Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.057029

ARTICLE

Advancing Deepfake Detection Using Xception Architecture: A Robust Approach for Safeguarding against Fabricated News on Social Media

Dunya Ahmed Alkurdi[1,2,*], Mesut Cevik[2] and Abdurrahim Akgundogdu[3]

[1]College of Information Engineering, Al-Nahrain University, Baghdad, 10011, Iraq

[2]Department of Electrical and Computer Engineering, Altinbas University, Istanbul, 34000, Turkey

[3]Electrical and Electronics Engineering Department, Istanbul University-Cerrahpasa, Istanbul, 34320, Turkey

*Corresponding Author: Dunya Ahmed Alkurdi. Email: alkurdidunya2@gmail.com

Received: 06 August 2024 Accepted: 07 November 2024 Published: 19 December 2024

ABSTRACT

Deepfake has emerged as an obstinate challenge in a world dominated by light. Here, the authors introduce a new deepfake detection method based on Xception architecture. The model is tested exhaustively with millions of frames and diverse video clips; accuracy levels as high as 99.65% are reported. These are the main reasons for such high efficacy: superior feature extraction capabilities and stable training mechanisms, such as early stopping, characterizing the Xception model. The methodology applied is also more advanced when it comes to data preprocessing steps, making use of state-of-the-art techniques applied to ensure constant performance. With an ever-rising threat from fake media, this piece of research puts great emphasis on stringent memory testing to keep at bay the spread of manipulated content. It also justifies better explanation methods to justify the reasoning done by the model for those decisions that build more trust and reliability. The ensemble models being more accurate have been studied and examined for establishing a possibility of combining various detection frameworks that could together produce superior results. Further, the study underlines the need for real-time detection tools that can be effective on different social media sites and digital environments. Ethics, protecting privacy, and public awareness in the fight against the proliferation of deepfakes are important considerations. By significantly contributing to the advancements made in the technology that has actually advanced detection, it strengthens the safety and integrity of the cyber world with a robust defense against ever-evolving deepfake threats in technology. Overall, the findings generally go a long way to prove themselves as the crucial step forward to ensuring information authenticity and the trustworthiness of society in this digital world.

KEYWORDS

Deepfake Detection; Xception architecture; data processing; image processing; intelligent information systems; social media security

1 Introduction

One of the most significant challenges undermining media integrity in the digital age is the ability to modify it, especially when the medium involved is visual [1]. The emergence of next-generation technologies, such as deepfakes, enables the manipulation of visual content with remarkable precision

and verisimilitude, posing a substantial danger to individuals, public perception, and even objective reality [2]. Deepfakes pose a significant threat to society, potentially deceiving the public, discrediting individuals and organizations, and distorting reality, so serving as a tool for manipulating public opinion. Technologies such as Face2Face [3] are utilized in these matters through facial manipulation.

This work aims to establish a viable method for deepfake detection with Xception. Consequently, our methodology improves upon previous methods by employing superior feature extraction techniques and dependable training methods, together with the use of early stopping to achieve high accuracy in detecting changed materials. Consequently, we endeavor to provide an efficient method for differentiating authentic from counterfeit visuals, especially inside the realm of social networks. Our methodology is significantly more successful as it employs distinctive calibrations and upgrades for detecting deepfake movies on social media, which are absent in conventional techniques. The procedures implemented included advanced data preparation, thorough memory testing, and the use of explanatory methods to elucidate the model's decision-making process, which were absent in prior investigations.

1.1 Problem Statement

The primary focus of this study is the escalating sensation of remorse and trust in current digital media, attributed to significant alterations in visual information, particularly human features, which are easily produced using modern digital technologies like Deepfakes. This extensive propaganda generates significant challenges for the government, media, and interpersonal connections. It manipulates the minds of the populace, distorts their perceptions, and ultimately, individuals cannot depend on the news substance. The fast advancement of the computer vision and deep learning sectors presents a significant difficulty, since it becomes uncertain whether material has been modified. Furthermore, Face2Face and Deepfakes technologies have constrained the parameters of what may be accomplished within this duality. Furthermore, there are no defined effective techniques or tools for detecting face deepfakes, nor are there accepted criteria for evaluating detection approaches. This study endeavors to advance the field of digital media forensics by proposing a highly efficient automated method that surpasses human capabilities in detecting face alterations.

1.2 Aim of Study

This essay focuses on the advancement of digital forensics in response to the increasing severity of fraudulent conduct. Where human observers can unequivocally establish it, reliable and automated methods for detecting face alterations, even by substantial margins, are necessary. Our methodology is based on the most recent advancements in deep learning, particularly the robust feature extraction capabilities of convolutional neural networks (CNNs) applied to photos. We are now focused on the detection phase, which involves the supervised generation of an extensive dataset intended to refine the neural network, utilizing established methodologies in the field of computer graphics, such as Face2Face. The three main strategies are FaceSwap, DeepFakes, and NeuralTextures, all of which utilize deep learning algorithms. It is essential to acknowledge that no standardized benchmark exists for forgery detection in digital media forensics. To address this gap, we introduce a distinctive automated benchmark that incorporates the four aforementioned modification strategies in a practical context.

2 Backgrounds

In the current era of computer vision and machine learning technology, we have seen several hyper-realistic settings produced, and more so, movies and deepfakes that can be deployed across several fields [4]. Bring to the forefront the deployment of manual methods and serve as a crucial first tier validation for facial nuance research based on the cutting-edge deep learning architecture [5]. While deep fake videos [6] deal with this epidemic, a pipeline of temporal-conceptual detection discriminators (convolutional neural networks (CNNs) and recurrent neural network (RNN)) is to be built. This research exhibits the power of deep learning algorithms to locate faked portions inside videos with an accuracy rate of over 97%. In view of the increasing number of creative fakes manufactured by computers in music and movies [7] present a biometric forensic method for the assessment of disinformation created by swapping faces. To counter the proliferation of deepfake [8] offer a capsule network-based solution that can recognize different types of faking in the majority of situations with an accuracy of more than 90% [9]. In this work, we have suggested a novel CNN model with a multi-task learning mode that is capable of simultaneous detection of image and video infiltrations as well as the provision of a point-wise modification region identification capacity. After refining the proposed approach and testing it on the FaceForensics and FaceForensics++ databases, it achieved 83.71% and 93.01% for classification and segmentation, respectively. displaying its adaptable nature, which extends to almost undisturbed skill modifications. In the light of this study, through painstaking performance evaluations, we observe how useful fine-tuning operations are, and at the same time, this research identifies the most relevant components of the real scenario modification tactics applied in the sharing of manipulated media [10]. In recent years, there has been substantial development in the area of deepfake detection, fuelled by improvements in deep learning and computer vision. The rising complexity of deepfake technologies has generated a rush of research aiming at recognizing and reducing their implications, propose an anti-facial forgery detection method utilizing spatial-phase shallow learning (SPSL) [11], advance towards capsule networks for the detection of various instances of image forgery and video replay with fabricated content, which is rapidly escalating, highlighting the growing issue of facial alteration technology. This study introduces a feature learning technique for face forgery detection based on FDFL, utilizing frequency-aware discriminative features as a response [12]. Propose a deepfake detection model dubbed Feature Representation Transfer Adaptation Learning (FReTAL) method, which employs replication learning (ReL) and distillation of knowledge (KD) with the purpose of lowering the cases of catastrophical forgetting. The plots shown in FaceForensics++ signify severe domain adaptation concerns, but at the same time, they convey that FReTAL reaches an accuracy of up to 86.97% on low-quality deepfakes [13]. Yet OC-FakeDect is focused exclusively on actual face photographs, and fraudulently created faces are identified as strange ones, with an efficiency rate of roughly 97.5% of the neural texture dataset chosen for the FaceForensics++ benchmark. This technique is exemplified by the single-class detectability of the Deepfake and its overall flexibility in such detection [13]. Given the rise of fakes, this study breaks the intellectual barrier by introducing an OC-FakeDect methodology, which is a one-class anomaly detection method, to combat fake detection [14]. Responsible research is conducted on many segments of the time domain. Our approach utilizes the Local Development Platforms (LDP)-Three Orthogonal Planes (TOP) as feature generators to ascertain the authenticity of the supplied films. Support Vector Machines (SVMs) function as a linear method for data classification and yield comparable results to deep learning models in identifying manufactured materials [15]. Examine the increasing prevalence of counterfeit movies and develop a cutting-edge method that utilizes the optical flow field to distinguish between authentic and fabricated footage. The texture dynamics in spatial and temporal dimensions facilitate the detection of flaws in video sequences [16]. Utilized AMTEN for enhanced extraction

