UDC 004.932 **doi:** https://doi.org/10.20998/2522-9052.2024.2.10

Velusamy Rajakumareswaran[1], Surendran Raguvaran[2], Venkatachalam Chandrasekar[3], Sugavanam Rajkumar[4], Vijayakumar Arun[5]

[1] Erode Sengunthar Engineering College, Tamil Nadu, Thuduppathi, India
[2] SRM Institute of Science and Technology, SRM Nagar, Kattankulathur 603203, Chennai, TN, India
[3] Jain University, Bangalore, India
[4] Sona College of Technology, Salem, India
[5] School of Engineering, Mohan Babu University, Tirupati, India

# DEEPFAKE DETECTION USING TRANSFER LEARNING-BASED XCEPTION MODEL

**Abstract. Justification of the purpose of the research.** In recent times, several approaches for face manipulation in videos have been extensively applied and availed to the public which makes editing faces in video easy for everyone effortlessly with realistic efforts. While beneficial in various domains, these methods could significantly harm society if employed to spread misinformation. So, it is also vital to properly detect whether a face has been distorted in a video series. To detect this deepfake, convolutional neural networks can be used in past works. However, it needs a greater number of parameters and more computations. So, to overcome these limitations and to accurately detect deepfakes in videos, a transfer learning-based model named the Improved Xception model is suggested. **Obtained results.** This model is trained using extracted facial landmark features with robust training. Moreover, the improved Xception model's detection accuracy is evaluated alongside ResNet and Inception, considering model loss, accuracy, ROC, training time, and the Precision-Recall curve. The outcomes confirm the success of the proposed model, which employs transfer learning techniques to identify fraudulent videos. Furthermore, the method demonstrates a noteworthy 5% increase in efficiency compared to current systems.

**Keywords:** DeepFake; editing faces; Improved Xception model; Precision-Recall.

## Introduction

Since the advent of photography and videography, image and video modification has been practiced. In image or video transformation and animations, artifact systems for photos and videos, as well as strong editing tools, played a critical role. Furthermore, because these visual distortions are immediately visible and identifiable with human-focused vision, there is little scope for them to be created. With fast improvements in technology and the utilization of image processing methods, the creation of face warping has started. In this, targeted faces look like a real image, appeared on a big scale in latest months starting with animations, commercial components for marketing and entertainment objectives. Fake multimedia became a major issue, particularly when DeepFakes began circulating on social networking sites, falsely accusing well-known people [1].

Attacking and changing digital data from scratch is now feasible due to the quick development of artificial intelligence and techniques like Generative Adversarial Networks (GAN) [2]. The Deepfake pictures phenomena, or generally Deepfakes, emerged as a consequence of these technologies' ability to produce remarkably accurate outcomes. It has lately been a major source of public issues.

The term "DeepFake" denotes a deep learning technology capable of generating counterfeit videos by swapping the facial features of one individual with those of another [3].

In terms of the amount of alteration, the most common ways for creating false face content may be divided into four categories [4-6]:

i) identity swapping,
ii) attribute modification,
iii) complete face synthesis,
iv) expression switch.

Face swapping or face synthesis, in combination with voice dubbing, speech conversion, or speech synthesising is used in most deepfake production techniques [8]. Framework or soft biometrics within visual information, and spoof identification for audio, were the focus of detection algorithms [7–13]. Digital forensic tools are excellent at identifying targeted forgeries, but they fall short when it comes to out-of-sample or unanticipated fabrication processes.

The detection of Visual DeepFakes is far more difficult than that of Audio DeepFakes. The majority of present DeepFakes remain in the form of vision-based and may be readily altered to have a realistic appearance [14]. DeepFakes was mostly used to impersonate actors and politicians in order to make amusing things. As accuracy improves, it has extended to the point where false content is circulated, causing a commotion in society and disrupting peace and tranquilly. The shifting away from ideas while on an election or indirectly accusing an individual in a crime [15] has recently been circulating in the media. Through synthesising a face which is re-enacted through motions, activities of other person, these modifications of face properties of famous politician, CEO or any other person become believable in a quiet way that any human eye cannot differentiate a real or false movie.

With latest advances in deep learning, it is now fairly simple to create a model from a person's existing images and videos [1]. People use several real-time face recognition techniques to identify face on these photographs. The foundation of DeepFake lies in a deep neural network trained on facial data. After the appropriate post-processing stage, this network transfers the source's facial movements to the target, providing a high amount of realism. However, the models still lag in terms of accurate fake video detection. This issue needs to be avoided by strengthening countermeasures.

The main contributions of this study are outlined as follows:

- This work introduces a novel transfer learning approach called the Improved Xception Model for detecting fake videos. To our knowledge, no previous studies have employed the Xception model for this purpose.
- To enhance the model's accuracy, Xception undergoes training using network weights, followed by fine-tuning of the pre-trained network weights.
- The suggested approach achieves better accuracy, ROC values, and minimal loss compared to existing cutting-edge prediction algorithms.

The following is the outline of this work: The one section examines similar work in the field of fake video detection. In Section 2, the suggested methodology is applied to identify fraudulent videos. The outcome and discussion in Section 3 demonstrate how effective the suggested effort was.

## 1. Related works

Hashmi et al. [1] presented the Conv-LSTM framework that utilizes face landmarks as well as convolutional features to identify visual fraud in movies and photos effectively. After extraction of 512 face landmarks, they were compared. The subject's location in the video seems to be the most significant constraint. Because the prediction accuracy is poor in side angles, zoomed angles are preferable over side angles. In the point of computing complexity, the system is considered to be heavyweight.

Guarnera et al. [2] suggested a novel method for extracting a Deepfake fingerprint from photos. The approach is built around an Expectation-Maximization algorithm which was given training for recognizing and retrieving a fingerprint. While generating images, it reflects the Convolutional Traces leftover through GANs. The CT shows excellent discriminative capability, outperforming best in the Deepfake detection test while demonstrating resistant to various assaults.

Tolosana et al. [3] give a wide analysis of facial image modification strategies incorporating DeepFake approaches, in addition to ways to recognize such changes. Quatern different systems of face modification are conferred in aspect. They give a specific focus on DeepFakes, noting its advances as well as problems in identifying fakes between all elements mentioned in this study.

A novel fusion deep learning technique for incorrect update sorting that combines convolutional and recurrent neural networks was presented by Nasir et al. [16]. This model produced high detection accuracy that was significantly better than non-hybrid baseline techniques, and it was successfully validated on two datasets.

