Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048238

ARTICLE

Customized Convolutional Neural Network for Accurate Detection of Deep Fake Images in Video Collections

Dmitry Gura[1,2], Bo Dong[3,*], Duaa Mehiar[4] and Nidal Al Said[5]

[1]Department of Cadastre and Geoengineering, Kuban State Technological University, Krasnodar, 350072, Russian Federation

[2]Department of Geodesy, Kuban State Agrarian University, Krasnodar, 350072, Russian Federation

[3]School of Innovation and Entrepreneurship, Liaoning University, Liaoning, 110031, China

[4]Department of IT, Middle East University, Amman, 11831, Jordan

[5]College of Mass Communication, Ajman University, P.O Box 346, Ajman, United Arab Emirates

*Corresponding Author: Bo Dong. Email: dongbo@lnu.edu.cn

Received: 01 December 2023 Accepted: 22 February 2024 Published: 15 May 2024

ABSTRACT

The motivation for this study is that the quality of deep fakes is constantly improving, which leads to the need to develop new methods for their detection. The proposed Customized Convolutional Neural Network method involves extracting structured data from video frames using facial landmark detection, which is then used as input to the CNN. The customized Convolutional Neural Network method is the date augmented-based CNN model to generate 'fake data' or 'fake images'. This study was carried out using Python and its libraries. We used 242 films from the dataset gathered by the Deep Fake Detection Challenge, of which 199 were made up and the remaining 53 were real. Ten seconds were allotted for each video. There were 318 videos used in all, 199 of which were fake and 119 of which were real. Our proposed method achieved a testing accuracy of 91.47%, loss of 0.342, and AUC score of 0.92, outperforming two alternative approaches, CNN and MLP-CNN. Furthermore, our method succeeded in greater accuracy than contemporary models such as XceptionNet, Meso-4, EfficientNet-BO, MesoInception-4, VGG-16, and DST-Net. The novelty of this investigation is the development of a new Convolutional Neural Network (CNN) learning model that can accurately detect deep fake face photos.

KEYWORDS

Deep fake detection video analysis; convolutional neural network; machine learning; video dataset collection; facial landmark prediction; accuracy; models

1 Introduction

The ease with which modern technology can be acquired has led to the widespread dissemination of deep fake films on social media platforms [1]. An instance of a deep fake can be observed when an image or video substitutes the image of the subject with that of another individual. Deep fake technology has been used for disseminating false information by politicians [2–6]. As a result, deep fakes may have a severe impact on our society and spread false information, especially on social media [7,8]. Fake news can be used to distribute misleading information or completely distort legitimate news

stories [9,10]. There are many instances of false news. According to Alcott et al. [11], during the 2016 US presidential election, the movement of Clinton supporters was impacted by the dynamics of top false news spreaders. Meanwhile, the movement of Trump supporters was influenced by the typical centre- and left-leaning news broadcast by top influencers. Additionally, it was reported that fake news about the Brexit vote in the United Kingdom was used to manipulate public opinion [12,13]. Deep fake technology is primarily based on machine learning neural networks and is particularly associated with the creation of fake images, videos, and audio using Generative Adversarial Networks [14]. Deep fake, however, is routinely used to produce electronic evidence and false news, misleading the public and upsetting the social order.

Deep fakes [15] can produce false pictures and films that are challenging to spot with the naked eye, causing societal unrest [16]. Deep fake videos [17] have had a significantly larger impact than was initially anticipated. While the technology is safe for leisure use, there is a risk that it may be used for political or criminal purposes, which could have serious consequences [16,18–20]. The use of various neural network designs by academics to distinguish between fake and authentic films is not yet widespread. Deep fake video forensics research is typically still in its early stages. Deep learning has shown to be a potent and useful approach in many fields [16–23].

Problem Statement. Facial recognition is a challenging task in image processing, as it presents numerous obstacles to the development of accurate algorithms for identifying faces. The first problem that has to be solved is face detection. There are two main challenges associated with face detection [24,25]. Face emotions and a variety of facial characteristics are present. Through facial expressions, individuals convey their emotions and intentions. It is significant to remember that they may significantly alter an appearance. Many people have glasses, while others have a moustache or beard, and yet others have scars from a previous existence. Face features are those characteristics.

The motivation for this study is that the quality of deep fakes is constantly improving, which leads to the need to develop new methods for their detection. The process of detecting deep fakes involves the use of two primary types of classifiers: Deep classifiers and shallow classifiers. Shallow classifiers can accurately differentiate between counterfeit and legitimate images and films by identifying the anomalies in their characteristics. For example, other attributes such as the reflections in the eyes might potentially be overlooked. The teeth may also exhibit analogous variances that might be employed comparably. This study suggests deploying specialised Convolutional Neural Networks [26,27] that utilise a Deep Learning approach to identify counterfeit films. The main accomplishment of this research is the creation of an innovative Convolutional Neural Network (CNN) learning model that exhibits outstanding precision in identifying altered facial images generally known as deep fakes.

1.1 Literature Review

Mukta et al. [28] conducted a comprehensive literature analysis and assessed the efficacy of current deep fake detection techniques. The authors classified the models into two distinct groups: (1) traditional models that utilise machine learning and (2) models that are founded on deep learning. Deep learning models may be classified into three categories: Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based models.

Contemporary machine-learning methods are essential for comprehending the fundamental reasoning behind any activity that may be understood from a human perspective [29–31]. These methodologies provide enhanced command over data and activities, rendering them very appropriate for deepfake applications. Furthermore, it is much more straightforward to alter the structure and

hyperparameters of the model. A decision tree functions as a visual depiction of the decision-making process in several machine-learning approaches, including decision trees, random forests, and other tree-based methodologies. Therefore, the tree-based technique has no difficulties in terms of interpretability. Several studies employ support vector machines, logistic regression, KNN classifiers, as well as alternative boosting models like XGBoost and ADABoost to detect deepfakes. Deep learning methods are widely employed in computer vision because they can quickly identify and extract features from input data through feature selection and extraction [32].

