Journal of Metaverse
Research Article
Received: 2023-08-30 | **Reviewing:** 2023-09-06 & 2023-12-12 | **Accepted:** 2023-12-13 | **Online:** 2023-12-14 | **Issue Date:** 2024-06-30
Year:2024, **Volume:**4, **Issue:**1, **Pages:**11-20, **Doi:** 10.57019/jmv.1338907

Face Warping Deepfake Detection and Localization in a Digital Video using Transfer Learning Approach

Rachel Selva Dhanaraj
National Institute of Technology Tiruchirappalli
Tamil Nadu, India
rselva00607@gmail.com
0000-0002-4834-4174
(Corresponding Author)

M Sridevi
National Institute of Technology Tiruchirappalli
Tamil Nadu, India
msridevi@nitt.edu
0000-0003-0657-7188

***Abstract*—Generative AI (GenAI) can generate high-resolution and complex content mimicking the creativity of humans, thereby benefiting industries such as gaming, entertainment, and product design. In recent times, AI-generated fake videos, commonly referred to as deepfakes, have become more commonplace and convincing. An additional deepfake technique, face warping, uses digital processing to noticeably distort shapes on a face. Tracking such warping in images and videos is crucial and preventing its use for destructive purposes. A technique is proposed for detecting and localizing face warped areas in video. The input video is extracted to perform various image pre-processing techniques that refine the video into a format that is more likely to classify the classes efficiently. Transfer learning is employed, and the pre-trained model is adopted to train using Convolutional Neural Network (CNN) with the source videos to identify face warping. Based on the experimental results, it was determined that the proposed model detects and localizes the warped areas of the face satisfactorily with an accuracy of 89.25%.**

***Keywords*—Deepfake, Face Warping, Transfer Learning, Convolutional Neural Network, Generative Artificial Intelligence**

I. INTRODUCTION

The capabilities of GenAI [1] have been significantly enhanced by recent breakthroughs in the field, such as Generative Pre-trained Transformer (GPT) and Midjourney. The advancements of GenAI have opened up new possibilities for solving complex problems, creating art, and assisting scientists. Deepfake is the outcome of artificial intelligence technology, as various new applications and services are on the horizon. When digital images, audio and videos are simulated or forged, with the utilization of the machine learning's generative model, it is referred to as deepfake. As appraising the digital image content or assessing the forged regions would be a commendable act, as in a judicature when digitalized videos are considered as evidence or malicious purposes, the same could stand as a lifesaver. The other positive note with the application of deepfake is, the editing of movie clips without shooting them again, and also the creation of audio-voice of individuals who have lost theirs accidentally [2, 3]. Deepfakes has some additional worries attached to it. It is not common for celebrities to use deepfakes, the use of which has appeared on the internet: the introduction of Nicholas Cage in films he didn't play such, as "The Matrix" and "Fight Club" or Jim Carrey's admirable music video where he was seamlessly integrated into Kubrick's "The Shining" instead of Jack Nicholson. The video of Obama fabricated by Buzzfeed in association with 'Monkey paw Studios', or the video in which alleged statements made by Mark Zuckerberg claiming the platform's capability to plunder users' information [4].

Face warping [5, 6, 7] is one of the deepfake techniques that has become popular in recent times. The digital manipulation of a face that results in a significant distortion of any shapes depicted on the face is known as face warping. Face warping can be used for both creative and face distortion correction. Face warping can be divided into two groups: facial expression manipulations and face identification manipulations. A noteworthy technique for the manipulation of facial expressions is the Face2Face method. With the usage of the community hardware, the said methodology swaps the facial expression of an individual with another in real-time. "Synthesizing Obama" a follow-up work, animates the facial features of an individual on the basis of an input segment of audio. Moving on to the second category of face forgery is that identification manipulation. This mechanism replaces the target's face instead of faking facial expressions. Thus, gives rise to a category known as the swapping of faces. It received renowned popularity beyond the widespread use of consumer-level use of applications like Snapchat. Face swapping is also done by deepfakes via deep learning [8]. Though the former relies on a simple Computer- Graphics form which runs in real-time, deepfake is a sluggish task as it needs to be instructed for pair of videos [9].

However, the more worrying aspect is the malicious use of face warping, as this sector dominates the positive ones. The mechanism of processing subversive videos and images is very simple in today's world, as it only requires an identity photo or a video to complete the forgery. Thus, posing a severe threat to the common man and affecting public figures. To note a few, a CEO was duped of $243000 using voice deepfake. Recently liberated software called deepNude exhibits very upset trends. As it can transform a person into infanticide porn. Similarly, Zao, the Chinese-app is yet another example, which swaps the face of individuals onto the physique of film stars and incorporates them into prominent TV clippings. These forms of counterfeiting do not only possess a huge threat to privacy but also various aspects of human lives [3]. Therefore, a reliable prediction method for

Cite (APA): Dhanaraj, R., Sridevi, M. (2024). Face Warping Deepfake Detection and Localization in a Digital Video using Transfer Learning Approach. *Journal of Metaverse*. 4 (1), 11-20. **Doi:** 10.57019/jmv.1338907

AI-edited images is needed to determine if digital videos have been tampered with. This paper is an attempt to create a model that can detect and locate face warping on video quite well. The task of automatic face identification based on the similarity of facial images in computer vision is challenging. Easy detection of fake faces requires strong local changes in representation and lighting, global pose changes, temporal changes, partial occlusion, and affine transformations, but in the current scenario, it's not required.

A. Contributions

The major contributions of this work are as follows:

- Pre-processing techniques and various data transformations like image alignment and normalization, Image degradation and Illumination normalization are applied to the dataset for more accuracy.
- Exploiting knowledge gain through Transfer Learning.
- Detect and localize the forged areas in the videos.
- Achieving 94.5% accuracy in the top 5 makes the model more efficient in terms of computation time and accuracy.

The organization of this paper is structured as follows. An outline of research related works is illustrated in Section 2. Section 3 introduces the Proposed Methodology. Section 4 discusses Results and Performance Analysis, followed by the Conclusion in Section 5.