Jung et al. [17] introduced a novel technique for noticing Deepfakes created by the Generative Adversarial Network (GANs) model using a DeepVision method to examine a substantial change in blinking patterns. But flashing is also linked to mental disorders as well as dopamine action, which is an investigated problem. Patients with cognitive sicknesses or problems in bravery transfer ways will not be capable of using the veracity check.

With the help of contrastive loss, Hsu et al. [18] present a deep-learning framework designed to detect fraudulent images. Initially, a range of advanced GANs is waged to produce sets of counterfeit and genuine image pairs. Following this, the simplified DenseNet undergoes adaptation into a dual-stream network structure, accepting paired data as input. Finally, a classification layer is incorporated into the improved Xception structure to ascertain whether the input image appears genuine or counterfeit.

The first openly accessible collection of Deepfake films created from videos in the VidTIMIT database was provided by Korshunov and Marcel [19]. Based on GANs, false videos are generated using open-source software. Researchers discovered that Deepfake video detection techniques are required since cutting-edge face identification structures centered on VGG and Facenet neural networks are still susceptible to fake videos.

In order to detect audio spoofing and visual deepfakes, Chintha et al. [20] introduced an efficient digital forensic method. Bidirectional recurrent structures, entropy-based cost functions, and convolutional latent representations are combined in the suggested methods. Sensible semantic information is extracted from the recordings by carefully creating latent representations, including video and audio. Additionally, they pinpoint spatial and temporal anomalies in deepfake renditions by feeding them into a recurrent structure.

Verdoliva et al. [21] investigated techniques for fake video recognition. The focus will be on deepfakes produced by deep learning-based methods and on innovative data-based forensic techniques to counter them. The findings assist in highlighting the shortcomings of current forensic methods as well as the most urgent issues, new challenges, and areas for further investigation.

Caldelli et al. [22] proposed method introduces a technique for discerning both counterfeit and authentic videos. It uses CNNs that have been trained to identify potential motion differences in the temporal structure of a video sequence by using optical flow fields. The outcomes compare favorably with state-of-the-art methods that frequently rely exclusively on individual video frames.

Some authors consider examining the internal GAN pipeline for distinguishing between actual and false images by detecting distinct artifacts. The authors speculated in [23] that the color difference between actual camera images and fake synthesized images seems to be significant. They suggested a color-feature-based detection method to classify fake images.

As the neuron activity design between layers might record high minor aspects which is crucial for the modified face identification technique, Wang et al. [24] hypothesized about observing neuron behaviour might be useful in identifying false faces. The suggested method used deep facial recognition framework for extraction of structures nerve cell exposure characteristics for real and fake faces, and thereafter classification is carried out using trained SVM.

The use of steganalysis in fake detection systems was also studied. A pixel co-occurrence matrix and convolutional neural network (CNN) based detection approach was suggested by Nataraj et al. [25] in this research. First, a database containing a variety of objects and Cycle-GAN-generated situations was used to evaluate their proposed strategy.

## 2. Methodology

This section describes first about dataset used for deepfake detection from video frame. Next, the traditional CNN and its limitations are discussed. Following this, Xception model used for fake video detection is described in detailed manner.

### 2.1. Dataset

The data utilized in this work comes from FaceForensics++ [26]. The FaceForensics++ dataset is made up of 1000 original video sequences that have been altered through the use of four automated face alteration techniques: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. With the help of data gathered from 977 YouTube videos, which all include searchable faces that are mostly frontal and occlusion-free, realistic frauds can be produced by automated tampering procedures. Because binary masks are provided, the data can be used for segmentation and image and video classification. To help create and improve new data, a total of 1000 Deepfakes models are also offered.

### 2.2. Training dataset and Testing dataset

As per a typical neural network guideline, the dataset is divided into two groups: 80% of the data are train data, and the remaining 20% are test data. Models are trained on the training dataset, and predictions are run on the test dataset. Furthermore, from the training dataset, 90% of the data are used as training data and 10% are used as validation data in order to calculate the model's performance, model loss, and ROC. During training and testing the model with video frame, every frame in the video is converted to $299 \times 299$ pixels for fitting the model. The landmark features extracted from the video frames can be used for training the model. Additionally, Adam optimization is used to train the transfer learning-based model at various learning rates in order to improve accuracy and prevent overfitting of the suggested job.

### 2.3. Traditional Convolutional Neural Network

Convolutional neural networks is a type of artificial neural network used in different fields like radiology. CNN learn spatial hierarchies of data autonomously and flexibly with the help of back propagation and different layers of convolution, pooling and fully connected layer. Fig. 1 shows the traditional Convolutional Neural Network. In this, we have an input image of 3x10x10 size. To perform convolution operation, 3x3x3 kernel size is used. This kernel slides over input image to produce desired output.

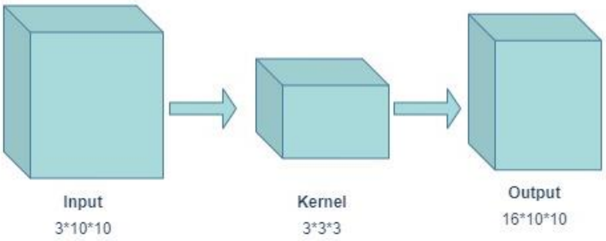

**Fig. 1.** Traditional Convolutional Neural Network

This convolution operation needs 432 parameters and 43200 computations to produce output image of $6 \times 10 \times 10$. This results in complexity. To reduce this complexity, Depthwise Separable Convolution can be used. This is illustrated in Fig. 2. The above mentioned procedure is separated by depthwise separable convolution into a depthwise convolution and a pointwise convolution. $3 \times 3 \times 1$ convolution is performed in depthwise convolution. Then, following this, pointwise convolution with kernel size of $1 \times 1 \times 3$ is performed. For depthwise convolution, 27 parameters and 2700 computations are needed. For pointwise convolution, 48 parameters and 4800 computations are needed. So, totally we need 75 parameters and 7500 number of computations for producing output. When compared to traditional CNN, Depthwise Separable Convolution performs convolution with less operations, so it reduces network's computational cost. Through this, it is clear that deeper models are effective when compared to wider ones. This Depthwise Separable Convolution is used in Xception model for better performance of the model and it is shown in Fig. 3.