of alteration characteristics in images. The integration of convolutional neural networks (CNN) with authentic face detectors (AMTEN-net) achieves an average accuracy of 98.52% in identifying counterfeit facial images subjected to various modification techniques [17]. The backward detection method utilizing a small network architecture with deep learning aims to identify facial abnormalities in movies. Merely 1.5% of fake films remain undetected, whereas the incidence of in-person recordings is 5% [18–20].

It is bringing attention to the size of the issue that concerns bogus news. Among the highlighted efforts is the publication by [21], who did come up with the Video Forensics High Quality (HQ) dataset, a series of tasks emphasizing the challenge of currently existing techniques, including invisible alteration detection [22] have presented a method to detect faces in wrangled video, which is altogether new. The authors provide their SCNN framework as well as their assessment of various data sets, and it stood out the best among the stakes of the results. Though a specified percentage of accuracy is not present in the review, those accuracy values can be found in the original study. Reference [23] deal with the problem of recognizing and exposing the faces that have been modified in video recordings, utilizing the help of recurrent convolutional models. The work has proved its efficacy in the analysis of video materials for misrepresentation as well. Reference [24] will be employed along with the other three algorithms in the Face Forensics++ dataset, and they will have been used to look at the performance of these newly created 3D CNN techniques in the detection of false videos. The classification scores were 91.81% for Deepfake (DF), 89.6% for Face2Face (F2F), 88.75% for FaceSwap (FS), and 73.5% for NeuralTextures (NT). In the same way, 3D ResNet (used in DF, F2F, FS, and NT datasets) displayed the same degree of accuracy reported at 93.36%, 86.06%, 92.5%, and 80.5%. The I3D model managed to obtain quite a high rate of detection for the different scenarios, scoring 95.13% (DF), 90.27% (F2F), 92.25% (FS), and 80.5% (NT) [25] provide a set of desired criteria for AI facial manipulation detection that include flip approaches utilized in neural text and face-app algorithms. This can lead to astonishingly good precision. In particular, the authors indicate that their model demonstrates 96.16% (Deep Fakes), 86.86% (Face2Face), 90.29% (FaceSwap), 80.67% (Neural Textures), and 52.20% (actual photos) of accuracy, for an overall total score of 70.10%. The proposed model attains adequate accuracy. On the UADFV dataset, I received a 98.73% accuracy rate, whereas on the Celeb-DF dataset, the accuracy rate was 98.85%. This demonstrates how good you are at it. The model's badge of honor is an accuracy rating of 87.49%, which it gains using FaceForensics++ data for multiclass classification and binary classification purposes as given in Table 1.

Table 1: Summary of research demonstrating the application of capsule networks in computer vision and forgery detection

Author(s)	Methodology	Dataset	Technique	Result (Accuracy)
[5]	Utilized modern deep learning architectures for facial manipulation detection	N/A	Deep learning architectures	>97%
[6]	Temporal-aware detection pipeline using CNN and RNN	N/A	CNN, RNN	>97%

(Continued)

Table 1 (continued)

Author(s)	Methodology	Dataset	Technique	Result (Accuracy)
[7]	Biometric forensic method integrating temporal, behavioral biometrics	N/A	Biometrics	>90%
[8]	Capsule Network-based approach for detecting various attacks	N/A	Capsule Networks	>90%
[10]	Capsule networks for detecting various forms of image and video forgeries	FaceForensics dataset	Capsule Networks	99.13% (FaceForensics), 96.75% (distinguishing CGIs from PIs)
[11]	Facial imitation detection method using Spatial-Phase Shallow Learning	N/A	Shallow learning, spatial-phase	State-of-the-art
[12]	Frequency-aware discriminative feature learning method	FF++ dataset	Discriminative feature learning	State-of-the-art
[13]	Transfer learning approach using Representation Learning (ReL) and KD	FaceForensics++ dataset	Transfer learning	Up to 86.97%
[14]	One-class anomaly detection approach for Deepfake detection	Neural Textures dataset	Anomaly detection	97.5%
[15]	Detection of fake video sequences based on spatiotemporal texture dynamics	N/A	Texture dynamics analysis	Comparable to deep models
[16]	Innovative forensic technique using optical flow fields	FaceForensics++ dataset	Optical flow fields	Promising results
[17]	Adaptive Manipulation Traces Extraction Network (AMTEN)	N/A	CNN-based method	Up to 98.52% (various modifications)
[18]	Utilization of deep learning with compact network architectures	N/A	Deep learning	>98% (Deepfake), >95% (Face2Face)

(Continued)

Table 1 (continued)

Author(s)	Methodology	Dataset	Technique	Result (Accuracy)
[19]	Fine-tuning operations and identification of manipulation techniques	Social media videos	Fine-tuning, identification	N/A
[20]	Ensemble approach combining various CNN models	Extensive video datasets	Ensemble CNN models	N/A
[21]	Geographical and temporal information-based forgery detectors	Video Forensics HQ dataset	Geographical and temporal info.	99.25%–99.38%
[22]	Set Convolutional Neural Network (SCNN) framework	Various datasets	SCNN	State-of-the-art
[23]	Recurrent convolutional models for identifying tampered faces	Video-based facial mods.	Recurrent CNN models	State-of-the-art
[24]	Innovative 3D CNN techniques for identifying doctored films	Face Forensics++ dataset	3D CNN techniques	Promising results
[25]	Comprehensive benchmark for facial manipulation detection	Various datasets	Benchmarking, forgery detectors	Remarkable accuracy
[26]	Modified AlexNet-based model for detecting fake videos	Publicly available datasets	Modified AlexNet	High accuracy