CNN is a well-known and highly regarded deep-learning model. The approach is widely acknowledged and employs pre-trained convolutional neural network (CNN) models to directly extract distinguishing information from each frame of the sequences. The researchers conducting deepfake detection studies use many Convolutional Neural Network (CNN) approaches [33]. The Recurrent Neural Network (RNN) is a prevalent model in the realm of deep learning that is employed to handle sequential input. Several recurrent neural network (RNN) models have been discovered to generate and detect deepfake pictures and videos. Transformer models, including EfficientNet+ViT, M2TR, CViT, ViT, ViT+Distillation, and Video Transformer, play a vital role in the advancement and identification of deepfake technology.

In the study, a unique deep fake detection method is proposed that uses unsupervised contrastive learning [34]. The authors create two distinct copies of a picture and feed them into an encoder and a projection head. CNN facial recognition models are used in previous articles [35,36] to differentiate between real and fake photos of humans.

DeepfakeStack, a deep ensemble learning approach, was developed by Rana et al. [26] to address the problems raised by Deep Fake Multimedia. This method integrates multiple cutting-edge deep learning-based classification models to improve composite classifiers. A real-time deep fake detector can be created with this approach.

The study conducted by Suratkar et al. [37] presents a method for detecting deep fake videos using the Convolutional Neural Network (CNN) architecture and transfer learning mechanism. The proposed method in the article [38] utilises residual noise, defined as the disparity between the original image and its denoised counterpart, to execute the recommended strategy. The study investigates Xception and MobileNet as two approaches for classification tasks, commonly used for detecting deepfake movies. The user's input is " [23]." Eight deep fake video classification models were trained, tested, and evaluated utilising four methods for producing misleading videos and two advanced neural networks. Subsequently, the models were interconnected and assessed. Every model demonstrated accurate classification performance when evaluated on the specific dataset used during its creation. This study employs four distinct datasets generated by different deep fake technologies to train and evaluate the Face Forensics++ technique. The findings indicate that the precision of the method ranges from 91 to 98 per cent across all datasets, contingent upon the specific deep fake technology employed.

Current international events are predominantly accessed through social media platforms. Frequently, erroneous information emerges and disseminates on social media platforms. Moreover, it exerts a detrimental influence on the stability of society. Several studies employing diverse methodologies have established efficient frameworks for identifying fraudulent news on social media platforms. Nevertheless, there are some limitations and deficiencies. Moreover, due to its significant significance, it was discovered that the accuracy of the detection models was inadequate. While several review studies have analysed the impacts of fake news, most of them have focused on particular and recurring attributes of algorithms used to detect false information. The predominant focus of research in this field has primarily been on the categorization of the datasets, features, and classifiers employed. The

study did not investigate the constraints of the dataset, its attributes, how these attributes are included, and the influence of these factors on detection models, particularly because most detection models employed a supervised learning approach. This review study analyses current research and explores the challenges encountered by algorithms developed to identify fake news, as well as their implications on their efficacy [39].

1.2 Problem Statement

This study aims to create a novel Convolutional Neural Network (CNN) learning model that can precisely identify deep fake face images. The suggested approach entails the extraction of structured data from video frames through the utilisation of facial landmark detection. The gathered data is subsequently utilised as input for the Convolutional Neural Network (CNN). Automated feature extraction involves feeding video image frames directly into a Convolutional Neural Network (CNN). The resulting output is then linked to both an activation layer and a dense neural layer to produce the final result. The primary aims of this inquiry are as follows:

1. To apply CNN deep learning approaches to detect deep fake face photos with high accuracy.
2. To utilize the facial landmark predictor model to identify all facial characteristics, such as the eyes, mouth, and nose, to improve the accuracy of detection.
3. To develop a universal frame model that can more accurately identify deep fake faces, taking into consideration the many facial expressions that may be present.
4. To compare the performance of three different approaches for deep fake detection.

2 Methodology

This study was carried out using Python and its libraries. For testing purposes, the batch size was set to 32 and the starting learning rate at 0.0001. After 40 epochs, this procedure was terminated. Convergence in CNN training was used to choose this maximum value. The success of the study's designs is quantitatively examined using the metrics of accuracy, loss, and ROC AUC.

Below, the information about the Software and Hardware configuration testing was added.

1) Software Configuration Testing involves thoroughly analysing the application under test (AUT) concerning various operating system versions, software upgrades, and other pertinent aspects. This test is arduous since it necessitates the installation and removal of many software programs that will be utilised for testing objectives. An optimal strategy to optimise time is to utilise virtual machines for software configuration testing. A virtual machine emulates real-world setups and provides a comparable user experience to that of a physical computer.

2) Hardware Configuration Testing typically conducted in a laboratory environment, hardware configuration testing involves a set of physical machines connected to various hardware components.

2.1 Data Description

We use 242 films from the dataset gathered by the Deep Fake Detection Challenge [37], of which 199 are made up and the remaining 53 are real. Ten seconds are allotted for each video. There were 318 videos used in all, 199 of which were fake and 119 of which were real.

The main goal of the Deepfake Detection Challenge Dataset is to assess the advancements achieved in the realm of deepfake detection technologies. Facebook collaborated with prominent business leaders and esteemed academic specialists to establish the Deepfake Detection Challenge (DFDC) to expedite the advancement of cutting-edge methods for identifying deepfake videos.