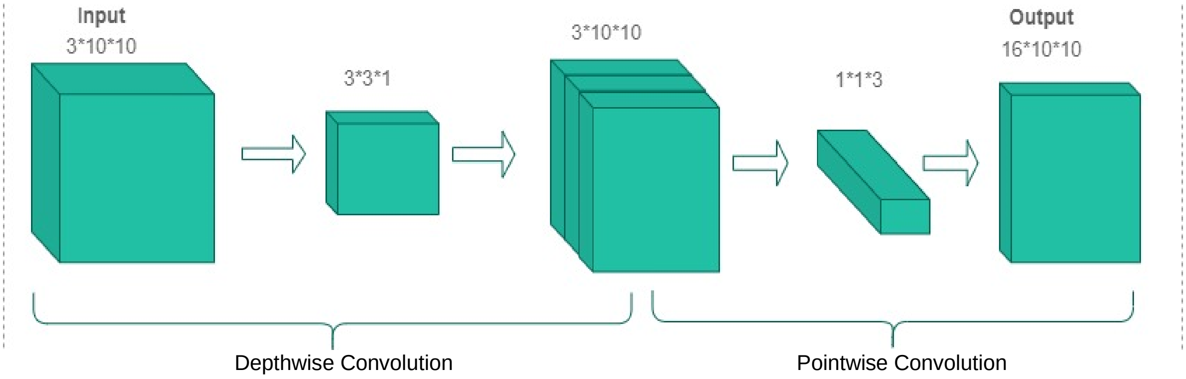

**Fig. 2.** Depthwise Separable Convolution

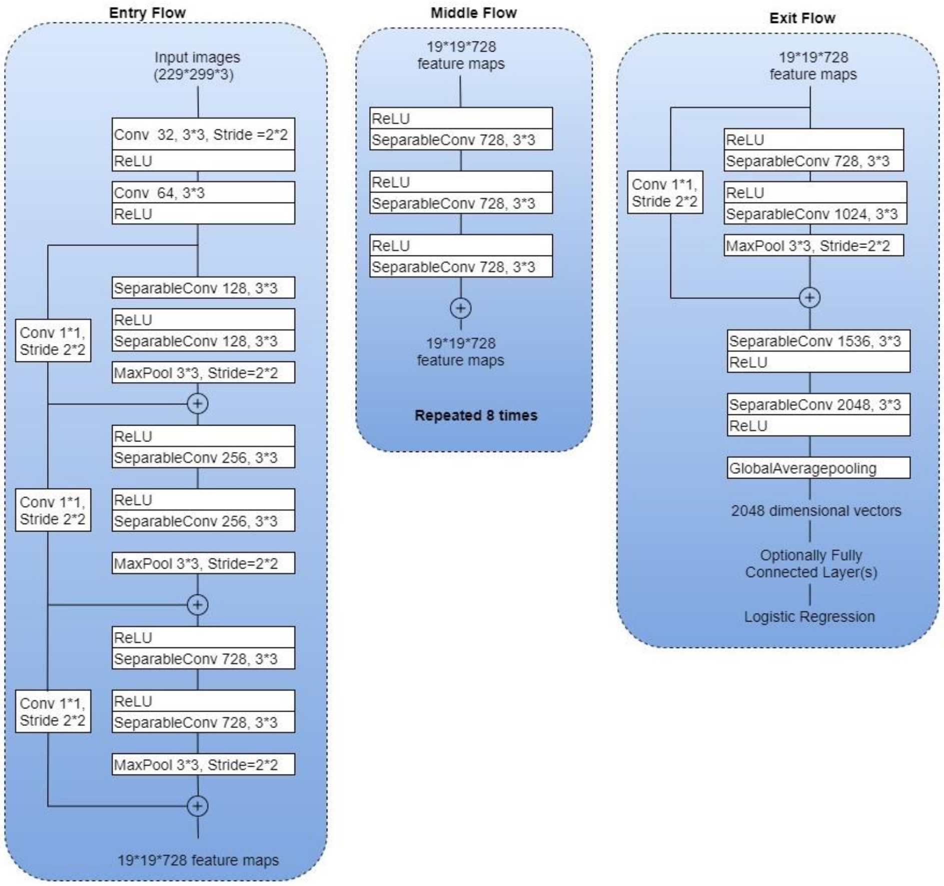

**Fig. 3.** Architecture of Xception Model

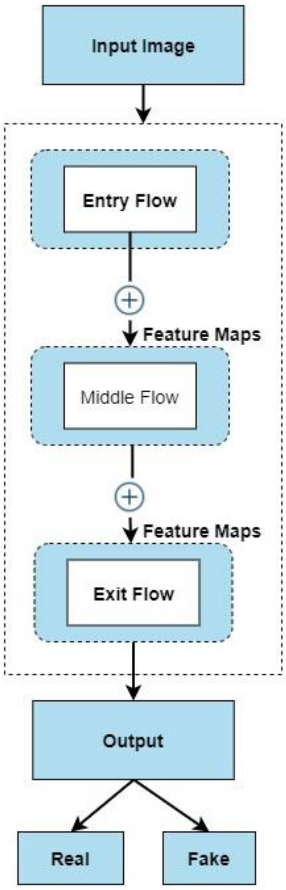

**Fig. 4.** Overview of proposed Improved Xception Model

## 2.4. Transfer Learning based Fake video Detection

Transfer learning is the process of applying previously acquired subject-matter knowledge to address disparate but related domain issues. Transferring current knowledge to address object-domain learning challenges with fewer training sample images is the goal. Transfer learning lowers the amount of data needed for model training in the supervised training mode. The solution not only resolves the issue of inadequate data in the target field causing difficulty in fitting the model, but it also expedites the model's convergence and enhances its accuracy. The suggested work employs transfer learning for fraudulent video detection because of these benefits.

**2.4.1. Improved Xception Model.** This section discusses the proposed work in detail. First, the features are extracted from the video frame and then the classification of fake and real video are carried out. In this work, deepfake is detected using a proposed model named the Improved Xception model. The proposed Xception model consists of three flows:

(i) The Entry flow,

(ii) The Middle flow;

(iii) The Exit flows.

The landmark features from video frame are extracted first. Then fake video is detected in the last layer of the model. Along with the weights of the ImageNet, Xception model is utilized. The original Xception model [30] is modified to achieve higher accuracy. In the standard Xception model, the final layers are logistic layer and pooling layer. These 2 layers are replaced by Global Average Pooling layer (GAP) and following this, dropout layer is added. Lastly, for prediction, a logistic layer is added at network's end. Fig. 4 shows an overview of proposed Improved Xception Model.