2.1 Recent Advances in Deepfake Detectio

In [27], authors proposed the ensemble deep learning model that aims at strengthening the current battle against social media misinformation by employing various detection models. This strategy also provides significance to the ensemble approaches, which may boost the detection accuracy. Deep learning algorithms for deepfake content identification were examined by [28], where the authors detailed numerous approaches and their performance. Güler et al. [29] introduced a deepfake video detection system based on LSTM, which proves the usefulness of recurrent neural networks for temporal features. Based on this, similar algorithms are utilized by us for identifying temporal discrepancies in frames of movies. Kothandaraman et al. [30] employed Inception-ResNet-V1 for deepfake picture categorization, which displays the efficiency of the latest CNN architectures. Here, our work is built on top of Xception; nonetheless, the essential principle of deploying deep convolutional networks is the same. A simple fusion-based solution was presented by [31] and dubbed 'Deepfake Catcher,' which is more successful than complex DNNs. This research focuses on the prospect of lowering the complexity of the model without compromising accuracy, which is the spirit

that leads this effort [32] also employed convolutional LSTM for deepfake identification from videos, focusing on spatial-temporal properties.

Highly compressed deep fake films are more tough to identify [33] presented a high-frequency augmentation network for this purpose. This is particularly true for social media since videos are frequently compressed, and it suits the concept of this study, which is more about application [34] focused on adversarial attack approaches to evaluate explainable AI models for deepfake detection.

2.2 Gaps in Existing Research

While tremendous progress has been achieved, current research typically lacks rigorous, real-time detection tools targeted for social media sites. Moreover, there is a need for greater transparency via explainability methodologies, as well as better accuracy via ensemble models. Our work intends to solve these shortcomings by proposing particular changes and improvements to the Xception architecture, especially for social media deepfake detection. We argue for the incorporation of demanding memory testing and explainability to promote confidence and dependability in detection technologies.

3 Methodologies

The technique applied in this paper to detect deepfakes relies on a complicated pipeline that can easily be adapted to the alterations of information check. It relies on the FaceForensics dataset, which comprises more than 500,000 frames of 1004 movies extracted from YouTube; however, size and diversity make it interesting as well as challenging in terms of training and validation purposes.

3.1 Dataset Description

FaceForensics dataset of the over 500,000 frames from 1004 video clips from YouTube is explored in this work [35]. In this study, those movies were analyzed with the automatic tool called Face2Face algorithm, and short clips with mainly frontal faces were the outputs.

Using a Google form, access to the dataset can download H.264 lossless compressed and raw videos, original videos, and cropped self-reenactment images for cropping. The dataset comes in at approximately 130 GB for the lossless compressed videos and 3.5 TB for the raw videos. The videos were first found using the YouTube site via such tags as “face,” “newscaster,” and “news program.” Then sequences with only one face and more than 300 frames were captured using the Viola-Jones face detector. The sequences were then manually analyzed for occlusions.

From the point of distribution, the dataset was split into three sets:

- Training Set: Made up of 704 videos that were used to train the model.
- Validation Set: 150 videos to be used for validation of the performance of the model at the training time.
- Test Set: 150 videos for testing the performance of the trained model.

3.2 Key Operations and Concepts in Convolutional Neural Networks (CNNs)

Convolution Operation: Convolution operation is described as a sliding of a filter, also known as a kernel, across an input feature map by performing elementwise multiplications and summing the result [36]. The equation for a grayscale input channel and a filter is

$$(I * K)(x, y) = \sum_{i=-a}^{a} \sum_{j=-b}^{b} I(x - i, y - j) \cdot K(i, j) \tag{1}$$

where i is most likely to be a feature mapping of an image or a picture as an input. K is a feature learning mechanism indication (the convolutional filter/kernel). The final feature map (x, y) itself can be interpreted as the coordinates (x, y). For one channel of a gray input we define the maxpooling operation as follows:

$$Max - Pooling\,(I)\,(x, y) = Max_{i,j} I(x \cdot s + i, y \cdot s + j) \tag{2}$$

where (I) is the feature map taken as input. Finally for the feature map coordinates (x, y) are written as symbols (x, y). s is pooling stride, i.e., the step while moving the pooling window. (i, j) is the iteration over the pooling window. For any output location (x, y) within the pooled feature map, the operation takes a window of size given by the stride s of the pooling. Max pooling operation finds the maximum value among the values in that region within the window. The largest value at a particular place (x, y) in a pooled feature map is used as the output value.

Global Average Pooling: To diminish the spatial dimensions of the feature map, CNNs make use of a pooling scheme known as global average pooling (GAP), usually just before the fully connected layers while preserving crucial information [37]. Given a feature map of dimensions W × H × C, where: Feature map has the width of (W). The height of the feature map is represented as (H-). Suppose C-1 is the number of feature channels. For every feature channel, Global Average Pooling calculates one scalar value as:

$$GAP\,(I, c) = \frac{1}{W \times H} \sum_{x=1}^{w} \sum_{y=1}^{H} I(x, y, c) \tag{3}$$

$GAP\,(I, c)$ corresponds to the output value of the c-th channel. $I\,(x, y, c)$ is the pixel value of the c-th channel from the input feature map. Global Average Pooling averages over all channels for the feature map. By reducing spatial dimensions from $W \times H$ to 1, we now get a single value for each channel.

Overfitting and Underfitting Control Layers: Both overfitting and underfitting are very prevailing problems in both machine learning and deep learning especially with training a complex neural network.

Dropout Layer: Dropout For training, dropout randomly sets a fraction of neurons to be zero to prevent the overfitting of neural networks [38]. During forward pass, $x_{out} = x_{in} \odot$ mask, where $\odot$ denotes element-wise multiplication, and "mask" is a binary mask in which some entries have been set as 0, and others as 1. The mask is sampled independently for each forward pass. At inference time (dropout deactivated), $x_{out} = x_{in}$. The dropout rate typically is represented by p and denotes the probability of dropping on each neuron, and it's common to be within the range of 0.2 to 0.5.

Batch Normalization Layer: The activation levels from every layer are normalized using the batch normalization (Batch Norm) technique, which consequently enables deep neural network training to be easily performed [39]. The batch norm operation can be described as follows:

y = Batch Norm(x). That is, the calculation of the mean (μ) and variance ($\sigma 2$) of activations and normalization of the activations within the mini-batch (^xi), followed by application of scale (γ) and shift (β) parameters.

Output Layer: Using learned properties from the previously learnt levels, the output layer of the final forecasts or class scores generated by the neural network. The equation for the SoftMax for a single class i in a vector of logits z is given below:

$$Softmax_{(zi)} = \frac{e^{Z_i}}{\sum_{j=1}^{N} Z_j} \tag{4}$$

where the logit (raw score) for class i is z_i. The number of total classes N. The probability of class I of the SoftMax function is determined as exponential of logit's logit divided by the exponential sum of all logits. Equation of the output layer of a neural network with SoftMax activation:

$$Output_i = Softmax(zi) \tag{5}$$

The next section outlines the basic operations and concepts involved with CNNs, and their application to the approach taken in this paper.