Facebook generated and distributed a distinctive dataset for the challenge, comprising over 100,000 clips. The DFDC has fostered collaboration among professionals from diverse international locations, allowing them to assess and contrast their deepfake detection models, investigate novel methodologies, and exchange specialised knowledge. The DFDC dataset comprises two variations: An initial dataset featuring 5000 videos illustrating two face alteration algorithms, along with a research article; and a comprehensive dataset containing 124,000 videos demonstrating eight facial modification techniques, also accompanied by a research paper. Participants in a Kaggle competition utilised the entire dataset to develop improved algorithms for identifying altered drugs. The dataset was generated by Facebook, employing remunerated performers who consented to the use and alteration of their appearances for the development of the dataset.

Fig. 1 displays several examples of both phoney and real photographs [40].

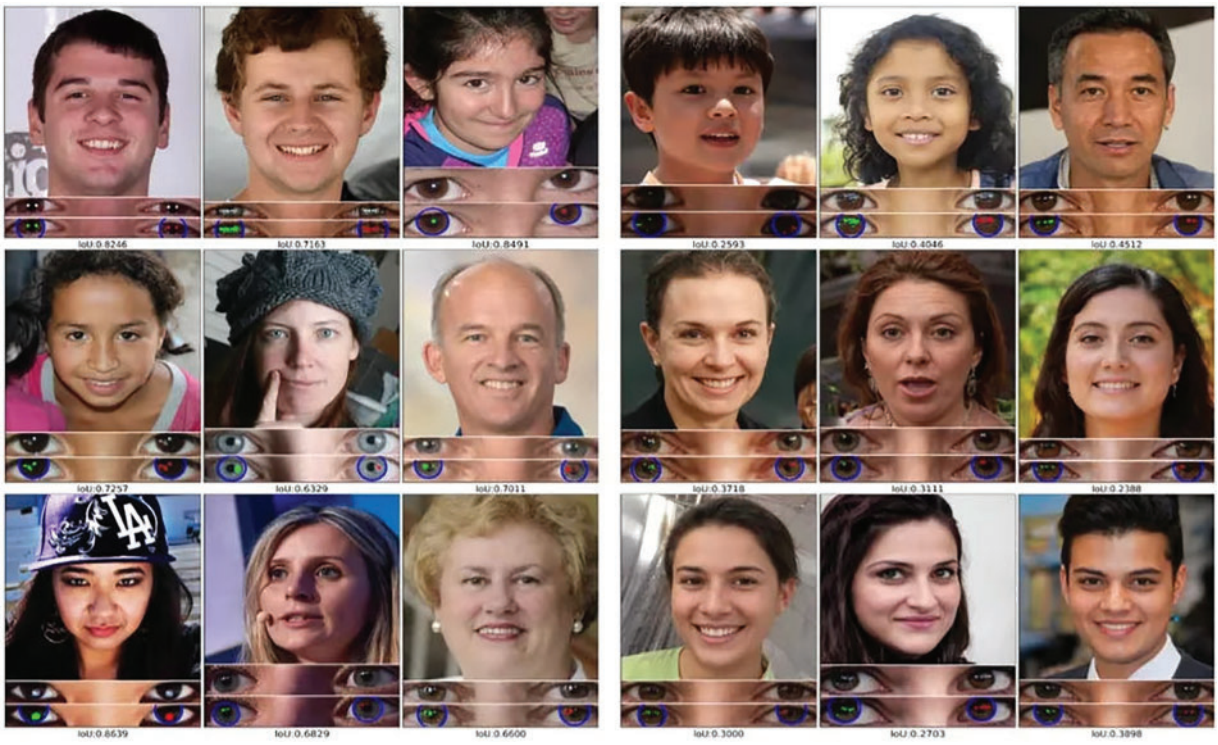

Figure 1: Sample images

Evaluating our machine learning algorithm is an essential part of this study. We used statistical values described by Zheng et al. [41].

Classification Accuracy. Most frequently, when talking about "accuracy," categorization accuracy is meant. The ratio of accurate predictions to the total number of input samples is a good indicator of accuracy.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made} \tag{1}$$

Taking into account the following description, 98 per cent of the samples in our training set originate from class A, while the remaining 2 per cent come from class B.

Logarithmic Loss (LL). The use of logarithmic loss, commonly referred to as log loss, is applied to rectify erroneous categorizations in data. It greatly enhances the process of classifying many categories [42]. The calculation of LL is calculated by the following formula, where N is the number of instances that fall into M classes:

$$LL = \frac{-1}{N} \Sigma\Sigma Y_{ij} \cdot (logP_{ij}) \tag{2}$$

where p_{ij} is the likelihood that sample i belongs to class j; LL has no upper bound and occurs in the interval [0, ∞); Y_{ij} displays whether or not sample i belongs to class j.

Whereas LL that is further from 0 indicates less accuracy, log loss that is closer to 0 shows more accuracy. In general, better categorization is obtained by lowering LL.

Confusion Matrix. As the name suggests, this produces a matrix as output that enumerates the overall effectiveness of the model (Fig. 2).

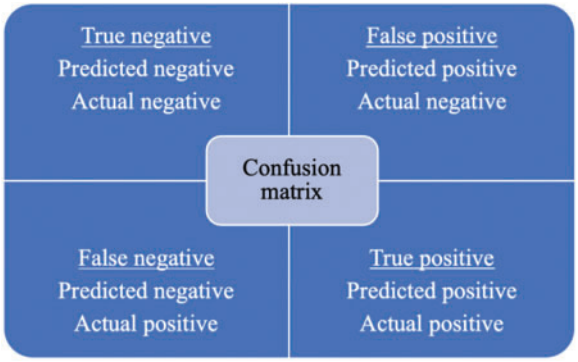

Figure 2: Confusion matrix

Averaging across the "main diagonal", which is effectively the whole matrix, can be used to gauge accuracy.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Sample} \tag{3}$$

By accurately forecasting the outcomes for each training sample in class A, our model has the potential to attain a training accuracy of 98 per cent. By employing a sample collection consisting of 60% samples from class A and 40% samples from class B, it is feasible to achieve a test accuracy of 60%. Evaluating the accuracy of categorization might create the impression that we have attained substantial levels of precision.