**2.4.2. Landmark feature extraction and detection of fake video.** In this section, facial landmark features representing the frames of the video is extracted. Then, the deep visual features of the video frame are extracted. The base layers in original Xception model is used for feature extraction process. This detects the region of face in frame I. The face shape can be represented with the help of face parametric function z(). For determining the initial 2D landmark locations, the landmark vertices z() are used on the frame. Dense vertices can be utilized for determining depth data in the face. Also, it is helpful for generating better performance in the facial movements. Then, both information is concatenated. The invisible landmarks are set to zero. A total of 512 landmark vertices are used in the model. The feature map is extracted using transfer learning based pretrained model which improves accuracy of the model. From the training set, the network learn about detecting the face in every frame and aligning landmarks in 2D space with help of landmarks. The 512 landmarks define how each portion of the face changes. This is compared to the movement observed by landmarks in each component of the faked one. This work focuses on the motions of ideal components like as the eyes, nose, lips and eyebrows.

**2.4.2.1. Convolutional layer.** The convolution layer in CNN is the most important layer utilized for feature extraction. This convolution layer uses a convolutional technique to create feature maps over the input image. Different features can be retrieved over this layer by using numerous kernels. Eyes, nose, lips, and eyebrows are the main focus of extracted visual traits, both deep and raw. One definition of the output of the L$^{th}$ convolutional layer is [29],

$$C_{OP} = ReLU(I * W_L + b_L), \tag{1}$$

where $W_L$and $b_L$ indicates the weight and bias of the L$^{th}$ layer; $I$ indicates the input picture; and $C_{OP}$ indicates the output of the Nth convolution layer. ReLU activation function is applied to the result after convolution process. The Rectified Linear Unit (ReLU) is used to stimulate neurons following this procedure. In neural networks, this ReLU is crucial since it converts input at each network node into output.

With a higher learning rate, it enables the neural network to learn nonlinear dependencies and mitigate vanishing gradient. Its rate of convergence is also faster. In the output layer of networks, linear activation functions are typically utilized for predictions. We can express the ReLU function over input I as follows:

$$ReLU\ (I) = \max\ (0, I). \tag{2}$$

Adam optimization, the optimization technique employed in this work, aids in weight updating through the utilization of training data. This Adam optimization makes use of the advantages of Root Mean Square Propagation (RMSProp) and Adaptive Gradient (AdaGrad) techniques. For every parameter θ, it calculates the unique adaptive learning rate. Similar to momentum, the Adam optimizer uses the exponentially decaying average of previous gradients, $m_i$ [29]:

$$V_i = \beta_1 V_{i-1} + (1 - \beta_1) g^2 i; \tag{3}$$

$$m_i = \beta_2 m_{i-1} + (1 - \beta_2) g_i. \tag{4}$$

In Eq-n (3) and (4), $V_i$ indicates the variance and $m_i$ denotes the mean values. The following can be represented by the Adam updated rule using these variables:

$$\phi_{i+1} = \theta_i \frac{\mu}{\sqrt{v_i + \varepsilon}}. \tag{5}$$

Weights are changed and the appropriate learning rate is selected for accurate prediction based on this optimization technique.

**2.4.2.2. Max-pooling layers** This pooling layer will take the output from the preceding convolution layer as its input. In general, pooling can be divided into two types: maximum pooling and average pooling. This max pooling layer can be used to suppress noise. It can do de-noising and dimensionality reduction in addition to removing the noisy activations. In contrast, dimensionality reduction is actually carried out using average pooling as a noise suppression technique. As a result, max pooling outperforms average pooling. Downsampling (DS) is done on the feature map in this max pooling layer using the output that was obtained from the preceding convolutional layer.

The output of the pooling layer [29] is denoted as,

$$P_l = \underset{l \in S}{\text{Max}}\, C_{OP}, \tag{6}$$

where $P_l$ indicates the pooled feature map. S indicates the pooling region in the feature map.

**2.4.2.3. Global Average Pooling Layer and dropout Layer.** Conventional convolutional neural networks reduced dimensionality and performed non-linear transformation on high-dimensional feature data extracted from the convolutional layer using the entire connection layer. The resulting data was then sent into the classification layer for classification. A relationship between the convolutional structure and the conventional neural network classifier was created by this structure.

On the other hand, over-fitting and parameter redundancy brought on by the entire connection layer would reduce the network's capacity for generalization and increase the amount of time required for model training. In order to reduce the number of parameters and computational load while simultaneously enhancing the model's performance, we decided to replace the full connection layer with the global average pooling layer.

A regularization technique called global average pooling produced a feature vector by averaging every pixel in each feature map that the convolutional layer produced. The global average pooling layer preserved the convolutional structure and improved the correspondence between the mapping features and the final category when compared to the full connection layer.

Moreover, there are no parameters for the global average pooling layer that need optimization. Consequently, dropout layer with dropout of 0.5 reduces model complexity and prevents over-fitting. The feature vector from global average pooling output $GAP_O$ can be represented as,

$$GAP_O = \overrightarrow{P_l}, \tag{7}$$

where $\overrightarrow{P_l}$ denotes the vector of feature map.

**2.4.2.4. Logistic layer.** The final layer of the Xception model is called a logistic layer, and its purpose is to forecast the likelihood of bogus video in the dataset^

$$L_{Op} = \begin{cases} 0, if\ Op\ is\ real; \\ 1, if\ Op\ is\ fake. \end{cases} \tag{8}$$

The output will be 0 if the video is real and the output will be 1 if the output is fake.

**2.4.2.5. Hardware Requirement.** Using the Xception model via transfer learning, the method for identifying DeepFakes is implemented on the NVIDIA DGX-1 system. Eight V100 GPU accelerators, each with 32 GB of RAM, are included with this platform. Python is the primary programming language used by the system, and Tensorflow in Python is used for its implementation.

## 3. Result and discussion

Xception models have been applied recently to address a variety of issues in several disciplines, including detection and categorization. In this work, a model for detecting bogus videos based on transfer learning is provided. By contrasting it with current state-of-the-art techniques, the suggested method's effectiveness is examined. For efficiency analysis, the phrases listed below are utilized:

- ROC;
- Model loss;

- Precision-Recall curve;
- Training time.

### 3.1. Receiver Operator Characteristic (ROC)

The Receiver Operator Characteristic (ROC) is a crucial metric for problems involving classification and detection. This curve helps distinguish between signal and noise in data and can be used to draw a graph between the True Positive Rate (TPR) and the False Positive Rate (FPR). TPR, also known as sensitivity, expresses the degree to which the negative class was precisely estimated. We can see how much of the negative class the model incorrectly estimated by looking at the FPR or specificity. The ROC curve is summarized by the Area Under the Curve (AUC), which also serves as a measure of a model's ability to distinguish across groups. The model's output in differentiating between positive and negative groups is stronger in this study, as seen by the high AUC value of 0.986.

In Fig. 5, the enhanced Xception model exhibits a notably high true positive rate in comparison to alternative approaches such as Resnet and Inception.