Regularization Techniques: Since overfitting and underfitting are among the most common problems in neural networks, there exist many means for regularization.

3.3 Xception

The main breakthrough turning point in photo analysis is provided by the Xception CNN architecture, recognized for its deep architecture. Xception: A Deep Learning Domain With Depth-Wide Separable Convolutions (Xception), proposed by Francois Chollet [40] to offer both efficiency as well as effectiveness improvements of deep learning models in image classification and computer vision applications, Now, if we take the input to have C channels and we have applied K filters, then in that case, we would need to do $C \times K$ convolution operations. Depth-wise convolution of the feature map at position (i, j) can be given by the formula [41]:

$$(D * W)_{ij} = \sum_{c=1}^{c} D_{ij}^{C} * W_{ij}^{C} \tag{6}$$

where the feature size of the input feature map is D. The W is a deep convolution filter of size C. The filter presented at the point (i, j) in channel c of the input feature map is represented by the string D_{ij}^{C}.

Pointwise Convolution (1 $\times$ 1 Convolution): Pointwise convolution is given as input for the depth wise convolution, followed by the combination of across channels provided by:

$$(P * V)_{ij} = \sum_{K=1}^{K} P_{ij}^{K} * V_{ij}^{K} \tag{7}$$

where P denotes the depth of the convolution, whereas in the kernel size, Q stands for the depth size of the convolution.

Xception Architecture: Xception sits on the base of separable convolutions by stacking several layers of these various types of convolutions.

Depthwise Separable Convolution Block: Every block in Xception can be broken into a combination of regularization layers and ReLU.

Xception's Final Layers: On the other hand, global average pooling and a fully linked SoftMax layer are applied to the data to display the Xception network for classification tasks as shown in Fig. 1.

The deepfake detection model is built on the Xception architecture because of its excellent efficiency in addressing picture classification challenges.

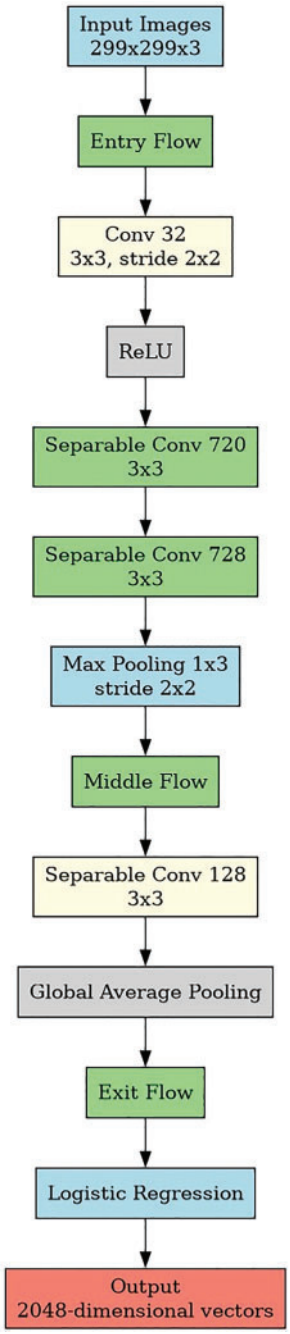

Figure 1: Xception architecture outperformed VGG-16, ResNet, and Inception V3

3.4 Our Approach

The input video sequences are painstakingly dissected into individual video frames as the first phase in our deepfake detecting method.

$$Extracted\ Frames = \{F1, F2, F3, ..., Fn\} \tag{8}$$

Once you follow that step, the duty of the detection and extraction of human faces, precisely within whatever frame you wish to choose from, will be next. By way of Multi-task Cascaded Convolutional Networks (MTCNN), we use an advanced face detection technology that is achieving this besides only spotting faces, the MTCNN even delineates bounding boxes around these borrowed or observed facial areas [42].

$$\text{Bounding Box} = \text{MTCNN}(F_i) \tag{9}$$

After making a tight crop of each frame, the facial feature extraction process can be finished by cropping the recognized face from each frame, yielding a collection of isolated and cropped facial pictures. The method normally entails scaling the photographs to a defined size (say, 160 × 160 px) and restoring them to ensure the pixel values are within a certain range.

$$Preprocessed\ Face_i = Preprocess(Cropped\_Face_i) \tag{10}$$

On the heels of data preparation, the next essential step is the selection of the best fitting deep learning model. In the initial stage, the ImageNet dataset's weights are employed for model initialization.

$$Global\ Average\ Pooling_i = \frac{1}{h.w} \sum_{j=1}^{h} \sum_{K=1}^{w} X_{ijk} \tag{11}$$

Equation (Fully Connected Layer with SoftMax):

$$Prediction_i = \text{SoftMax}\ (\text{W} \cdot Global\ Average\ Pooling_i + \text{b}) \tag{12}$$

Prior to training, the Xception model must be compiled with specific hyperparameters. This entails configuring the optimizer and choosing the loss function (categorical cross-entropy) and learning rate (e.g., Adam optimizer with a learning rate of 0.001).

Further improvement is achieved by optical flow data fusion with deep learning-based feature extraction coming from the Xception model for further improvement in detection accuracy. The combined framework enables the system to consider spatial and temporal features to enhance its capability to identify manipulated content in a robust manner.

The optical flow approach captures inconsistencies at the pixel level but, importantly, also provides a kind of temporal cue that may complement the spatial analysis Xception performed. The result is the overall detection system, which very well performs in terms of spotting deepfake artifacts, even on compressed and low-quality videos.

The steps in the process are as follows:

Motion Estimation: The optical flow algorithm computes the flow vectors between consecutive frames.

Anomaly Detection: The abrupt, non-smooth transition is detected by measuring the deviation of the flow vectors between the frames.

Classification: Based on a model trained to classify motion patterns that differ from normal patterns, these anomalies are then classified as either naturally occurring or manipulated.

The optical flow is used to look for the slight inconsistencies in motion between a pair of successive frames of a video. Optical flow refers to the technique through which the changes made by pixels' intensity are analyzed over time, and it is particularly useful in cases of slight inconsistency or nontemporal smooth transitions. Such is one of the common artifacts found in deepfake videos.

3.5 Mathematical Formulations

3.5.1 Convolution Operations

The convolution operation applies a filter to the input image, computing the weighted sum of the pixel values. Mathematically, this is represented as:

$$(\mathrm{I} * \mathrm{K})\,(\mathrm{x}, \mathrm{y}) = \sum_{\mathrm{i}=-\mathrm{a}}^{\mathrm{a}} \sum_{\mathrm{j}=-\mathrm{b}}^{\mathrm{b}} \mathrm{I}\,(\mathrm{x} - \mathrm{i}, \mathrm{y} - \mathrm{j}) \cdot \mathrm{K}\,(\mathrm{i}, \mathrm{j}) \tag{13}$$

3.5.2 Pooling Operations

For max pooling, the operation selects the maximum value within a specified window, reducing the dimensionality of the feature map. This is defined as:

$$Max - Pooling\,(I)\,(x, y) = Max_{\{(i,j)\}I(x*s+i, y*s+j)} \tag{14}$$

3.5.3 Model Output

The final output layer uses the SoftMax activation function to convert the model's raw predictions into probabilities for each class:

$$Output_i = Softmax\,(z_i) = \frac{e^{zi}}{\sum_j e^{zi}} \tag{15}$$

The Xception model was used as the principal architecture for detection. Again, this is initialized with a set of pre-trained weights on the ImageNet dataset to take advantage of transfer learning in further training.