II. RELATED WORKS

By embracing the new technology of GenAI, will usher in a new era of creativity, efficiency, and progress. Among the benefits of GenAI are faster product development, improved customer experiences, and greater employee productivity. The breakthroughs in media promotion techniques have made it a child's play for intruders to formulate forged images and videos. Recent technologies, obtained from social networking sites, authorize the real-time generation of a fabricated video. With the rise in the number of attacks, the methodologies developed by researchers to detect forged images have become obsolete as they are primarily focused on certain domains. Therefore, the need of the hour is the development of effective tools to intuitively ascertain forged videos. The associated literature for various deepfake [10] techniques is included in this subsequent section. A comparative approach on the various techniques used is tabulated in Table 1.

A. Image Processing based method

The method mentioned in [2] envisages a new technique to detect artificially generated videos or fake face images called deepfakes. Deepfake generation algorithms can only generate a limited resolution and prescribed size, which then needs to be processed in the form of blurring and transformation to match the necessary results, in this case, the faces that must be swapped in the original video. Now, in the deepfake videos, special artifacts are left behind on the Region of Interest (ROI) by the additive blur transformations, which then could be efficiently captured by using Haar Wavelet transformation to ascertain the divergence between the ROI and the rest of the image. The effectiveness of the proposed scheme was an outstanding efficacy of 90.5%.

TABLE I. COMPARATIVE APPROACH USED BY VARIOUS TECHNIQUES

Title	Approach
Effective and Fast Deepfake Detection Method Based on Haar Wavelet Transform, M.A. Younus, T. M. Hasan., 2020 [2]	Employs Haar Wavelet transformation with Special artifacts left behind on the ROI
Exposing Deep Fakes Using Inconsistent Head Poses, X.Yang, Y. Li, S. Lyu, 2019 [11]	Uses Support Vector Machine with Interlacing amalgamated face zones
Capsule forensics: Using Capsule Networks to Detect Forged Images and Videos, Nguyen, H. H., et al. 2019 [14]	Utilize Capsule network and CNN
Deepfake Video Detection by Combining CNN and RNN, Y.Al-Dhabi,S. Zhang. 2021 [13]	Applies Inter-frame and Intra-frame features in deep neural networks with CNN and RNN
A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow features, Pallabi Saikia, 2022 [15]	Employs Optical flowbased feature extraction approach using CNN and RNN
Exposing Deepfake Videos By Detecting Face Warping Artifacts, Yuezun Li, et al 2019 [17]	Follows Ensemble method with GAN technique
Few-Shot Training GAN for Face Forgery Classification and Segmentation Based on the Fine-Tune Approach, Lin, Y.-K.; Sun, H.-L.2023 [18]	Uses False pixel percentage threshold in GAN method
A GAN-Based Model of Deepfake Detection in Social Media Preeti, Manoj et al, 2023 [16]	A comparative case study on GAN was conducted

B. Machine Learning based method

The approach in [11] depends on the overall observation that deepfakes on images and videos are fabricated by interlacing amalgamated face zones into the authentic image, and in accomplishing the same, generating errors that could be divulged as and when, from the face images 3D head poses are estimated. Evaluations were performed to illustrate this phenomenon and, thereby, developed a classified methodology based on this intimation. A Support Vector Machine (SVM) category was evaluated with the features of this cue, with the aid of real-face image set and deepfake. User identification has utilized face as its mainstream tool. However, it only takes a few seconds, to interchange the faces between two images of facial appearance, with the credit going to the popularity of the face-swapping applications.

C. Deep Learning based method

The authors in [2] revolve around the observation that presents deepfake algorithms can only produce resolution limited images, which must be distorted more to match the authentic face in the input video. Such transformations have been found to leave inherent artifacts in the resulting deepfake videos that can be captured by CNN. With this method, you don't need to use images generated by deepfake as negative training samples. This is because it targets the affine face

distortion artifact as a discriminating attribute to compare between fake and the real image.

1) CNN based method

The method used in [13] works on the fusion of CNN and Recurrent Neural Networks (RNN) using a pre-trained Resnext model for extracting descriptors, and these descriptors are used for Long Short-Term Memory (LSTM) training. CNN and RNN are used together to collect inter-frame and intra-frame features that are employed to locate if a video is real or fake. It depicts how the system achieves competitive results using a simplified architecture. [14] instituted the use of a capsule network to assess different kinds of spoofs, from replay generated videos with the aid of a deep convolutional neural network. Without stopping here, they also demonstrated the same could be used in realms besides computer vision. Also, the thing to note is the use of random noise which proved to be beneficial in the training phase. The prospect would be to refrain from confrontational machine attacks, specifically over the initiated random noise, and to oblige the methodology efficiently against mixed attacks. The author in [15] extracted temporal characteristics and integrated them into an association model for classification using the optical flow-based feature extraction approach. This association model is built on a CNN architecture and RNN combination. When applied to opensource datasets such as DFDC, Celeb-DF, and FF++, hybrid models show strong performance, demonstrating their effectiveness in handling these particular data sets. With a sample size of only 100 samples, the approach achieves an accuracy of 66.26%, 79.49%, and 91.21% in DFDC, Celeb- DF, and FF++ respectively.

2) GAN based method

Generative Adversarial Network (GAN) [16] is a model of prominent generation, impressively used in various applications. GenAI poses significant and rapidly evolving risks. In addition to generating artifacts supporting increasingly complex scams, a wide range of threat actors have already utilized the technology to create “deepfakes” or copies of products. The paper presents research on methods primarily used to implement deepfakes. It covers deep fake implementation utilizing a deep convolution-based GAN model, deep fake manipulation, and detection methods. A comparative study on analysis between the proposed GAN and other existing GAN models with parameters Inceptionv3 Score “IS” and Frechet Inception-v3 Distance “FID” is also incorporated. This document also describes the open questions and future trends that need to be taken into account in order to move forward in this field. Using MesoNet as a foundation, [17] trains a GAN and extracts discriminators as a dedicated deepfake detection module. Multiple discriminator architectures are tested using multiple datasets to investigate how different setups and training methods change the effectiveness of the discriminator. Finally, the model uses the ensemble method to increase the effectiveness of groups of GAN discriminators. Results show that the GAN discriminator does not work well on videos from unfamiliar sources, even when complemented by the ensemble method. [18] proposes a GAN-based deep learning method that allows to detect spurious regions with fewer training samples. The suggested architecture’s generator component is utilized to create predictive segments that display the bias of each pixel. To solve the classification problem, the false pixel percentage threshold is utilized to assess whether the input image is incorrect. The frames are extracted from the video and predicted if they are fake. With GenAI models, one of the breakthroughs is their ability to leverage unsupervised or semi-supervised learning approaches. By leveraging unlabelled data in this way, organizations can build foundation models more quickly and easily. This method has better classification and segmentation compared with other studies as revealed in the experimental results.