Area Under Curve. To calculate AUC, understanding a few key concepts is necessary. Sensitivity for True Positive Rate: Calculating TPR is done by dividing TP by (FN+TP).

$$Sensitivity = \frac{True\ Positive}{False\ Negative + True\ Positive} \tag{4}$$

True Negative Rate: TNR is determined using the following formula: TN/(FP+TN). To put it another way, the FPR is the proportion of negative data values that are correctly identified as such.

$$Specifity = \frac{True\ Negative}{True\ Negative + False\ Negative} \tag{5}$$

False Positive Rate: The formula for FPR is FP/(FP+TN). FPR is a proportion of negative data points that are incorrectly classified as positive in the context of all negative data points.

$$FPR = \frac{False\ Positive}{True\ Negative + False\ Positive} \tag{6}$$

3 Results

The steps of the suggested system are shown in Fig. 3.

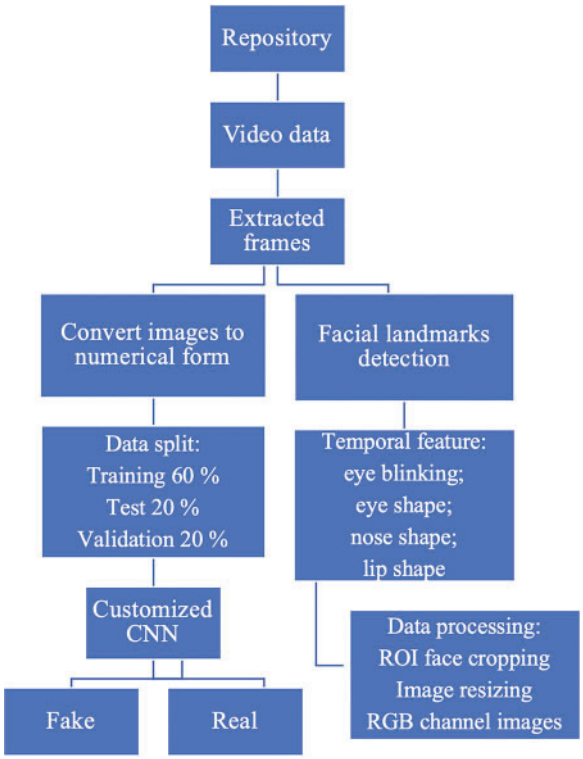

Figure 3: Flowchart of the proposed model

Individual picture frames can be recovered from video once it is initially provided as input. A facial landmarks detector may be used to locate the lips, nose, and eyes of a person. Using this information, it is possible to infer eye blinks and other facial characteristics. It is required to perform some sort of preprocessing before feeding the model with this input. Pre-processing, however, changes the images into their numerical form. In the first, it resizes the input image frames to 224 by 224 and crops the facial region of interest. It is important to ensure that every image is in the RGB channel. By applying this specialized deep learning method based on the CNN model, the classification step can determine if a particular video is a deep fake or not.

Frame Extraction and Facial Landmarks Detection involve separating a movie into discrete pictures and producing image frames. By separating a movie into discrete pictures, image frames are produced. This movie consists of 3735 frames in total. Each picture shows the same face. The face area is used to get the x, and y locations of 68 facial landmarks. A trained face landmark detector is present in the

specified digital library. To detect a face in a picture, 68 facial landmarks are first identified using the digital library.

Study of Temporal Facial Features. Preventing deep fakes from being compromised by unnatural eye blinking is a challenge. This phase involves capturing the subject's blinking pattern. The blink detector receives eye coordinates as input, which are obtained from the landmarks on the face. The aspect ratio of the eye is used to determine if someone is blinking (EAR). Six alternative landmark placements may be used to symbolize each eye independently. The eye is represented by six (x, y) coordinates, clockwise from the left corner (p1) of the image (p2, p3, p4, p5, p6):

$$EAR = \frac{\|pp2 - pp6\| + \|pp3 - pp5\|}{2\|pp1 + pp4\|} \tag{7}$$

The sign ‖p2-p6‖ indicates how far apart the points p2 and p6 are from one another. When the eye is open, the ear maintains a consistent level, but when the eye is closed, it decreases to zero. The facial landmarks detector may be utilised to extract certain facial characteristics, such as the precise coordinates of the eyes, nose, and mouth. Deepfake movies commonly utilise the face manipulation approach to generate a diverse range of eye shapes. These shapes are determined by examining the extracted characteristics from the generated eyes. Our research indicates that the physical structure of a person's eyes remains surprisingly consistent in a genuine movie. Regardless of any manipulation of the tape to create a deceptive impression, this was never the true situation. Furthermore, the oral area is the primary site for the majority of facial abnormalities, such as facial malformations. As a result of significant and inaccurate alterations, there are several discrepancies in lip forms. To train our classifier, we want to utilise the disparities in face attributes between frames. The eye coordinates obtained from facial markers are utilised as input to ascertain the eye shape using an eye shape detector. The eye shape detector utilises an eye form detector to calculate the Euclidean distance (d1) between the endpoints of the left eye and the Euclidean distance (d2) between the endpoints of the right eye. The Lip form detector utilises the lip coordinates (points 49–68) acquired from face landmarks. The length of the inner lips (d3) is dictated by the coordinates of d1, which correspond to a specific location on the inner lips. Similarly, the Euclidean distance between the coordinates of the outer lip may be employed to ascertain the length of the outer lips (d4). The Nose Shape detector utilises the facial landmark data it gathers to accurately ascertain the precise configuration of the nose. The base width of the nose is dictated by the distance (d5) between its two edges. The distance is utilised to calculate the elevation of the nose at its maximum point (d6). Therefore, the measurements of facial features, such as the distances between the eyes (d1, d2), the positions of the inner and outer lips (d3, d4), and the upper and lower parts of the nose (d3, d4), were acquired (d5, d6).