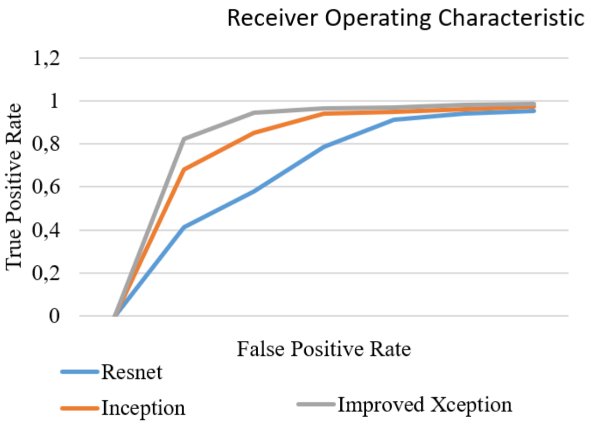

**Fig. 5**. ROC Curve for Resnet, Inception and Xception

This implies that the suggested approach successfully discerns between authentic and fraudulent videos. The suggested strategy outperforms previous algorithms in reliably identifying bogus frames in videos, as indicated by the ROC curve, which is closer to the top-left corner. The proficiency of the enhanced Xception model in correctly detecting videos is attributed to its robust training, involving a substantial amount of data.

### 3.2. Precision-Recall curve

The precision-recall curve is used to assess model performance, just like the ROC curve. When there is a significant imbalance in the courses, it is frequently utilized. Instead of showing a single value, precision-recall curves give a graphical depiction of a classifier's performance across a range of thresholds. In addition to helping us choose the ideal threshold for a given situation, a precision-recall curve makes it easier to see how threshold selection impacts classifier performance. The accuracy can be defined as the ratio of correctly labeled positive samples to the total number of correctly classed (or mistakenly classified) positive samples. The accuracy metric evaluates how well the model can appropriately interpret a result as positive. Values for precision range from 0 to 1.

The precision [29] can be denoted as,

$$Precision = \frac{True_{Positive}}{True_{Positive} + False_{Positive}}.$$

By dividing the number of actual positive outcomes by the total number of samples, the recall is used to calculate the number of accurate positive predictions.

The recall is represented by [29]:

$$Recall = \frac{True_{Positive}}{True_{Positive} + False_{Negative}}.$$

Fig. 6 displays the precision-Recall curves for the ResNet, Inception, and Improved Xception models.

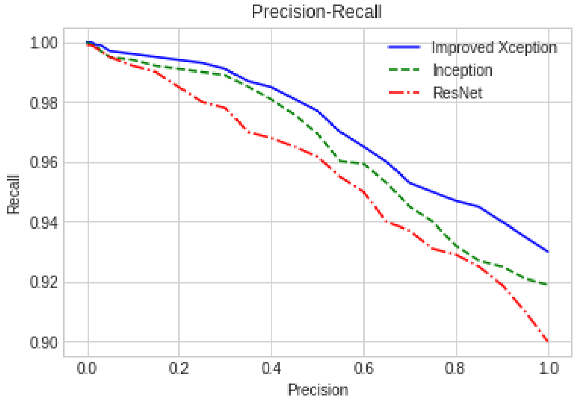

**Fig. 6.** Comparison of Precision-Recall curve for different methods

Precision is indicated by the X axis, and recall is indicated by the Y axis. The classifier's performance levels are shown by three precision-recall curves. The precision-Recall curve of Xception model shows the high precision and recall value, clearly indicating that the model outperforms than Inception and ResNet model.

### 3.3. Model loss

The Loss Function is a crucial element of neural networks, representing the model's prediction error. Both training and validation data are used to compute the loss. To train the model, training data is utilized. A key component of neural networks is the Loss Function, which shows the prediction inaccuracy of the model. The loss is calculated using data from both training and validation. Training data is used to train the model.

Fig. 7 depicts the average loss values for training and validation datasets throughout learning. Training loss surpasses validation loss due to the latter's lack of regularization and larger dataset size. After each training iteration, the network optimizes its parameters and computes the total batch loss[27]. The Enhanced Xception model quickly learns an accurate representation of normality, evidenced by low validation loss after a few epochs. This study reduces the loss function value by adjusting weight vector values and employing Adam optimization during training. Smaller validation sets mitigate overfitting risks, particularly obvious with larger training datasets prone to noise and outliers.

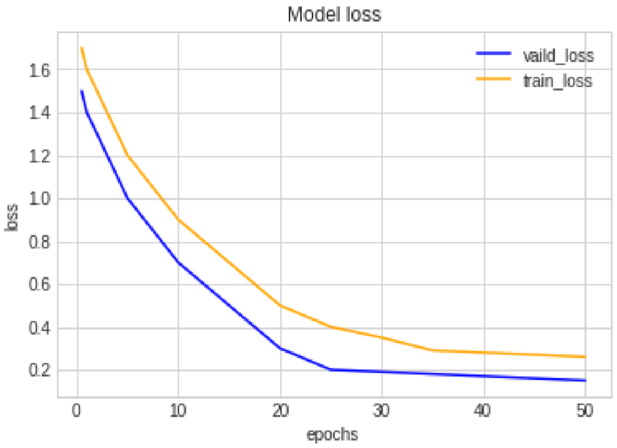

**Fig. 7.** Model loss of Improved Xception

**3.3.1. Accuracy.** The proportion of accurately predicted images out of the total number of predictions is defined as accuracy and it is calculated using the equation below,

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}.$$

Accuracy of training data and validation data of proposed method is shown in Fig. 8.

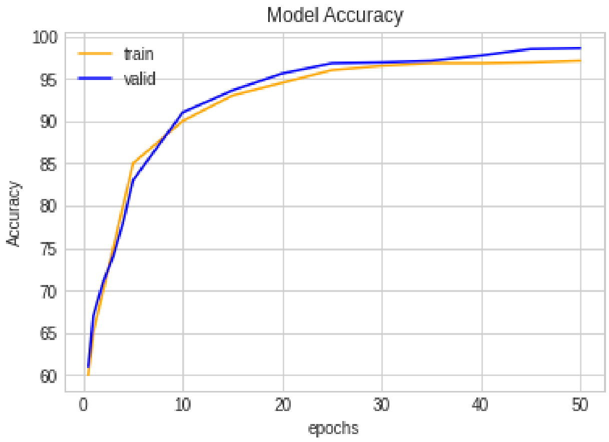

**Fig. 8.** Accuracy of Improved Xception

The global average pooling layer is additionally added in this Xception architecture for accurate classification of fake videos from the real ones. Also, due to robust training and extraction of deep features with the help of base layers in Xception model, the detection accuracy of suggested work is high on validation data. The validation accuracy is high when compared to training accuracy. This indicates the model performs better for new data.