The main research challenge encountered in carrying out this work was the lack of research literature and datasets on facial warping.

III. Proposed Methodology And Architecture

Face warping images and videos have become very common in the media, so a need for a technique that would address the issue of forgery in the images is indeed a necessity. As mentioned in the related works section, there are authors who have come up with methods to handle deepfake videos and not much research work is performed on face warping techniques and that too, they do have limitations of utilization of resources, time-consuming and accuracy.

The proposed method has made an attempt to detect face warping and even localize them. The method employs CNN efficiently to detect face warped videos. A CNN with transfer learning is used to achieve fixed network parameters. The goal is to transfer knowledge from the high-level feature vectors of the CNN network to the offline pre-processed target video, where the features are trained in a SoftMax classifier for face warping identification. Fig I manifests the proposed technique graphically.

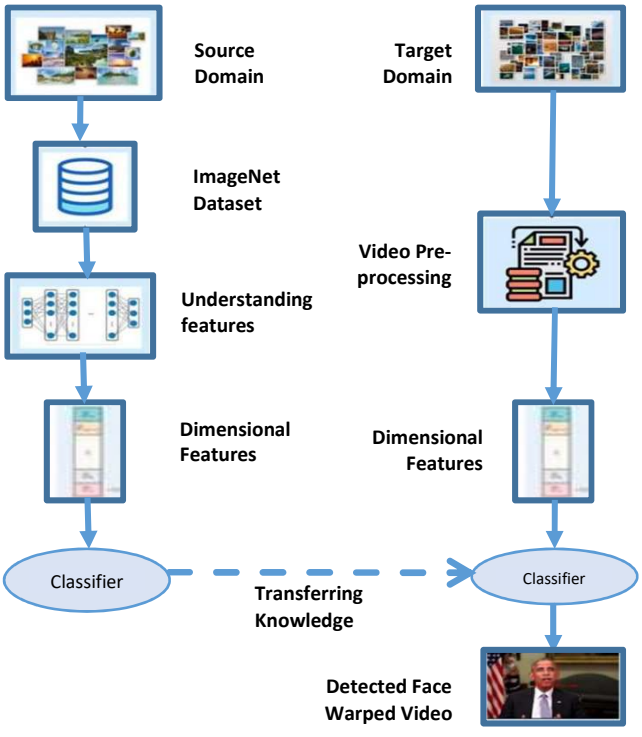

Fig I. Block diagram of Proposed Face Warping detector

The main steps in the method are as follows:

- Pre-process the target video dataset
- Train the classifier for the source dataset
- Transfer the knowledge from the obtained resultant model of the source dataset to the target dataset
- Train the classifier for the target dataset and obtain the detected face warped video

A. Video Pre-processing

The input video is in raw format and it needs to be processed so as to perform the required task. The flow of the process is represented in Fig II and the algorithm is provided in Algorithm I. First, the video is captured from any source, such as a camera, and split into frames for further processing. A cascade of images is performed to divide the face detection problem into several stages. For each block, a very rough and quick test is run [19]. If this passes, a little more detailed testing is performed. The algorithm can have 30-50 of these levels or cascades and will only recognize faces [20] if all levels pass. The advantage is that most of the images return negatives in the early stages. The algorithm wastes no time testing all the features. A rectangular bounding box is drawn on the selected faces. Furthermore, cropping is performed on the images and later color conversion techniques are used to convert into a format that is required by the proposed model. Data is later transformed by applying various methods like image alignment, image degradation and illumination normalization. Lastly, batch processing task is performed.

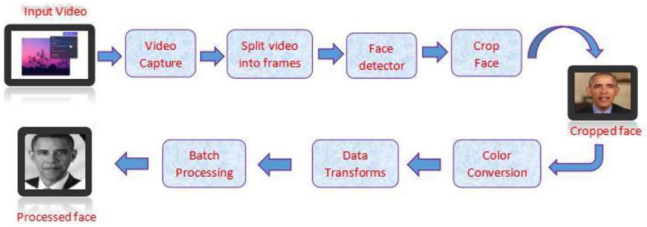

FIG II. VIDEO PRE-PROCESSING

1) Color Space Conversion

The representation of a color [21] that is translated from one space to another is color space conversion [19]. This typically happens when translating an image from one colour space to another, with the intention of keeping the translated image as similar to the source as feasible. There are many popular and widely used color spaces, such as RGB, CMYK, Y'UV, YIQ, Y'CbCr, HSV, etc. The types of color spaces depend on the medium we are using, i.e. digital or print format. . For example, digital devices use a color space called RGB. It is based on colored light. Different color spaces exist because they show color information in ways that facilitate particular calculations or because they improve the intuitiveness of color detection. For instance, the RGB color space describes a color as the proportion of blended red, green, and blue hues. According to other color models, colors are classified based on their tint (color shade), saturation, and luminance (intensity).

ALGORITHM I: Pre-Processing the Video Dataset

```
Input : Video dataset
Output: Processed Face
Procedure:
 1. Use VideoCapture() to capture a video object
    for the camera
    while true do
    Adopt the read() method to read the frames
    using the above created object
    end
   2. Employ imshow() method to display the
      frames in the video
   3. Pass in the image and cascade names as
      command-line argument
   4. Create the haar cascade to initialize with
      face cascade using Cascade Classifier
      method
   5. Read the image
   6. Detect faces using
      faceCascade.detectMultiScale method
   7. Draw the rectangle using the built-in
      rectangle() function
   8. Apply Image.crop() method to crop a
      rectangular portion of any image
   9. Use cvtColor() method to convert an image
      from one color space to another
   10. Image alignment is performed by applying
       a deep funnelling technique
   11. Apply normalize() function to normalize
       the image
   12. Apply Otsus thresholding for image
       degradation
     a. Acquire the histogram of the image
     b. Compute the threshold value T
     c. Replace image pixels into white in
        those regions, where saturation is
        greater than T and into black in those
        regions, where saturation is lower than
        T
   13. Apply batch processing on the images
End
```

2) Data Transforms

The following transformations are applied on the images to get the desired and accurate output.