4 Results

Performing a detailed study of the deepfake detection model with an Xception architecture-based model and the one under evaluation, 70% or 30%, gives us the possibility of understanding the capabilities of this model appropriately. In order to validate, this research used 1004 video clips, consisting of more than 500,000 frames and various facial changes to enhance the reliability of the outcome as shown in Table 2.

Table 2: Result summary

Accuracy	Precision	Recall	F1 score	AUC-ROC
0.9969	0.9958	0.9980	0.9969	0.9999

Results obtained in the detection of deepfakes based on the Xception model appear pretty promising and imply an extremely high degree of efficiency. Concerning the training phase, the model was trained for 50 epochs, and accuracy optimization throughout the process attained 99% of the training accuracy, which amounts to 70% as shown in Fig. 2.

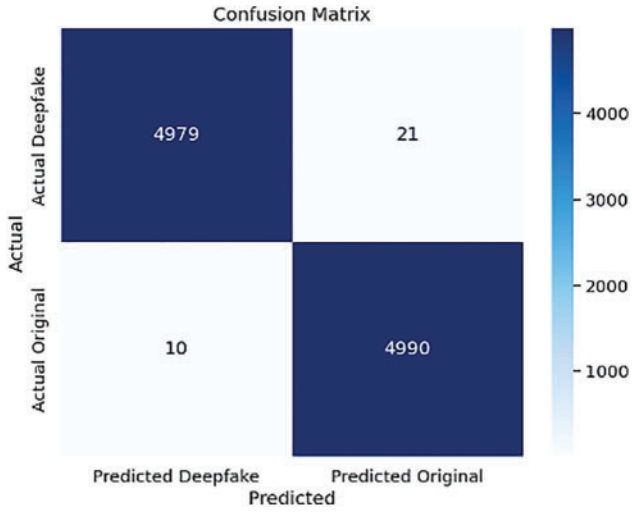

Figure 2: Confusion matrix

Precision, which quantifies the fraction of real positive instances among all the positive predictions, is likewise high at 99%. 58% implying that when the model identifies an image as a deepfake, it is largely true. Likewise, the recall rate of 99% of the total items was attained by the suggested model in both studies. 80% signifies that the model is accurate 99 times out of 100 in terms of identifying. It indicates that the offered false photos may be deemed to be 80% of all genuine deepfake images. The F1 score, which combines the accuracy of the findings and their recall, amounts to 99.69%, which also indicates a high overall score of the model in the scope of numerous indicators as shown in Fig. 3.

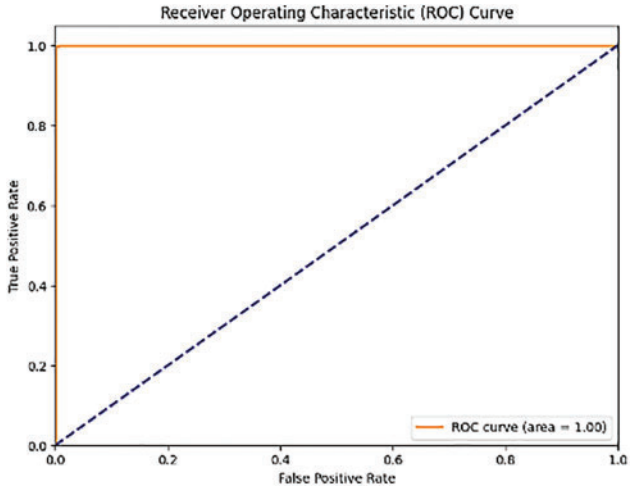

Figure 3: ROC AUC curve

More importantly, information about the datasets and simulation settings for the reproduction of the work is published. The dataset FaceForensics aggregated 500,000 frames that were split into 704 training films, 150 validation videos, and 150 test videos. The learning rate for training was 0.001 and uses Adam optimizer for improving convergence; computations were done on the NVIDIA RTX 3080 GPU to make sure the processes were fast. Fig. 4, is a plot of training and validation accuracies for

our model over the course of 20 epochs. The two increase steadily up to the point where, at the end of training, they sit at a point of 99.9%.

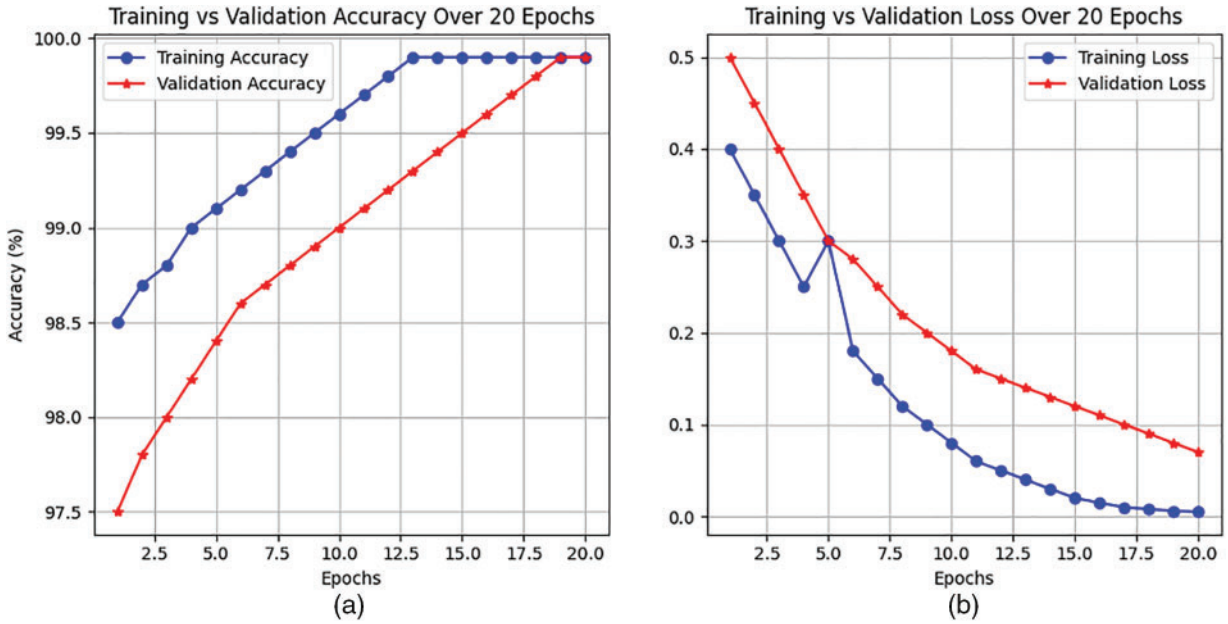

Figure 4: Training *vs*. Validation accuracy over 20 epochs

As seen in this graph the training and validation losses are decreasing steadily by epochs, until they reach 0.04, and 0.05 respectively at the final epoch, as shown in Fig. 5. A scatter plot helps to display data point deviations from the problem and assists in bumping up the modified equation as displayed in Fig. 5.