**3.3.2. Comparison of Accuracy with Existing Techniques.** The recommended approach beats both Inception and ResNet in Table 1 at Epoch 50, showing a noteworthy 5.54% improvement over ResNet and a substantial 6.56% gain in accuracy over Inception. As a result, it has been confirmed that the suggested strategy can achieve 5% greater accuracy than the current methods.

*Table 1* – **Comparison of Accuracy with Existing models**

| Epochs | Inception | Resnet | Proposed Improved Xception |
|---|---|---|---|
| 0 | 50.14 | 55 | 62.43 |
| 10 | 66.2 | 73.5 | 76.8 |
| 20 | 71.15 | 81.2 | 85.1 |
| 30 | 82.4 | 83.5 | 90.3 |
| 40 | 87.3 | 89.2 | 94.3 |
| 50 | 92.2 | 93.12 | 98.26 |

This underscores the effectiveness of the suggested method in achieving enhanced accuracy, emphasizing its potential superiority in the specified task when compared to Inception and ResNet.

### 3.4. Training time

In neural networks, freezing a layer refers to managing the updating of the weights. A frozen layer indicates that there is no way to change the weights any more. This is used to cut down on training computation time without significantly sacrificing accuracy. Certain layers of the model are frozen in order to maintain specific pretrained model features. Figure 9 displays the training time with an increase in layers.

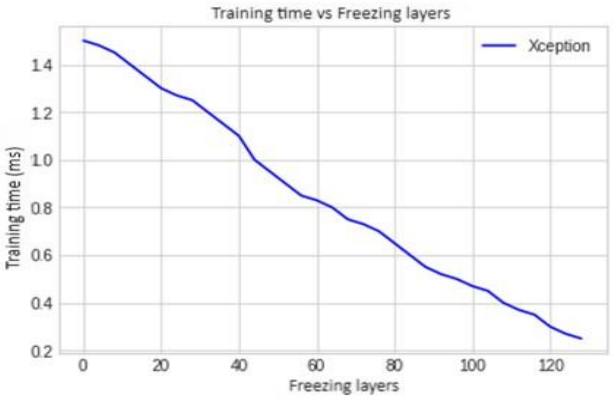

**Fig. 9.** Training time vs Freezing layers

With fewer frozen layers, training requires more time. However, as training progresses, the performance of the individual pre-trained models remains unchanged. However, the accuracy decreases with increasing numbers of frozen layers. A compromise between accuracy and training time is achieved in the Xception model by using 116 layers.

**3.4.1. Comparison of Detection Time.** Fig. 10 compares the detection times of the proposed model with ResNet and Inception. The proposed model achieves a lower time of 34.15 ms, signaling enhanced efficiency under the same configuration. Detection times for ResNet and Inception are 54.23 ms and 36.81 ms, respectively. More training time is needed when there are fewer frozen layers. Still, each pre-trained model's performance doesn't change while training goes on. With more frozen layers, the accuracy does, however, decline.

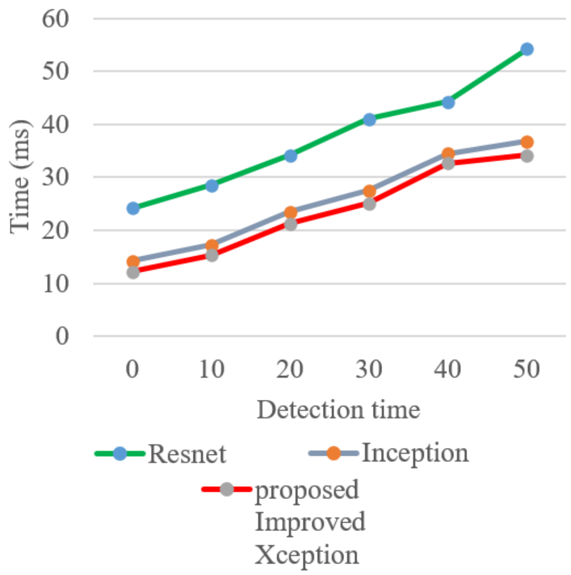

**Fig. 10.** Comparison of Detection Time

The 116 layers in the Xception model allow for a trade-off between training time and accuracy.

Using eight V100 GPU accelerators with 32 GB of RAM each, the method uses transfer learning to implement the Xception model on an NVIDIA DGX-1 platform. The outcomes of using fictitious faces in the video frames are shown in Fig. 11.

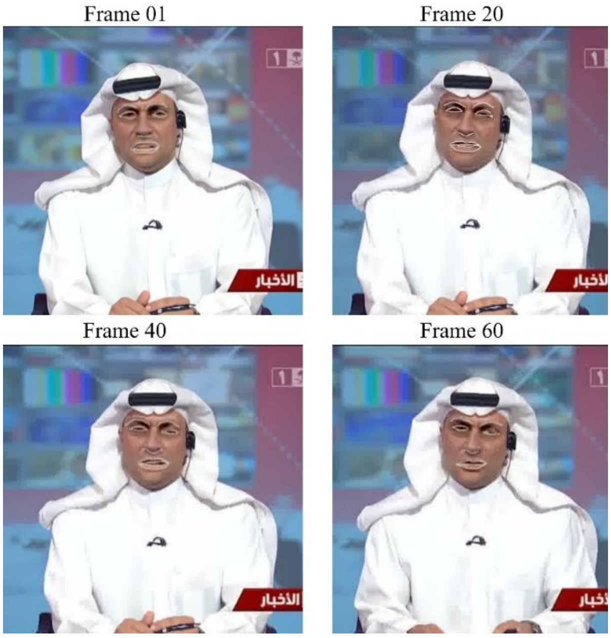

**Fig. 11.** Visualization results of fake video frame using Improved Xception method

With the use of landmark features that have been collected from video frames, the deepfakes are precisely identified. When attempting to extract features from the image, the landmark features are essential. As shown in Figure, the variations are plotted with respect to movements such as eyes, eyebrow, nose and lips. The original face in the video is covered by the fake image and it clearly indicates that the facial parameters is not compatible with original face movements.

Based on the differences between each frame, the suggested method divides each video frame into actual and fraudulent ones. In order to accomplish feature extraction and identification, the changes of the eyes, nose, lips, and eyebrows in each frame were extracted. Fig. 11 shows the four bogus frames (1, 20, 40, and 60) that are included in the video. To assess the model's correctness, it is trained using various video frames.