Data Preparation. Preprocessing allows for the removal of unwanted artefacts and the enhancement of essential features crucial to the application under development. Some of these elements could change based on the application. To account for variations in photo size when acquiring images from cameras for use in our AI algorithms, we define a baseline size for all images.

The Region of Interest (ROI) is automatically recognized in pictures using computer vision in Face Crop. Next, a rectangular crop is used, concentrating on either the biggest face or all of its faces. Even if the image is scaled up or down, DNNs can detect faces at a resolution of 300 × 300 pixels. By "cropping" a photo, we imply selecting and erasing the ROI (region of interest) from the image. The face of a picture might need to be cut off for a face-detection program. When cropping a picture, the attempt is to remove any elements that are unrelated to the subject at hand. This step is also known as selecting an area of interest, or ROI.

Image Resizes. Resizing works best when used to lower the size of the photo to fit a specific dimension or to minimize the file size. In this instance, we reduced the photographs' resolution to 224×224 pixels.

Training, Validation, and Testing Data Split. This dataset consisted of several subsets that had been intentionally created for training, validation, and testing. The database was divided into three segments: 60% of the data was assigned for training, 20% for validation, and another 20% for testing. The collection comprised 2399, 750, and 600 images. The proportion of genuine and fraudulent films remained constant across all categories. The validation approach was utilised to ascertain the most optimal design for the Convolutional Neural Network (CNN) in this methodology. The validation set was employed to choose the optimal architecture for training the model, while the training and test sets were merged to evaluate the model's performance post-training.

Convolutional Neural Network (Customized). The motivation for creating of Customized Convolutional Neural Network method is to improve the quality of deep fake determination. A convolutional neural network [43,44] is a DNN [45] that is widely used to recognize patterns in images. Also, for uniformity and faster processing, each image is reduced in size to fit inside the 220 px by 3 px boundaries. Some details on the visual frames are sent to CNN. In order to enhance the network's ability to recognise unique features, input parameters of 32, 64, and 128 filters of progressively bigger sizes are utilised. The model is constructed using a modified Convolutional Neural Network (CNN) architecture. The model consists of 40 epochs and includes layers such as Conv2D, ReLU, Batch Normalisation, MaxPooling2D, and a densely connected layer. Ultimately, we apply a substantial coating on top of the first layer, followed by the suitable layers for Batch Normalisation and Dropout. To assess the influence of adjacent pixels, we employ a filter. As expected, we generate a filter with the size supplied by the user (a suggested standard is 33 or 55) and then position it diagonally from the upper left corner to the lower right corner of the image. Every pixel in the image has a value thanks to a convolutional filter. Each filter generates a feature map once it has passed over the image. In other words, the proposed Customized Convolutional Neural Network method involves extracting structured data from video frames using facial landmark detection, which is then used as input to the CNN. Customized Convolutional Neural Network method is the date augmented-based CNN model to generate 'fake data' or 'fake images'.

The image is subjected to several filters using a convolutional layer to extract various kinds of information. It is simpler to compute and gives sparsity with ReLU [46]. The three types of layers to consider while creating a CNN are convolutional, pooling, and fully connected layers/dense layers. Each of these levels uses the provided data in a certain way and contains several variables that may be changed.

Convolution Layers. Convolution Layers are the filtering layers in which extra feature maps are used or the original image is filtered. The great bulk of the user-defined parameters for the network are located here. The most popular type of convolution is 2D, and it is frequently referred to as conv2D. A filter or kernel multiplies two-dimensional input data elementwise in a conv2D layer. Consequently, every piece of information will be accommodated within a solitary display pixel. Each time the kernel traverses a site, it performs a consistent operation, transforming a 2D feature matrix into another 2D feature matrix.

Pooling Layers. Max-pooling involves selecting the highest value within the filter zone, whereas average pooling involves choosing the mean value within the filter region. They are commonly employed to minimise the network's size.

Dropout Layers. Dropout is commonly used to fully link layers since they have a large number of parameters, which raises the possibility of excessive co-adaptation and overfitting. Both pooling layers and convolutional layers, such as Conv2D, can be utilised either before or after dropout. Dropout is commonly implemented after the pooling layers, based on a fundamental heuristic. However, there may be some cases where this rule does not apply. Dropout may be implemented on any specific cell or element inside a feature map.

Fully Connected Layers/*Dense Layers*. Prior to CNN's classification output, fully connected layers (FCLs) are integrated. Moreover, they are employed to standardise findings before categorization. It has a resemblance to the output layer of a Multilayer Perceptron (MLP).

The suggested algorithm's flowchart is displayed in Fig. 4 below.

The algorithm depicted in Fig. 4 comprises the following steps: Obtaining a video dataset as input, extracting frames from the video, identifying facial features using a landmark prediction model, preprocessing the frames, partitioning the data into three segments, configuring parameters, employing a customised CNN model by incorporating additional layers from training, conducting testing on an independent test set, computing performance metrics, and determining the classification outcomes as either counterfeit or genuine.