- *Image Alignment and Normalization:* Real face image data frequently has issues with people' appearances, which can change dramatically for huge poses from a person's profile to their frontal perspective. Face alignment [19] is reportedly used to input faces to deep networks to match diverse position variants of face data into a canonical pose, enhancing the effectiveness of human feature extraction techniques since we are minimizing pose variability and using photos that are aligned with deep funnelling, that improves recognition performance. Additionally, zero mean and one standard deviation method is used as needed to hasten the convergence of the network during training. This ensures the consistency of the data distribution and the input parameters for normalized data.
- *Image degradation:* Most data acquired through online media suffers from colour compression because of the poor performance of mobile devices in supporting a restricted number of colors. The drawbacks are addressed using both global quantization and area-based quantization. Otsus's thresholding methodology [22], have been used for both, but the former leads to the generation, with the help of RGB image based on a specific layer, a threshold vector with multiple layers. As the quantization values of a plane change, the value

of the threshold vector changes. For instance, a 6-bit quantization indicates that the picture is quantized using 6 thresholds that are produced from the complete raw format image. This takes into consideration how an RGB image's greyscale varies from layer to layer. To produce a vector with 6 thresholds in 3 layers, thresholds are generated for the red, green, and blue layers (i.e., if the choice for the quantization layer is six layers to quantize each layer). It will be important to monitor how well deep networks behave when the grey level is decreased to get actual data from mobile devices. This study contends that the quantization procedure makes things more homogeneous.

- *Illumination Normalization:* It is anticipated that spatial disparities in the sensitivity of the camera systems will be present for images of faces captured using various spectral bands. For face images of the same class, rgbGELog and the widely used lighting normalization technique LSSF [23] are employed to reduce the impact of variability.

B. Train the classifier for the source dataset

A CNN [24, 25] is a deep learning algorithm that takes the input image from the source dataset, assigns importance to different objects in the image, and distinguishes them from each other. By using the appropriate filters, CNN can detect spatial and temporal connections in images. CNNs generally consist of three layers: convolutional, pooling, and fully linked layers which is referred from Fig III. From the input picture, the convolution function extracts high-level characteristics such as edges. By using dimensionality reduction, a pooling layer makes it possible to lower the amount of computational power needed to analyze the data. It also extracts key features that are positional and rotational invariant, thereby keeping the process of training the model effectively. The fully connected layer is added to make way to obtain highlevel features from nonlinear combinations represented by the output of convolutional layers.

The suggested model makes use of Xception [24], a deep convolutional neural network architecture with 71 layers of depth-separable convolutions and a linear stack of these layers with rest connections. Traditional convolutions can be substituted with depthwise separable convolutions, which are reportedly significantly faster to compute. On the majority of traditional classification problems, the Xception architecture performs better than VGGNet, ResNet, and Inception-v3. The data initially moves via the input stream, after which it moves through the intermediate stream eight times, before arriving at the output stream designated by Algorithm II. Batch normalization is applied after all convolutional layers and separable convolutional layers. Therefore, Xception is used to classify and obtain a model from the source dataset that is ImageNet. The obtained resultant model is transferred and used by the next module.

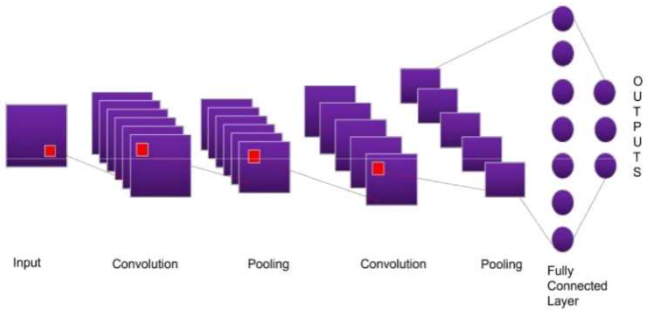

FIG. III. ARCHITECTURE OF CNN

ALGORITHM II: ALGORITHM FOR CREATING XCEPTION CNN MODEL

```
Input : Processed Video dataset
Output: Classified Video
Procedure:
 1. Import all necessary libraries needed for
    creating layers
 2. Write all the necessary functions for the
    following modules
    a. Create Conv-BatchNorm module
        i.   Input tensor x, number of filters,
             kernel size of convolutional layer,
             strides of convolutional layer
       ii.   Apply a convolutional layer to
             tensor x
      iii.   Set use bias=False()
       iv.   Apply Batch normalization
    b. Create SeparableConv-BatchNorm module
        i.   Input tensor x, number of filters,
             kernel size of Separable
             convolutional layer, strides of
             Separable convolutional layer
       ii.   Apply a Separable convolutional
             layer to tensor x
      iii.   Set use bias=False()
       iv.   Apply Batch normalization
 3. Write a function for each one of the 3 flows
    - Entry, Middle and Exit
    a. Create Entry block
        i.   Employ convolutional layer with 32
             filters
       ii.   Use RELU activation function
      iii.   Employ convolutional layer with 64
             filters
       iv.   Use RELU activation function
        v.   Employ Skip connection by using the
             ADD function
           A. Apply two separable convolutional
              layers
           B. Apply MaxPooling layer
    b. Create Middle block
        i.   Apply Skip connection by using the
             ADD function
           A. Use RELU activation function
           B. Apply separable convolutional
              layer
       ii.   Repeat the above steps 8 times
    c. Create Exit block
        i.   Apply Skip connection by using the
             ADD function
           A. Use RELU activation function
           B. Apply separable convolutional
              layer
           C. Apply MaxPooling layer
       ii.   Apply separable convolutional layer
      iii.   Use RELU activation function
       iv.   Employ GlobalAveragingPooling layer
        v.   Apply a fully connected layer with
             softmax activation function
End
```

C. *Transfer the knowledge from the obtained resultant model of the source dataset to the target dataset*

This module transfers the model obtained from module 2 to the classifier in module 4. The Xception model is an excellent fit for the goal of this work because it was trained on ImageNet [26], a sizable dataset containing one million and 200,000 (1.2 million) generalised data instances and 1000 different class labels (of faces, objects, places, things, animals, etc.). Transfer learning is often thought to be most appropriate in circumstances where the training data is insufficient in the literature. The transfer of information from one area (source) to another, nearly unrelated domain (target) is of greater importance, nevertheless. The algorithm is elaborated in algorithm 3.

