–444.

[4] Domingos P. A few useful things to know about machine learning. Communications of the ACM. 2012; 55(10): 78-87

[5] Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks. 2015; 61: 85-117

[6] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Computer Vision –

ECCV 2014. Springer; 2014. p. 818-833

[7] Chollet F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017; p. 1251-1258

[8] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019; p. 4401-4410

[9] Wang T, Liu Z, Zhu J-Y, Liu H-Y. Neural face editing with GAN-based approaches. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020; p. 6183-6192

[10] Chesney R, Citron D. Deepfakes and the new disinformation war. Foreign Affairs. 2019; 98(3): 147-155.

[11] Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019; p. 1-11

[12] Mirsky Y, Lee W. The creation and detection of deepfakes: A survey. ACM Computing Surveys. 2021; 54(1): 1-41

[13] Farid H. Digital forensics. Scientific American. 2008; 298(6): 66-71.

[14] Korshunov P, Marcel S. Deepfakes: A new threat to face recognition? Assessment and detection. IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS). 2018; p. 1-6.

[15] Agarwal S, Farid H, Gu Y, He M, Nagano K, Li H. Protecting world leaders against deep fakes. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019; p. 38-45.

[16] Guera D, Delp EJ. Deepfake video detection using recurrent neural networks. IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). 2018; p. 1-6.

[17] Zhang Z, Lin L, Zhang Y. Real and fake face detection based on convolutional neural network with improved data augmentation. Journal of Electronic Imaging. 2019; 28(3): 1-11.

[18] Sabir E, Cheng P, Jaiswal A, AbdAlmageed W, Masi I, Natarajan P. Recurrent convolutional strategies for face manipulation detection in videos. arXiv preprint arXiv:1905.00582. 2019.

[19] Afchar D, Nozick V, Yamagishi J, Echizen I. MesoNet: A compact facial video forgery detection network. IEEE International Workshop on Information Forensics and Security (WIFS). 2018; p. 1-7.

[20] Dang H, Liu F, Stehouwer J, Liu X, Jain AK. On the detection of digital face manipulation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020; p. 5781-5790.

[21] Faceswap. Available online: https://faceswap.dev

[22] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1800-1807, doi: 10.1109/CVPR.2017.195.

[23] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: CVPR09. IEEE; 2009. Available from: https://www.image-net.org/

[24] Karki M. Deepfake and Real Images [Data set]. Kaggle. 2023. Available from:https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images

[25] I. E. Naqa and M. J. Murphy, “What is machine learning?”, in machine learning in radiation oncology,” machine learning in radiation oncology, Cham: Springer, pp. 311, 2015.

[26] Atwan J, Wedyan M, Albashish D, Aljaafrah E, Alturki R, Alshawi B. Using Deep Learning to Recognize Fake Faces. Int J Adv Comput Sci Appl. 2024; 15(1):Article 113.Available from: http://dx.doi.org/10.14569/IJACSA.2024.0150113

[27] Atwan J, Wedyan M, Albashish D, Aljaafrah E, Alturki R, Alshawi B. Using Deep Learning to Recognize Fake Faces. Int J Adv Comput Sci Appl. 2024; 15(1):Article 113.Available from: http://dx.doi.org/10.14569/IJACSA.2024.0150113