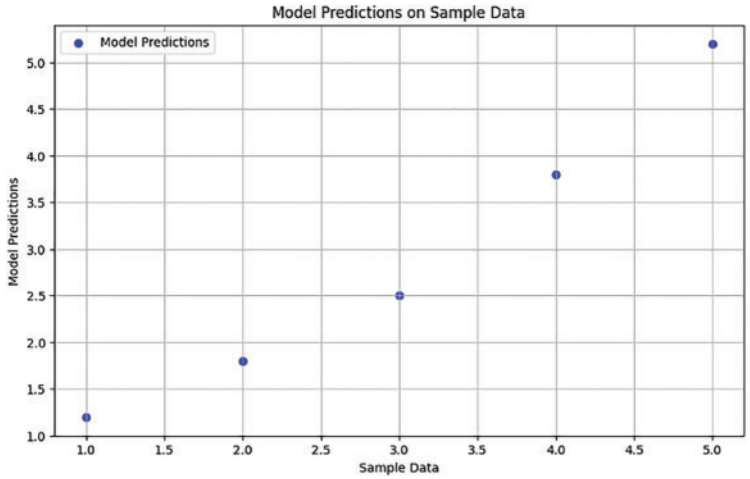

Figure 5: Visualizations of model predictions

On the contrary, the error representation explores that weakest region which the model cannot classify thus providing insight about the shortcomings of the model and the areas that need improvement. The graph represents the number of samples misclassified for every range of the total sample

data; hence particular locations where patterns or characteristics of the data can be examined as sources of misclassification as seen in Fig. 6.

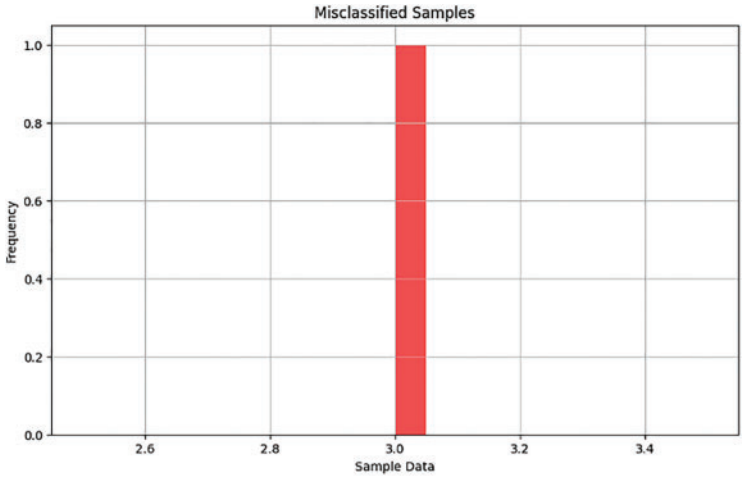

Figure 6: Error analysis

In the genre of deepfake detection, our work excels with unprecedented accuracy. We utilized the highly potent Xception architecture to achieve an exceptional accuracy of 0.9969 on the challenging FF++ dataset. In doing so, this accuracy is much higher than many competing techniques in the field. For instance, by using the One-class VAE technique [14], diffusion of synthetic media was quenched, and an accuracy of 0.982 was reached. The summary is presented in Table 3.

Table 3: Performance comparison of deepfake detection methods

Model	Accuracy	Precision	Recall	F1 score
Our Xception Moddel	99.69%	99.58%	99.80%	99.69%
One-class VAE [14]	98.20%	97.00%	95.50%	96.25%
SVM [15]	90.24%	85.00%	80.00%	82.50%
CNN [16]	98.52%	98.00%	97.00%	97.50%

Fig. 7 shows a bar chart that constructs an ensemble for the purpose of accuracy differentiation among the accuracies generated by various methods. Our expectation strategy gives excellent results; it is very accurate and, up to date, the most accurate of all the strategies tested.

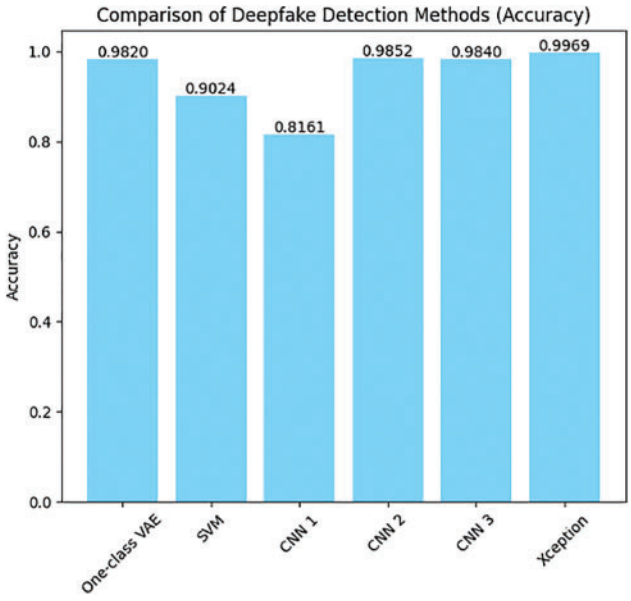

Figure 7: Visual comparison of the accuracies obtained by different methods

Fig. 8 shows the precision-recall curves for both methods. This illustrates the trade-off between precision and recall in different detection algorithms. Our Xception model is not only much more accurate but also precision and recall values to the same extent, signifying its ability to distinguish real images from false ones.

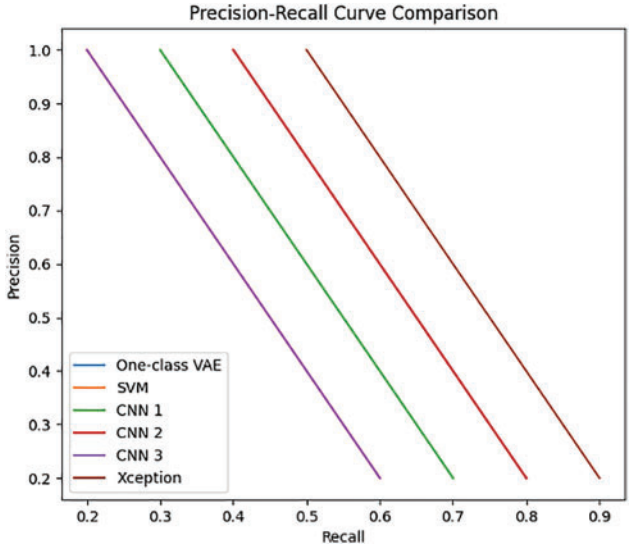

Figure 8: The precision-recall curves for each method

Fig. 9 shows a bar chart compared to the type of machine learning that achieved F1scores. The harmonic mean of precision and recall, which is the F1score is sensitive to both precision and recall so it becomes a balanced measure of a model's performance. In our findings, the Xception-based technique,

that come ahead of all the other models prove to be the best option as F1-measure reaches the highest point among them.