ALGORITHM III: TRANSFER LEARNING

```
Input: Video source dataset
Output: New model
Procedure:
 1. Extract the source dataset ImageNet
 2. Choose the pre-trained CNN network model
 3. Import the dataset and load it into the
    network for training
 4. Apply image augmentation to the training data
    if required
 5. Validation data is chosen by splitting it
    from the training data to prevent overfitting
 6. Prepare the network for training
  a. Replace Last Learnable Layer
     i.   Find the last learnable layer in the
          network to change the number of classes
          to match the new dataset.
    ii.   Set the filter size to 1,1 to match the
          original learnable layer
   iii.   Change the number of filters to the
          number of classes in the new dataset
        A. To achieve faster learning in the new
           layer, the learning rates are changed
        B. Delete the last original learnable
           layer and connect the new learnable
           layer.
  b. Replace Output layer
     i.   Create a new classification layer
    ii.   Delete the original classification
          layer and connect the new
          classification layer in its place
   iii.   Set the Output size
 7. Train the network
  a. The learning in the transferred layers is
     slowed down by initializing the learning
     rates to a small value
  b. The accuracy of the validation data is
     calculated once every epoch by specifying
     validation frequency.
  c. Choose a minimal number of epochs since many
     epochs are not necessary for transfer
     learning.
  d. Mini-batch size of the image is specified
     to divide evenly into the number of samples
     to be trained
 8. Export the network architecture with the
    trained weights
```

The rich properties of CNN were examined on several levels in the work of [27]. Their research shown that the lower levels respond to edge-like traits, whereas the following layers mix these features with more abstract ones before they are combined as global features at the highest level. This is comparable to how the human visual cortex can identify people by processing different features of their faces separately and combining them into a single global feature [28]. The output of the final layer, which consists of the high-level feature vectors of a pre-trained CNN, extrapolates to a new target dataset more effectively than fine-tuning some network layers. For face warped detection, the high-level feature vectors of the Xception model worked well.

D. *Train the classifier for the target dataset to obtain the detected deepfake video*

This module takes as an input the target dataset and the model obtained and inputs it into the classifier. Xception is used to classify and obtain the face warped video. The following steps are carried out in the training phase:

- Load the training and test target datasets
- Define the Xception CNN mode
- Define the loss function
- Train the network on the training and test data

IV. RESULTS AND PERFORMANCE ANALYSIS

The results obtained using the proposed method are elaborated and delivered in the subsequent section. Performance analysis tests are conducted to check on various aspects of the objective fidelity criteria.

A. *Dataset description*

The transfer learning technique uses the ImageNet dataset for training. Extensive, accurate, and diverse, ImageNet serves as a useful resource for visual recognition applications such as object detection, image classification, and object localization. This dataset contains 1000 object classes and 20,000 categories. Table 2 contains a detailed description of the ImageNet dataset. The test videos are taken from the internet itself of YouTube videos and Celeb- DF [12] dataset consisting of original and deepfake videos, as there is no dataset on face warping.

TABLE II: IMAGENET DATASET [27]

Feature	Statistics
Founder	Fei-Fei Li
Number of images with hand annotation	14 million
Number of images with bounding box	1 million
Categories	20,000
Subset Available	Yes
Total number of non-empty WordNet synsets	21,841
Number of synsets with SIFT features	1000
Number of images with SIFT features	1.2 million
Domain	Computer Vision
Benchmark	ImageNet Large Scale Visual Recognition Challenge
Applications	Object Recognition, Classification, Clustering

B. *Experimentation Details*

Python was used to implement the proposed method and its results and performance are presented in this section. The

proposed method is compared with other state-of-the-art methods like Inception-v3 [29], ResNet [25] and VGGNet [30] and is found to yield better results than the other methods. For the videos from the Celeb-DF dataset that were taken into consideration for experimentation, the performance of the proposed approach was assessed. It is assumed that the input video should be in mp4 or avi video format.The proposed method applies the model obtained from the pre-trained network which uses ImageNet as source dataset and employs transfer learning technique to transfer the knowledge to a new network with the target video dataset. The newly generated network can classify the video dataset as face warped with an accuracy of 89.25%. The output is shown in Fig 4. The proposed method identifies the video as fake and puts a boundary box over the area of the object. The accuracy of such kind is possible since the target dataset is pre-processed using various methodologies which prepare itself for fine tuning and the usage of Xception CNN model which delivers good performance. The video in Fig IV(a) has been taken from the internet media for testing purpose. Jordan Peele, in the role of "Obama" in the sculpture, is shown acting out his well-known impersonation of the late president. Anyone can make a highly convincing replica of a human subject using deepfakes' machine learning algorithm, which comes with a tonne of photographic data to teach the computer what the picture should be like. Jordan Peele's face has been distorted with Obama's. The proposed model is able to detect the face warped face of Obama with an average accuracy of 87.35%. Three cases are analyzed and their details are provided below.

Case I: Obama face detected as fake

In Fig IV(b), IV(c), IV(d) and IV(e), it is seen that Obama has spoken the words of Jordan Peele and used the same expression and the proposed model identifies that the words spoken and the expression by Obama as fake most of the times. Every time Obama utters words and changes expressions, it is detected as warped by the model. The accuracy of the face warped detection rate is 88.23%, which is quite high.

Case II: Obama face detected as real

In Fig IV(f), IV(g), IV(h) and IV(i), it is seen that the face of Obama is detected as real, which is false in very few cases. It is classified real whenever Obama does not utter words or does not change any expression. The rate at which the face is detected real is an average of 12%.