A comparative examination of Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN) models was conducted. The Multilayer Perceptron (MLP) is a neural network model commonly employed in computer vision tasks. However, Convolutional Neural Networks (CNN) have surpassed it in terms of performance in this field. The Multilayer Perceptron (MLP) is deemed unsuitable for modern sophisticated computer vision applications because of its dependence on completely linked layers, wherein each perceptron forms links with every other perceptron. An important constraint is that the overall number of parameters might significantly increase as a result of the multiplication of the number of perceptrons in layer 1 with the number of parameters in layer 2, and further amplified by the number of parameters in layer 3, and so forth. The presence of significant redundancy in dimensions makes this technique inefficient. Furthermore, it fails to incorporate geographical data. The inputs need to be converted into flattened vectors. An MLP with a constrained number of layers (2–3) has the potential to attain a significant level of precision when trained on the MNIST dataset. A Convolutional Neural Network (CNN) is the dominant and most preferred approach for solving computer vision issues. It has consistently outperformed rival methods in several ImageNet contests. The technique applies each filter on the whole image in a systematic manner, using preset dimensions and strides. This allows the filter to precisely detect and align patterns at any position within the image. The weights are often small in magnitude and distributed, leading to less resource usage and a streamlined training process compared to MLP, hence improving efficiency. In addition, CNN designs include the potential to handle networks with a larger number of layers. The layers exhibit a fragmented kind of connection rather than a full form of connectivity. Both matrices and vectors can be used as input. The layers exhibit a restricted or partial degree of interconnectedness, in contrast to total connection. There is no correlation between all nodes.

This research can discern whether a video is a deep fake or not. A deep fake may employ a voiceover or a face swap (or both). The labels "FAKE" or "REAL" in the label column in training data serve as indicators. The likelihood that the video is fake has been forecasted here (Fig. 5).

The y-axis displays total numbers, while the x-axis displays video class. In this story, there are two types of videos: Real and fake, denoted by the numbers 0 and 1. It can be seen from this graph that there are about equal numbers of counts for both classifications. In particular, the number of REAL values is 1854, and the number of FALSE values is 1881.

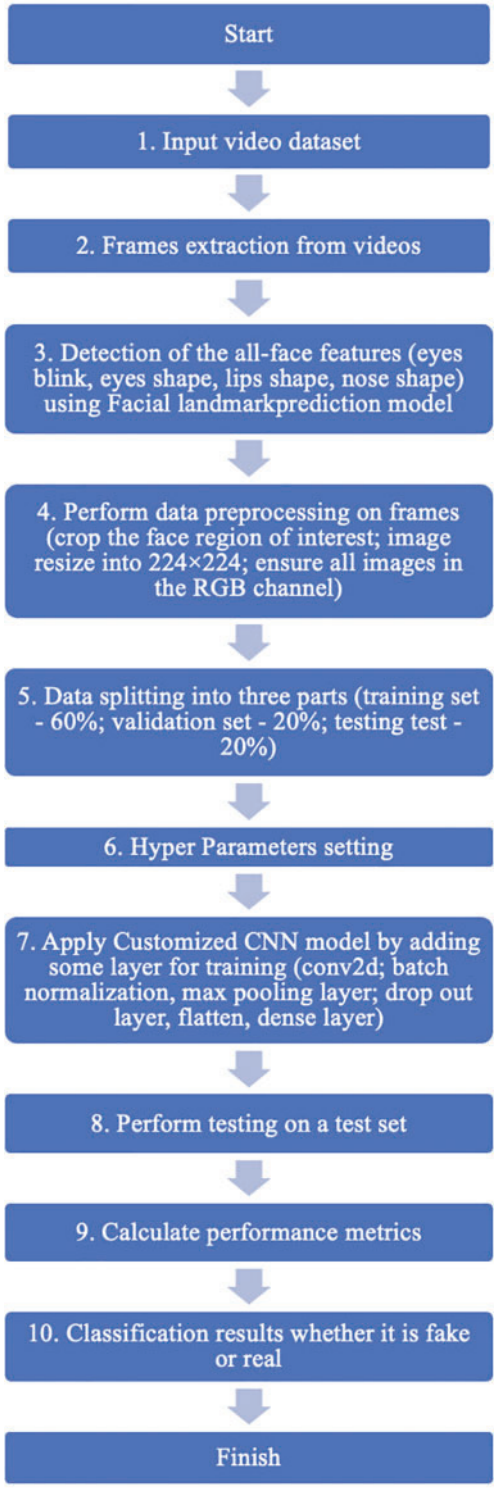

Figure 4: Flowchart of the proposed algorithm

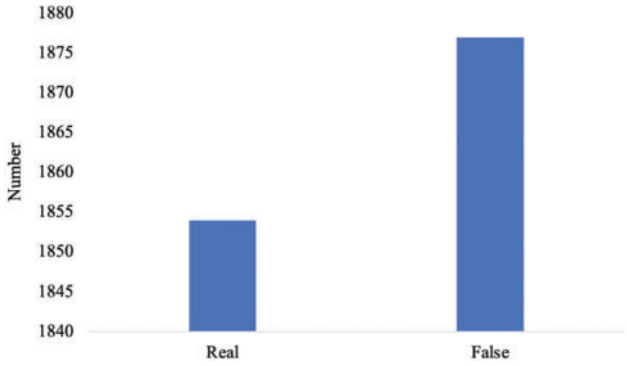

Figure 5: Distribution of the dataset for the compilation of deep fake videos

Fig. 6 shows the confusion matrix for test data.