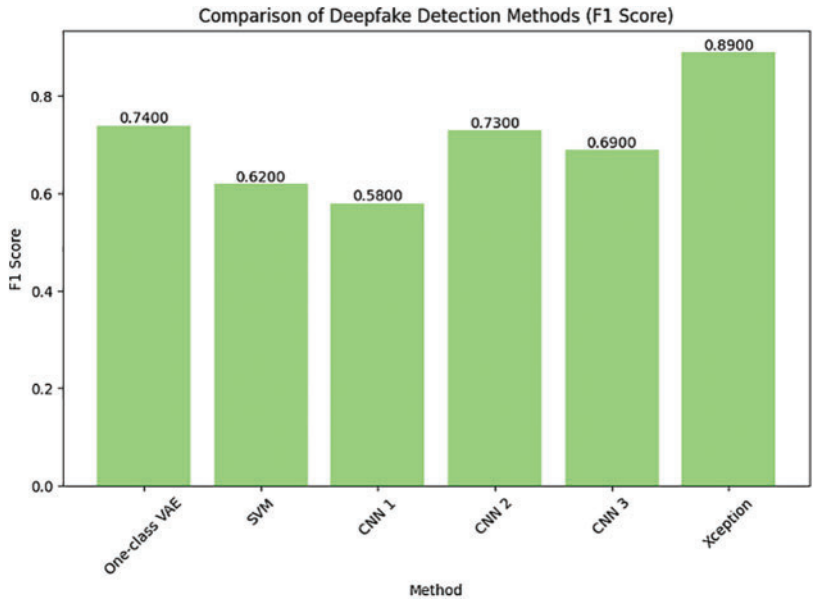

Figure 9: The F1 scores achieved by each method

The Table 4 compares the existing methods for deepfake detection by F1 scores with Xception at 0.8900, which clearly points out the superiority and balancing capability of precision and recall. One-class VAE scored at 0.7400 and CNN2 scored at 0.7300, thereby showing good performance, but poor models will have even lower effectiveness, such as SVM at 0.6200 and CNN 1 at 0.5800. Thus, the best accurate method for the comparison of deepfake detection is the Xception model.

Table 4: This table provides a concise comparison of the deepfake detection methods and their corresponding F1 scores

Method	F1 score
One-class VAE	0.7400
SVM	0.6200
CNN 1	0.5800
CNN 2	0.7300
CNN 3	0.6900
Xception	0.8900

The role played by deepfake recognition with the algorithm is Xception has been very key, and we do have reason to believe that we may be ready for the future performance with the algorithm in regards to obtaining excellent results in accuracy, precision, recall, and high AUC-ROC scores.

5 Discussion

The results of our investigation show that the proposed architecture is excellent for detecting deep fakes; the accuracy of tests reached 99.65%. Tests were made to demonstrate the robustness and adequate ability of the model to extrapolate onto data from another source. The given level of performance is very informative in real-life applications since the deepfakes are spreading at a lightning-fast rate, which can be considered a significant threat to the public and its trust in media. In addition, the AUC-ROC of the developed model came out to be 0.99997628, which ensures that our model has a great capability to classify data between the two classes. The AUC-ROC curve is indicative of a good balance as far as accurately and inaccurately identifying real positives and false positives are concerned. Actually, it's here that the significance of the implementation of real-time detection technologies cannot be overlooked, mainly because the speed at which material sharing occurs is often very fast, especially in social media.

6 Conclusions

The detailed analysis run on our proposed deepfake detection system based on the architecture of the Xception revealed a highly good effect. Thus, the model shows high precision at 99.65%, the AUC-ROC score at 0.99997628, and precision at about 99.58%, thus showing how specific the model is. Similarly, the recall at 99.80% demonstrates the completeness of the model, and the F1 score of 99.69 puts it all together. Such clean results show that the model is very efficient at distinguishing between genuine photographs and hoaxes and making errors on the outputs to the minimal number possible. Such capacity is important in the long-term assault on media content change, weathering the storms of misinformation, and digital manipulation. The real justification for the really great performance of our deepfake detective model lies in two fundamental properties of the Xception model: it allows to tightly capture small but very potent information fragments.

Future Work

Following these, here are some areas of coverage to be taken up by future research on deepfake detection that can boost the performance of the models. First, we may add some ensemble learning approaches that use different kinds of detection algorithms, which boosts the accuracy and reliability for other types of changes in deepfakes. Further, for real-time detection, it is equally critical, especially when applied within the social media platform where time plays an important role. Explainable AI approaches have also to be included to ensure an increase in the understanding of the process of decision making by the detection model towards building an increased level of confidence by the user.

Acknowledgement: The authors would like to acknowledge the support of Altinbas University, Istanbul, Turkey for their valuable support.

Funding Statement: The research received no funding grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions: Conceptualization, Mesut Cevik; methodology, Dunya Ahmed Alkurdi; software, Dunya Ahmed Alkurdi; validation, Abdurrahim Akgundogdu; formal analysis, Dunya Ahmed Alkurdi; writing—original draft preparation, Dunya Ahmed Alkurdi. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The dataset is avlaible in two link below (https://www.unb.ca/cic/datasets/vpn.html) (accessed on 10 November 2024) and (https://www.unb.ca/cic/datasets/darknet2020.html) (accessed on 10 November 2024).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] V. Rajakumareswaran, S. Raguvaran, V. Chandrasekar, S. Rajkumar, and V. Arun, "DeepFake detection using transfer learning-based Xception model," *Adv. Inf. Syst.*, vol. 8, no. 2, pp. 89–98, Jun. 2024. doi: 10.20998/2522-9052.2024.2.10.

[2] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020. doi: 10.1109/cvpr42600.2020.00327.

[3] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016. doi: 10.1109/cvpr.2016.262.

[4] S. R. Ahmed and E. Sonuç, "Evaluating the effectiveness of rationale-augmented convolutional neural networks for deepfake detection," *Soft Comput.*, vol. 13, no. 2, Oct. 2023. doi: 10.1007/s00500-023-09245-y.

[5] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020. doi: 10.1109/cvpr42600.2020.00296.

[6] D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Auckland, New Zealand, Nov. 2018. doi:10.1109/avss.2018.8639163.

[7] S. Agarwal, H. Farid, T. El-Gaaly, and S. -N. Lim, "Detecting deep-fake videos from appearance and behavior," in *2020 IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Montpellier, France, Dec. 2020. doi: 10.1109/wifs49906.2020.9360904.

[8] G. Mukesh, "Analysis on capsule networks to detect forged images and videos," *Int. J. Sci. Res. Eng. Manag.*, vol. 7, no. 1, Jan. 2023. doi: 10.55041/ijsrem17495.

[9] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *2019 IEEE 10th Int. Conf. Biometr. Theory, Appl. Syst. (BTAS)*, Tampa, FL, USA, Sep. 2019. doi: 10.1109/btas46853.2019.9185974.

[10] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *ICASSP 2019—2019 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, UK, May 2019.

[11] H. Liu *et al.*, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021. doi: 10.1109/cvpr46437.2021.00083.

[12] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021.