Case III: Jordan Peele face detected as real

In Fig IV(j), IV(k), IV(l) and IV(m), it is seen that the face of Jordan Peele has been detected real on an average of 73In the obtained results, it is clearly seen that the accuracy obtained is better off considering the fact that there are less resources available for face warping detection.

Fig V presents another example of a face warped video taken from the internet. Fig V(a) is the original video. Fig V(b) through V(e) has warping performed in the nose area, mouth area and eye area and the proposed model detects all this warping with a confidence ratio of 93%, 62%, 62% and 96% respectively. Fig V(f) through Fig V(i) display the true face with a confidence ratio of 75%, 66%, 92% and 79% respectively. The model detects face warping in the video with an accuracy of 77%. A bounding box is drawn around the region that is face warped with the confidence ratio displayed. It is seen that the model is able to detect face warping from videos satisfactorily. The proposed model that uses xception CNN is compared with other state-of-the-art CNN methods like Inception-v3, ResNet and VGGNet and the performance is shown in Fig VIII. The observation clearly shows that the proposed model performs better than the other methods.

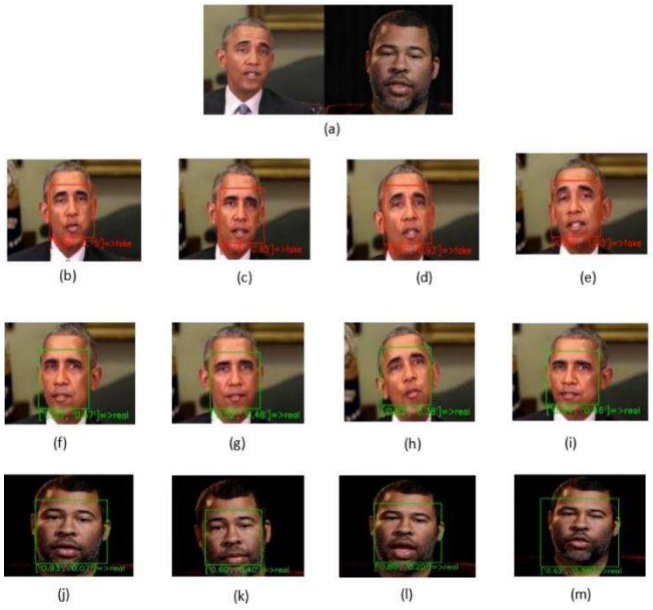

FIG. IV: DETECTED FACE WARPED VIDEO OF OBAMA (A) ORIGINAL VIDEO (B) 75% FAKE (C) 85% FAKE (D) 93% FAKE (E)100% FAKE (F) 53% REAL (G) 52% REAL (H) 62% REAL (I) 54% REAL (J) 93% REAL (K) 60% REAL (L) 80% REAL (M) 62% REAL

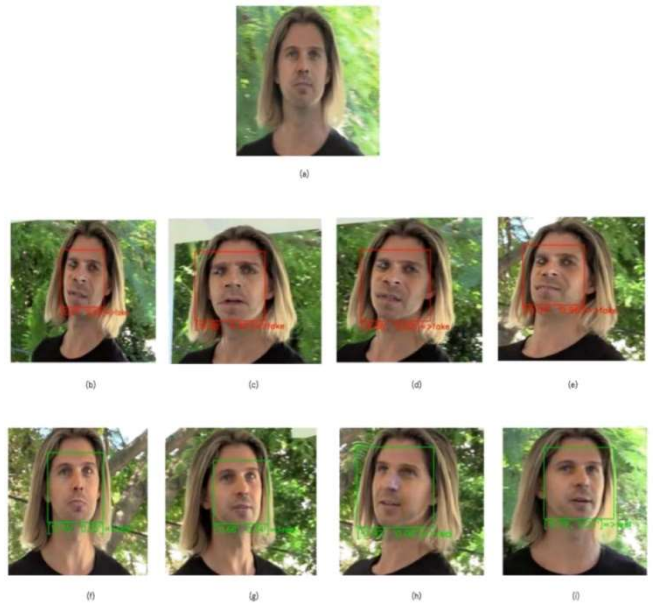

FIG. V: DETECTED FACE WARPED IMAGES FROM VIDEO (A) ORIGINAL VIDEO (B) 93% FACE WARPED (C) 62% FACE WARPED (D) 62% FACE WARPED (E) 96% FACE WARPED (F) 75% REAL FACE (G) 66% REAL FACE (H) 92% REAL FACE (I) 79% REAL FACE.

C. Performance Metrics

The effectiveness of the suggested method is carefully assessed by using a wide variety of metrics to fully gauge its success. These metrics include the confusion matrix, which gives a thorough overview of the classification results; the Area Under the Curve-Receiver Operating Characteristics

(AUC-ROC) [31], which quantifies the technique's discriminative power; accuracy, which quantifies the classification's overall correctness; precision, which gauges the technique's capacity to accurately identify positive instances; and recall, which measures the technique's ability to identify instances that aren't positive. A complete evaluation of the technique's performance is provided by the f-1 score, which achieves a balance between recall and precision and measures the technique's capacity to properly identify negative examples. The performance of the classification algorithm is evaluated, shown visually, and summarised using a table called the confusion matrix. One of the most crucial assessment measures for assessing the effectiveness of the classification model at various threshold values is the AUCROC curve. Key elements in assessing the efficacy of a method include the probability curve known as ROC and the measure of separability or discriminative power known as AUC. It demonstrates how well the model can distinguish between classes. The higher the AUC is, the more precisely the model predicts that the 0 class will be 0 and the 1 class will be 1. ROC curves are shown in the chart TP Rate vs. FP Rate. As a result, the TP Rate is on the y-axis and the FP Rate is on the x-axis, as illustrated in (1) and (2).

$$TP\ Rate = \frac{TP}{TP + FN} \tag{1}$$

$$FP\ Rate = \frac{FP}{FP + FN} \tag{2}$$

Accuracy refers to how close a measured value is to a standard or genuine value [20]. As illustrated in (3), accuracy is given as

$$Accuracy = \frac{TP\ +\ TN}{TP + TN\ +\ FP\ +\ FN} \tag{3}$$

Precision is the close proximity of two or more measures to each other. is called precision as given in (4) [20].

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall (True Positive Rate) is the proportion of successfully completed extubations that are correctly categorized as given in (5) [20].