**Fig. 12, a.** Prediction results of video frame using Improved Xception method for real image

The outcomes shown in Fig. 12, a show how the suggested method's predictions relate to actual visuals found in video frames. By using transfer learning to an already trained Xception model for DeepFake detection, this method effectively identifies subtle variations in video frames and offers crucial information about their legitimacy. This method not only increases detection precision but also demonstrates adaptability to the constantly shifting obstacles posed by manipulating videos.

**Fig. 12, b.** Indicates the fake video frames detected by the Improved Xception method

An identification of modified video frames using the suggested method is shown in Figure 12(b). The deep learning method makes use of a pre-trained model to capture complex patterns and identify minute characteristics that are suggestive of DeepFake material. Through the use of the Xception Model's pre-existing information, transfer learning expedites the model's adaptability to DeepFake detecting nuances and improves overall efficiency, hence cutting the training time dramatically.

The signatures for Fig. 11, 12, a, and 12, b are sourced from

https://github.com/ondyari/FaceForensics?tab=readme-ov-file [28].

## Conclusion

This study introduces a novel approach for detecting deepfakes, termed the Enhanced Xception model, which employs facial landmark characteristics to automatically identify falsified content within videos. The method undergoes rigorous training to ensure precise detection capabilities. On the FaceForensics++ dataset, the suggested method's transfer learning strategy produced a 5% improvement over current systems, which was mostly attributable to a greater detection rate. Efficiency in this sense refers to how accurate the suggested model is. Moreover, assessments based on ROC, accuracy, precision-recall curve, and model loss are used to evaluate the performance of the transfer learning model. Table 1 shows that after the 50th epoch, the suggested method outperforms both Inception and ResNet, exhibiting a notable 5.54% improvement over ResNet and a significant 6.56% accuracy improvement over Inception.

As such, the accuracy of the suggested method is confirmed to be 5% higher than that of the current methods. This demonstrates the efficacy of the suggested approach in achieving increased accuracy, hinting at its potential superiority over Inception and ResNet for the specified task.

### References

1. Hashmi, M. F., Ashish, B. K. K., Keskar, A. G., Bokde, N. D., Yoon, J. H. and Geem, Z. W. (2020), “An Exploratory Analysis on Visual Counterfeits Using Conv-LSTM Hybrid Architecture”, *IEEE Access*, vol. 8, pp. 101293–101308, doi: https://doi.org/10.1109/ACCESS.2020.2998330
2. Guarnera, L., Giudice, O. and Battiato, S. (2020), “Fighting Deepfake by Exposing the Convolutional Traces on Images”, *IEEE Access*, vol. 8, pp. 165085–165098, doi: http://dx.doi.org/10.1109/ACCESS.2020.3023037
3. Neves, J. C., Tolosana, R., Vera-Rodriguez, R., Lopes, V., Proença, H. and Fierrez, J. (2020), “GANprintR: Improved Fakes and Evaluation of the State of the Art in Face Manipulation Detection”, *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 1038–1048, Aug. 2020, doi: https://doi.org/10.1109/JSTSP.2020.3007250
4. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. and Ortega- Garcia, J. (2020), “DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection”, *Information Fusion*, doi: https://doi.org/10.1016/j.inffus.2020.06.014
5. Verdoliva, L. (2001), “Media Forensics and DeepFakes: An Overview”, *arXiv preprint*, doi: https://doi.org/10.48550/arXiv.2001.06564
6. Nguyen, H. H., Yamagishi, J. and Echizen, I. (2018), “Capsule-forensics: Using capsule networks to detect forged images and videos”, *arXiv preprint*, doi: https://doi.org/10.48550/arXiv.1810.11215
7. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Nießner, M. (2019), “Faceforensics++: Learning to detect manipulated facial images”, *arXiv preprint*, doi: https://doi.org/10.48550/arXiv.1901.08971
8. Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M. and Verdoliva, L. (2018), “Forensictransfer:Weakly-supervised domain adaptation for forgery detection”, *arXiv preprint*, doi: https://doi.org/10.48550/arXiv.1812.02510
9. Li, Y., Chang, M.-C., Farid, H. and Lyu, S. (2018), “In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking,” *arXiv preprint*, doi: https://doi.org/10.48550/arXiv.1806.02877
10. Nguyen, H. H., Fang, F., Yamagishi, J. and Echizen, I. (2019), “Multi-task learning for detecting and segmentingmanipulated facial images and videos,” *arXiv preprint*, doi: https://doi.org/10.48550/arXiv.1906.06876
11. Ciftci, U. A. and Demir, I. (2019), “Fakecatcher: Detection of synthetic portrait videos using biological signals”, doi: https://doi.org/10.48550/arXiv.1901.02212
12. Brundage, M. et al. (2018), “The malicious use of artificial intelligence: Forecasting, prevention, and mitigation”, *arXiv:1802.07228*, available at: http://arxiv.org/abs/1802.07228
13. Christian, Jon (2018), *The Outline: Experts Fear Face Swapping Tech Could Start an International Showdown*, available at: https://tinyurl.com/3hbzpw2r
14. Nasir, J. A., Khan, O. S. and Varlamis, I. (2021), “Fake news detection: A hybrid CNN-RNN based deep learning approach”, *International Journal of Information Management Data Insights*, vol. 1(1), 100007, doi: https://doi.org/10.1016/j.jjimei.2020.100007
15. Jung, T., Kim, S. and Kim, K. (2020), “DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern”, *IEEE Access*, vol. 8, pp. 83144–83154, doi: https://doi.org/10.1109/ACCESS.2020.2988660
16. Hsu, C.-C., Zhuang, Y.-X. and Lee, C-Y. (2020), “Deep Fake Image Detection Based on Pairwise Learning”, *Applied Sciences*, vol. 10(1), 370, doi: https://doi.org/10.3390/app10010370
17. Korshunov, P. and Marcel, S. (2018), “Deepfakes: a new threat to face recognition? assessment and detection”, *arXiv preprint*, doi: https://doi.org/10.48550/arXiv.1812.08685
18. Chintha, A., Thai, B., Sohrawardi, S. J., Bhatt, K., Hickerson, A., Wright, M. and Ptucha, R. (2020), “Recurrent Convolutional structures for audio spoof and video Deepfake detection”, *IEEE Journal of Selected Topics in Signal Processing*, vol. 14(5), pp. 1024–1037, doi: https://doi.org/10.1109/jstsp.2020.2999185
19. Caldelli, R., Galteri, L., Amerini, I. and Del Bimbo, A. (2021), “Optical flow based CNN for detection of unlearnt DeepFake manipulations”, *Pattern Recognition Letters*, vol. 146, pp. 31–37, doi: https://doi.org/10.1016/j.patrec.2021.03.005
20. Wang, R., Ma, L., Juefei-Xu, F., Xie, X., Wang, J. and Liu, Y. (2019), “FakeSpotter: A Simple Baseline for Spotting AI-Synthesized Fake Faces”, *arXiv preprint*, doi: https://doi.org/10.48550/arXiv.1909.06122
21. McCloskey, S. and Albright, M. (2018), “Detecting GAN-Generated Imagery Using Color Cues”, *arXiv preprint*, doi: https://doi.org/10.48550/arXiv.1812.08247
22. Nataraj, L., Mohammed, T., Manjunath, B., Chandrasekaran, S., Flenner, A., Bappy, J. and Roy-Chowdhury, A. (2019), “Detecting GAN Generated Fake Images Using Co-Occurrence Matrices”, *Electronic Imaging*, vol. 5, pp. 1–7, doi: https://doi.org/10.48550/arXiv.1903.06836