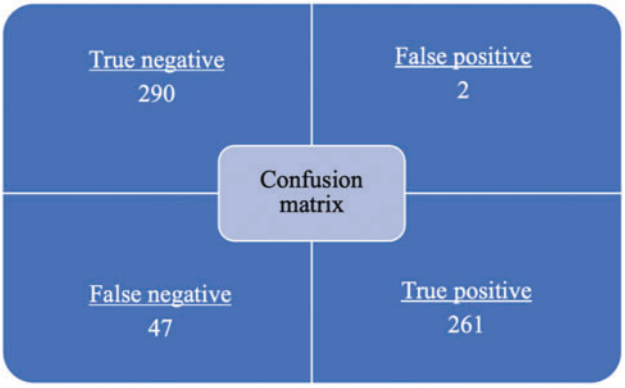

Figure 6: Confusion matrix, without normalization

If we confront a binary classification issue involving numerous instances categorized into two groups, fake and real, and we perform 600 model evaluations, the ensuing results are as follows. The four keywords are TP – 261; TN – 290; FP – 2; and FN – 47.

The comparison line graph in Fig. 7 shows how the three approaches' training and validation loss values differ from one another.

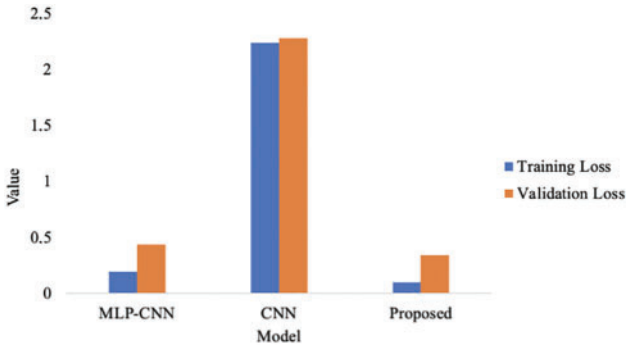

Figure 7: Training loss and validation loss comparison

Analyzing the data shown in Fig. 7, we can conclude that different models are characterized by different Loss values. The Loss values for the MLP-CNN model are training –0.1948 and validation −0.4383; CNN training −2.2433 and validation −2.281; Proposed training −0.1003 and validation −0.342.

The comparative bar graph for accuracy and AUC Score between the three approaches is shown in Fig. 8.

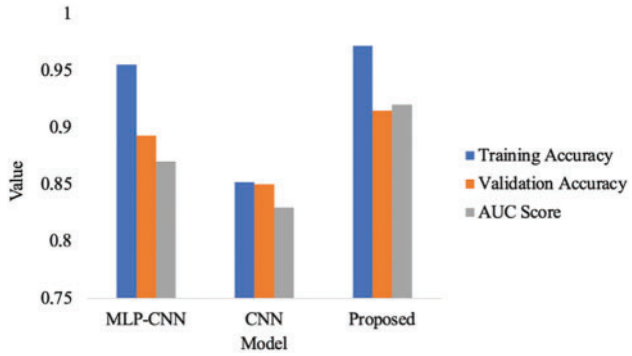

Figure 8: Training accuracy, validation accuracy and AUC Score comparison

The comparison graph illustrates that the training and validation accuracy of the Convolutional Neural Network (CNN) exhibit a high degree of similarity. The recommended Customized CNN beats these two methods in terms of training and validation data accuracy, but MLP-CNN generated favourable classification results. The contrasted line graph demonstrates that the MLP-AUC CNN's score of 0.87 is higher than CNN alone's AUC score of 0.83, another popular method. As we can see, MLP-CNN outperformed not only in terms of AUC score value but also outperformed the suggested Customized CNN, which had an AUC score of 0.92.

4 Discussion

Our proposed method outperformed two alternative approaches, CNN and MLP-CNN. Various techniques for identifying video deep fakes have been explored, with some focusing on manually crafted features [47] and others on physiological [48], among other approaches. For example, authors [47] introduced a non-pixel-based approach where feature vectors were generated using data extracted from the films' stream descriptors. Following that, the ensemble of SVM and random forest classifiers was trained using those feature vectors. On the Media Forensics Challenge (MFC) dataset, an AUC score of 98.4% was attained. Despite its strong performance, this strategy [47] cannot defend against video re-encoding assaults. The authors described a technique for locating the deepfake films using biological traits such as heart rate estimates [49]. Similar physiological parameters based on rPPG were employed by Qi et al. [50] and Fernandes et al. [51] to distinguish between authentic and false films.

Scientists suggest utilising deep neural networks to develop algorithms with the ability to detect deepfake videos, aiming to combat the possible misuse of these manipulated films. The researchers [52] proposed a framework that utilises the XceptionNet architecture and Bidirectional LSTM. The temporal sequence was assessed using Bidirectional LSTM, while the facial characteristics were extracted using XceptionNet. The model underwent training using a combination of the KL divergence and Cross Entropy loss functions to distinguish between genuine and fraudulent video characteristics. Similar to de Lima et al. [53], the temporal sequence descriptors were obtained from

the LSTM utilizing face characteristics that were collected from video frames using VGG-11. While it was computationally difficult, the technique [53] produced respectable detection accuracy on the Celeb-DF dataset.

In their study, Wang et al. [4] presented a distinctive model for complicated picture retrieval that is influenced by visual saliency. Firstly, the Itti visual saliency model is introduced. The composite saliency map in this model is formed by merging the direction, intensity, and colour saliency maps. The authors then proposed the concept of a multi-feature fusion paradigm to enhance the accuracy of visual pattern descriptions. To address the intricate nature of the image, the authors proposed a dual approach: (1) Assessing complexity by taking into account cognitive load; and (2) Classifying levels of cognitive complexity. Integrating the group sparse logistic regression model is crucial for the successful implementation of the photo retrieval system.