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

Specificity is a measure of a test's ability to identify genuine negatives as given in (6) [5].

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

f 1 Score represents the balance between precision and recall as specified in (7) [20].

$$f\_1\ Score = \frac{2\ *\ Precision\ *\ Recall}{Precision + Recall} \tag{7}$$

where TP represents the number of true positives, TN represents the number of true negatives, FP represents the number of false positives, and FN represents the number of false negatives.

The confusion matrix for the proposed model is presented in Table III that shows distinct scenarios for the performance of the proposed model. It reveals the number of face warped and real input videos and makes it easy to see if the system is mislabeling the two classes. In the model proposed above, there were 93 cases where the model predicted a video as face warped, 60 cases where the model predicted a real video, and 40 cases where the model predicted a real video as face warped, and that there are 7 cases where the model predicts a video with face warped faces as real.

TABLE III: CONFUSION MATRIX FOR THE PROPOSED MODEL

	Predicted Face Warped	Predicted Real
Actual Face Warped	TN = 93	FP = 7
Actual Real	FN = 40	TP = 60

The proposed method is compared with other state-of-the-art models like Inception-v3, ResNet and VGGNet for the various performance metrics and shown in Fig VI. As observed by the results, Accuracy of the proposed method is 82.5%, Recall is 70%, Specificity is 95%, Precision is 93.33% and f _1 score is 80%. It is inferred that the proposed method can detect face warped videos better than other state-of-the-art models. This is possible as improved CNN Xception model is used which detects face warped videos satisfactorily by applying depthwise separable convolution method. The other methods fail to achieve this efficiency as they are not computationally too heavy. Another main advantage of the proposed method is that the performance metrics are obtained far much better than the other CNN models.

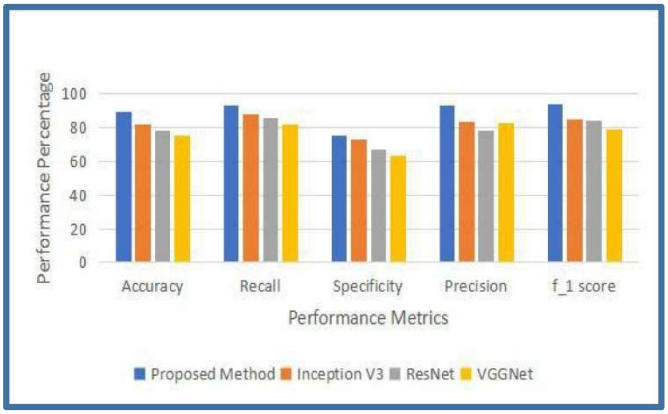

FIG. VI: PERFORMANCE ANALYSIS OF THE PROPOSED MODEL AND OTHER CNN MODELS

The Top-1 and Top-5 accuracy metrics of state-of-the-art CNN models are detailed in great depth in Fig VII. The Top-1 accuracy, which is regarded as the standard accuracy, requires an exact match between the anticipated response and the model's highest probability response. This detailed analysis offers a thorough knowledge of how various CNN models performed in terms of accuracy metrics.

Top-5 accuracy states that any model that provides the 5 most likely outcomes must also produce the desired outcome. In comparison to the prior CNN, the models mentioned in the proposed approach obtain the highest Top-1 precision of 79% and the highest Top-5 precision of 94.5%. Fig 8 shows accordingly, AUROC-based comparison studies between the suggested approach and the state-of-the-art methodologies. AUROC indicates if a model can accurately categorize video. The algorithm performs better at differentiating between authentic and deceptive videos by using AUROC.

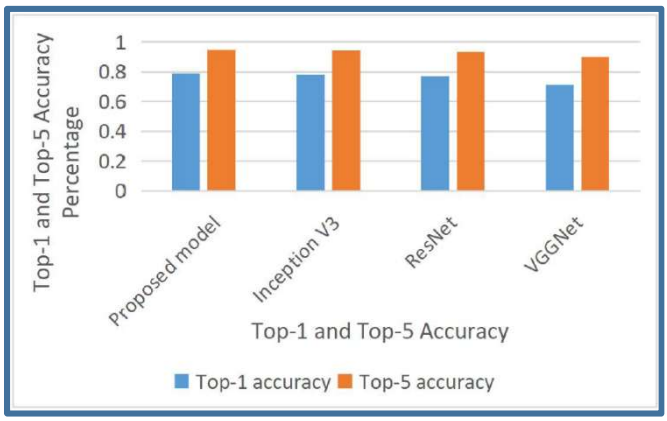

FIG. VII: TOP-1 AND TOP-5 ACCURACY OF THE PROPOSED METHOD AND OTHER CNN MODELS

The region beneath a particular curve is the AUROC. The weakest and best AUROCs are 0.5 and 1, respectively. A useless pattern is corresponding to AUROC 0.5. A subpar performance is one with an AUROC of less than 0.7. The decent performance for AUROC is between 0.70 and 0.80. Superb performance is an AUROC of at least 0.8. A perfect classifier corresponds to an AUROC of 1.

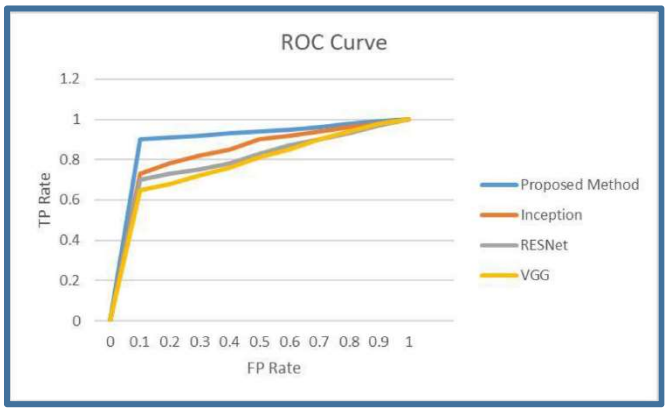

FIG. VIII: ROC CURVE FOR THE PROPOSED MODEL, INCEPTION-V3, RESNET AND VGG

The proposed model gives an AUROC of 0.8, Inception-v3 model produces an AUROC of 0.70, ResNet produces an AUROC of 0.60 and VGG produces an AUROC of 0.55 as shown in Fig VIII. It is clearly seen in the graph that the proposed model outperforms all the other state-of-the-art models by producing a good performance.