23. (2023), FaceForensics++. (n.d.). *Kaggle: Your Machine Learning and Data Science Community*, available at: https://www.kaggle.com/sorokin/faceforensics
24. Popat, K., Mukherjee, S., Yates, A. and Weikum, G. (2018), "Declare: Debunking fake news and false claims using evidence-aware deep learning", *arXiv:1809.06416*, doi: https://doi.org/10.48550/arXiv.1809.06416
25. Thangaraj, R., Anandamurugan, S. and Kaliappan, V.K. (2020), "Automated tomato leaf disease classification using transfer learning-based deep convolution neural network", *Journal of Plant Diseases and Protection*, vol. 128, pp. 73–86, doi: https://doi.org/10.1007/s41348-020-00403-0
26. Chollet, F. (2017), "Xception: Deep learning with Depthwise separable convolutions", *2017 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Honolulu, HI, USA, doi: https://doi.org/10.1109/cvpr.2017.195
27. Garcia Cordero, C., Hauke, S., Muhlhauser, M. and Fischer, M. (2016), "Analyzing flow-based anomaly intrusion detection using Replicator neural networks", *2016 14th Annual Conference on Privacy, Security, and Trust* (PST), doi: https://doi.org/10.1109/pst.2016.7906980
28. (2023), *Ondyari/FaceForensics: Github of the FaceForensics dataset*. (n.d.), GitHub, available at: https://github.com/ondyari/FaceForensics?tab=readme-ov-file

Received (Надійшла) 14.01.2023
Accepted for publication (Прийнята до друку) 24.04.2023

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

**Раджакумаресваран Велусамі** – Ph.D., доцент кафедри комп'ютерних наук і дизайну, Інженерний коледж Ероде Сенгунтар, Тудуппаті, Таміл Наду, Індія;
**Velusamy Rajakumareswaran** – Ph.D., Assistant Professor, Department of Computer Science and Design, Erode Sengunthar Engineering College, Thuduppathi, Tamil Nadu, India;
e-mail: dr.rajav@esec.ac.in; ORCID ID: https://orcid.org/0000-0002-2758-6107;
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57205750551.

**Рагуваран Сурендран** – Ph.D., доцент кафедри обчислювального інтелекту, Школа обчислювальної техніки, Науково-технічний інститут SRM, SRM Nagar, Каттанкулатур 603203, Ченнаї, ТН, Індія;
**Surendran Raguvaran** – Ph.D., Assistant Professor, Department of Computational Intelligence, School of Computing, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur 603203, Chennai, TN, India;
e-mail: prof.sraguvaran@gmail.com; ORCID ID: https://orcid.org/0000-0003-3998-7706;
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57211538473.

**Чандрасекар Венкатачалам** – Ph.D., професор, факультет інженерії та технології, Джайнський університет, Бангалор, Індія;
**Venkatachalam Chandrasekar** – Ph.D., Professor, Faculty of Engineering and Technology, Jain University, Bangalore, India;
e-mail: chandrasekar.v@jainuniversity.ac.in; ORCID ID: https://orcid.org/0000-0001-7258-0794;
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=56924794400.

**Раджкумар Сугаванам** – магістр (техн.), доцент, технологічний коледж Сона (автономний, афільований до університету Анна), Салем, Індія;
**Sugavanam Rajkumar** – MTech, Assistant Professor, Sona College of Technology (Autonomous); affiliated to Anna University, Salem, India;
e-mail: rajkumar.s@sonatech.ac.in; ORCID ID: https://orcid.org/0000-0001-7136-3771;
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=59097616000.

**Арун Віджаякумар** – доктор філософії, професор кафедри електротехніки та електроніки Інженерної школи університету Мохана Бабу, Тірупаті, Індія;
**Vijayakumar Arun** – Ph.D., Professor, Department of Electrical and Electronics Engineering, School of Engineering, Mohan Babu University, Tirupati, India;
e-mail: arunphd1986@gmail.com; ORCID ID: https://orcid.org/0000-0003-0016-9298;
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57191294342.

## Виявлення DeepFake за допомогою моделі Xception на основі трансферного навчання

В. Раджакумаресваран, С. Рагуваран, В. Чандрасекар, С. Раджкумар, В. Арун

**Анотація. Обґрунтування мети дослідження.** Останнім часом кілька підходів до маніпулювання обличчями у відео були широко застосовані та доступні для громадськості, що робить редагування облич у відео легким для всіх без особливих зусиль із реалістичними зусиллями. Незважаючи на користь у різних сферах, ці методи можуть завдати значної шкоди суспільству, якщо використовувати їх для поширення дезінформації. Тому також важливо правильно визначити, чи було спотворене обличчя у відеоряді. Щоб виявити цей глибокий фейк, у минулих роботах можна використовувати згорточні нейронні мережі. Однак для цього потрібна більша кількість параметрів і більше обчислень. Тому для подолання цих обмежень і точного виявлення глибоких фейків у відео пропонується модель на основі навчання передачі під назвою Improved Xception model. **Отримані результати.** Ця модель навчена за допомогою витягнутих орієнтирів обличчя з надійним тренуванням. Крім того, покращена точність виявлення моделі Xception оцінюється разом із ResNet і Inception, враховуючи втрати моделі, точність, ROC, час навчання та криву Precision-Recall. Результати підтверджують успіх запропонованої моделі, яка використовує методи навчання передачі для виявлення шахрайських відео. Крім того, метод демонструє помітне підвищення ефективності на 5% порівняно з поточними системами.

**Ключові слова:** DeepFake; редагування облич; Improved Xception model; Precision-Recall.