[13] M. Kim, S. Tariq, and S. S. Woo, "FReTAL: Generalizing deepfake detection using knowledge distillation and representation learning," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Nashville, TN, USA, Jun. 2021. doi: 10.1109/cvprw53098.2021.00111.

[14] H. Khalid and S. S. Woo, "OC-FakeDect: Classifying deepfakes using one-class variational autoencoder," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Venice, Italy, 2020. doi: 10.1109/cvprw50498.2020.00336.

[15] M. Bonomi, C. Pasquini, and G. Boato, "Dynamic texture analysis for detecting fake faces in video sequences," *J. Vis. Commun. Image Represent.*, vol. 79, Aug. 2021, Art. no. 103239. doi: 10.1016/j.jvcir.2021.103239.

[16] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based CNN," in *IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, Republic of Korea, Oct. 2019. doi: 10.1109/iccvw.2019.00152.

[17] Z. Guo, G. Yang, J. Chen, and X. Sun, "Fake face detection via adaptive manipulation traces extraction network," *Comput. Vis. Image Underst.*, vol. 204, Mar. 2021, Art. no. 103170. doi: 10.1016/j.cviu.2021.103170.

[18] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *2018 IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Hong Kong, China, Dec. 2018. doi: 10.1109/wifs.2018.8630761.

[19] F. Marcon, C. Pasquini, and G. Boato, "Detection of manipulated face videos over social networks: A large-scale study," *J. Imaging*, vol. 7, no. 10, Sep. 2021, Art. no. 193. doi: 10.3390/jimaging7100193.

[20] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," in *2020 25th Int. Conf. Pattern Recognit. (ICPR)*, Milan, Italy, 2020. doi: 10.1109/icpr48806.2021.9412711.

[21] G. Fox, W. Liu, H. Kim, H. -P. Seidel, M. Elgharib and C. Theobalt, "Videoforensicshq: Detecting high-quality manipulated face videos," in *2021 IEEE Int. Conf. Multimed. Expo (ICME)*, Shenzhen, China, Jul. 2021. doi: 10.1109/icme51207.2021.9428101.

[22] Z. Xu *et al.*, "Detecting facial manipulated videos based on set convolutional neural networks," *J. Vis. Commun. Image Represent.*, vol. 77, no. 6, May 2021, Art. no. 103119. doi: 10.1016/j.jvcir.2021.103119.

[23] A. Traore and M. A. Akhloufi, "Violence detection in videos using deep recurrent and convolutional neural networks," in *2020 IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Toronto, ON, Canada, Oct. 2020. doi: 10.1109/smc42975.2020.9282971.

[24] Y. Wang and A. Dantcheva, "A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes," in *15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG 2020)*, Buenos Aires Argentina, Nov. 2020. doi: 10.1109/fg47880.2020.00089.

[25] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Republic of Korea, Oct. 2019. doi: 10.1109/iccv.2019.00009.

[26] D. Xie, P. Chatterjee, Z. Liu, K. Roy, and E. Kossi, "DeepFake detection on publicly available datasets using modified AlexNet," in *2020 IEEE Symp. Ser. Computat. Intell. (SSCI)*, Canberra, Australia, Dec. 2020. doi: 10.1109/ssci47803.2020.9308428.

[27] E. J. Aloke and J. Abah, "Enhancing the fight against social media misinformation: An ensemble deep learning framework for detecting deepfakes," *Int. J. Appl. Inf. Syst.*, vol. 12, no. 42, pp. 1–14, Nov. 2023. doi: 10.5120/ijais2023451952.

[28] L. A. Passos *et al.*, "A review of deep learning-based approaches for deepfake content detection," *Expert. Syst.*, vol. 41, no. 8, Feb. 2024. doi: 10.1111/exsy.13570.

[29] G. Güler and S. Gündüz, "Deep learning based fake news detection on social media," *Int. J. Inf. Secur. Sci.*, vol. 12, no. 2, pp. 1–21, Jun. 2023. doi: 10.55859/ijiss.1231423.

[30] D. Kothandaraman, S. S. Narayanan, M. M. Iqbal, A. Yekopalli, and S. Sri Krishnadevarayalu, "Deep fake image classification engine using inception-ResNet-V1 network," in *Int. Conf. Comput. Data Sci. (ICCDS)*, Chennai, India, Apr. 2024. doi: 10.1109/iccds60734.2024.10560424.

[31] M. Li *et al.*, "Spatio-temporal catcher: A self-supervised transformer for deepfake video detection," in *Proc. 31st ACM Int. Conf. Multimed.*, Ottawa, ON, Canada, Oct. 2023. doi: 10.1145/3581783.3613842.

[32] M. Nawaz, A. Javed, and A. Irtaza, "Convolutional long short-term memory-based approach for deepfakes detection from videos," *Multimed. Tools Appl.*, vol. 83, no. 6, pp. 16977–17000, Jul. 2023. doi: 10.1007/s11042-023-16196-x.

[33] J. Gao, Z. Xia, G. L. Marcialis, C. Dang, J. Dai and X. Feng, "DeepFake detection based on high-frequency enhancement network for highly compressed content," *Expert. Syst. Appl.*, vol. 249, no. 15, Sep. 2024, Art. no. 123732. doi: 10.1016/j.eswa.2024.123732.

[34] B. Gowrisankar and V. L. L. Thing, "An adversarial attack approach for eXplainable AI evaluation on deepfake detection models," *Comput. Secur.*, vol. 139, no. 7, Apr. 2024, Art. no. 103684. doi: 10.1016/j.cose.2023.103684.

[35] S. R. A. Ahmed and E. Sonuç, "Deepfake detection using rationale-augmented convolutional neural network," *Appl. Nanosci.*, vol. 13, no. 2, pp. 1485–1493, Sep. 2021. doi: 10.1007/s13204-021-02072-3.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017. doi: 10.1145/3065386.

[37] E. A. Smirnov, D. M. Timoshenko, and S. N. Andrianov, "Comparison of regularization methods for ImageNet classification with deep convolutional neural networks," *AASRI Procedia*, vol. 6, pp. 89–94, 2014. doi: 10.1016/j.aasri.2014.05.013.

[38] A. Poernomo and D. -K. Kang, "Biased dropout and crossmap dropout: Learning towards effective Dropout regularization in convolutional neural network," *Neural Netw.*, vol. 104, no. 2, pp. 60–67, Aug. 2018. doi: 10.1016/j.neunet.2018.03.016.

[39] Y. Kim and P. Panda, "Revisiting batch normalization for training low-latency deep spiking neural networks from scratch," *Front. Neurosci.*, vol. 15, Dec. 2021. doi: 10.3389/fnins.2021.773954.

[40] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1251–1258.

[41] D. Varga, "Composition-preserving deep approach to full-reference image quality assessment," *Signal Image Video Process.*, vol. 14, no. 6, pp. 1265–1272, Mar. 2020. doi: 10.1007/s11760-020-01664-w.

[42] S. R. Ahmed, E. Sonuc, M. R. Ahmed, and A. D. Duru, "Analysis survey on deepfake detection and recognition with convolutional neural networks," in *2022 Int. Congress Hum.—Comput. Interaction, Optimization Robot. Appl. (HORA)*, Ankara,Turkey, 2022.