V. CONCLUSION

Almost anyone in a commercial enterprise creates a few types of content. As a result of GenAI, their jobs will undergo significant changes, regardless of whether they are working with text, images, hardware designs, music, video, or other media. Face warping on video has become a major challenge in today's real world since it is quite tedious to detect if a face itself is warped. The proposed method is able to detect and localize face warped areas from a given video by employing a transfer learning algorithm. The input target video undergoes video pre-processing techniques by applying different data transformations like color space conversion, image alignment and degradation, illumination normalization. A pre-trained model is generated by feeding the ImageNet dataset into the CNN classifier. The knowledge obtained from this generated pre-trained model is transferred to the new classifier model that is according to the requirements of the proposed work. The processed target video is then fed to the new generated model and classified by using the xception CNN classifier. The classifier is able to identify the face warped region of the video in an efficient manner and achieves a better classification result with good performance as depicted in the performance analysis section. The proposed method is compared with other CNN methods like Inception V3, ResNet and VGGNet using AUROC metrics. The proposed model gives an AUROC of 0.8, which is better than the other models. The results can be further improved by considering strong local variations of expression and luminosity, global pose changes, time shifting, partial occlusion as well as affine transformations. The future scope could be to build an own image processing algorithm which detects local changes in expression, global changes in pose, and affine transformations at arbitrary angles of the face.

ACKNOWLEDGMENT

None.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHORS` CONTRIBUTIONS

All authors have participated in drafting the manuscript. All authors read and approved the final version of the manuscript. All authors contributed equally to the manuscript and read and approved the final version of the manuscript.

CONFLICT OF INTEREST

The authors certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript.

REFERENCES

[1] Chan, C. K. Y., & Zhou, W. (2023). Deconstructing Student Perceptions of Generative AI (GenAI) through an Expectancy Value Theory (EVT)-based Instrument. *arXiv preprint arXiv:2305.01186*.

[2] Younus, M. A., & Hasan, T. M. (2020, April). Effective and fast deepfake detection method based on haar wavelet transform. In *2020*

International Conference on Computer Science and Software Engineering (CSASE) (pp. 186-190). IEEE.

[3] Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., ... & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, *223*, 103525.

[4] Guarnera, L., Giudice, O., Nastasi, C., & Battiato, S. (2020, September). Preliminary forensics analysis of deepfake images. In *2020 AEIT international annual conference (AEIT)* (pp. 1-6). IEEE.

[5] Gass, T., Pishchulin, L., Dreuw, P., & Ney, H. (2011, March). Warp that smile on your face: Optimal and smooth deformations for face recognition. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)* (pp. 456-463). IEEE.

[6] Pishchulin, L., Gass, T., Dreuw, P., & Ney, H. (2011). The fast and the flexible: Extended pseudo two-dimensional warping for face recognition. In *Pattern Recognition and Image Analysis: 5th Iberian Conference, IbPRIA 2011, Las Palmas de Gran Canaria, Spain, June 8-10, 2011. Proceedings 5* (pp. 49-57). Springer Berlin Heidelberg.

[7] Pishchulin, L., Gass, T., Dreuw, P., & Ney, H. (2012). Image warping for face recognition: From local optimality towards global optimization. *Pattern Recognition*, *45*(9), 3131-3140.

[8] Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., ... & Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *electronics*, *8*(3), 292.

[9] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1-11).

[10] Vasist, P. N., & Krishnan, S. (2022). Deepfakes: an integrative review of the literature and an agenda for future research. *Communications of the Association for Information Systems*, *51*(1), 14.

[11] Yang, X., Li, Y., & Lyu, S. (2019, May). Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8261-8265). IEEE.

[12] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3207-3216).

[13] Al-Dhabi, Y., & Zhang, S. (2021, August). Deepfake video detection by combining convolutional neural network (cnn) and recurrent neural network (rnn). In *2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE)* (pp. 236-241). IEEE.

[14] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019, May). Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2307-2311). IEEE.

[15] Saikia, P., Dholaria, D., Yadav, P., Patel, V., & Roy, M. (2022, July). A hybrid CNN-LSTM model for video deepfake detection by leveraging optical flow features. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.

[16] Kumar, M., & Sharma, H. K. (2023). A GAN-based model of deepfake detection in social media. *Procedia Computer Science*, *218*, 2153-2162.

[17] Li, Y., & Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*.

[18] Lin, Y. K., & Sun, H. L. (2023). Few-Shot Training GAN for Face Forgery Classification and Segmentation Based on the Fine-Tune Approach. *Electronics*, *12*(6), 1417.

[19] Olisah, C. C., & Smith, L. (2019). Understanding unconventional preprocessors in deep convolutional neural networks for face identification. *SN Applied Sciences*, *1*(11), 1511.

[20] Nirkin, Y., Masi, I., Tuan, A. T., Hassner, T., & Medioni, G. (2018, May). On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (pp. 98-105). IEEE.

[21] Guo, D., Fraichard, T., Xie, M., & Laugier, C. (2000, October). Color modeling by spherical influence field in sensing driving environment. In *Proceedings of the IEEE Intelligent Vehicles Symposium 2000 (Cat. No. 00TH8511)* (pp. 249-254). IEEE.

[22] Yousefi, J. (2011). Image binarization using Otsu thresholding algorithm. *Ontario, Canada: University of Guelph*, *10*.

[23] Xie, X., Zheng, W. S., Lai, J., Yuen, P. C., & Suen, C. Y. (2010). Normalization of face illumination based on large-and small-scale features. *IEEE Transactions on Image Processing*, *20*(7), 1807-1821.

[24] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258).

[25] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[26] Deng, J. (2009). A large-scale hierarchical image database. *Proc. of IEEE Computer Vision and Pattern Recognition, 2009*.

[27] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. *Advances in neural information processing systems*, *27*.

[28] Dakin, S. C., & Watt, R. J. (2009). Biological “bar codes” in human faces. *Journal of vision*, *9*(4), 2-2.

[29] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).

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

[31] Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, *4*(2), 627.