The researchers Yu et al. [54] presented a novel encryption method that utilises quaternion Fresnel transforms (QFST), computer-generated holograms, and the two-dimensional (2D) Logistic-adjusted-Sine map (LASM). Two instances of the quaternion Fresnel transform (QFST) are devised to efficiently handle the four images, and a corresponding computing technique for a quaternion matrix is created. The initial processing of the first four pictures, which are encoded using quaternion algebra, is collectively conducted in a vectorized manner using QFST. The initial complex amplitude is encoded via the Fresnel transform through the use of two separate virtual and independent random phase masks (RPM).

An image retrieval system employing semantic analysis is presented to enhance precision. This method combines a pre-existing C-Tree with a neighbouring graph called Graph-CTree [55]. The k-NN algorithm is employed to categorise a cluster of analogous photographs gathered using Graph-CTree to generate a collection of visual descriptors. An image ontology framework is constructed using a somewhat automated approach. The procedure entails the automated development of SPARQL queries by using visual keywords and extracting them from ontologies. These queries are subsequently employed to articulate the semantic data encapsulated inside pictures. The investigation utilised many photo datasets including COREL, WANG, ImageCLEF, and Stanford Dogs. The accuracy values derived from each dataset were 0.888473, 0.766473, 0.839814, and 0.826416, respectively.

We conducted a comparative study between our model and the most advanced existing methods to evaluate and measure the effectiveness of the suggested strategy. We performed a meticulous comparison between our technique and the methodologies outlined by Khalid et al. [56] and Ilyas et al. [57]. The level of accuracy of both the recommended and current models is presented in Table 1.

Table 1: Comparison with existing models

Model	Accuracy, %
XceptionNet	73.06
Meso-4	43.15
EfficientNet-BO	59.64
MesoLnception-4	77.88
VGG-16	81.03
DST-Net	90.94
Proposed	91.47

It can be seen that such models as XceptionNet, Meso-4, EfficientNet-BO, MesoLnception-4, VGG-16, and DST-Net have an accuracy of 43.15–90.94%. By achieving the greatest accuracy of 91.47%, the suggested strategy beats the current state-of-the-art models. The DST-Net outperformed the Meso-4 model in the video-only modality, which had the lowest accuracy overall.

In addition, our findings may be systematically compared to the data obtained by other researchers, as carefully evaluated in the study conducted by Hamed et al. [39]. Fig. 9 illustrates the relationship between the amount of research performed on the development of methods for identifying deepfakes and the accuracy of these methods.

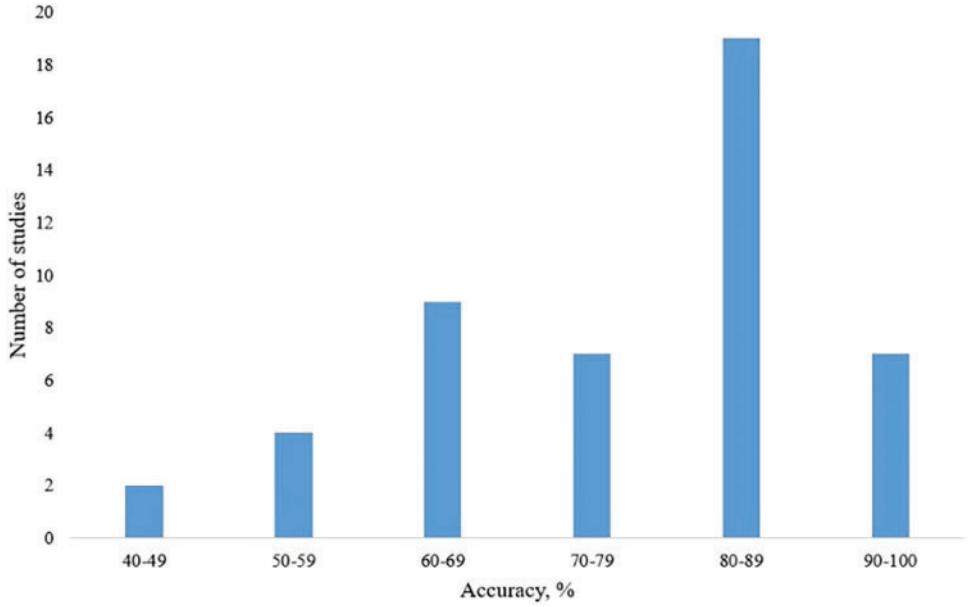

Figure 9: The ratio of the number

Based on our analysis of the data shown in Fig. 9 and our study findings, we assert with confidence that our results are among the most exceptional studies in the realm of devising techniques for detecting deepfakes. Based on the study and its findings, it can be inferred that more enhancements are needed to increase the accuracy of the results. The effectiveness and accuracy of our algorithms in detecting false news are influenced by many obstacles and limitations. The discussed topics encompass the datasets employed, concerns regarding overfitting and underfitting, incorporation of image-based features, representation of feature vectors, utilisation of machine learning models, and data fusion.

Consistent with the findings of Mukta et al. [28], we affirm that the progress in deepfake technology has underscored the need for stronger detection methods to tackle the weaknesses in current face-forensic technologies. The integration of forgery localization and deepfake detection tasks in multitask learning has demonstrated an improvement in detection accuracy. Another noteworthy field of study is on enhancing deepfake-generation algorithms to create more authentic videos with a restricted amount of source material. It is essential to address and minimize the risks associated with deepfakes in order to effectively identify them.

The effectiveness and accuracy of our algorithms in detecting false news are influenced by many obstacles and limitations. The variables include the choice of datasets, issues with overfitting and underfitting, the use of image-based features, the encoding of feature vectors, reliance on machine

learning models, and the integration of data fusion techniques. Our suggested strategy may effectively be employed on social media platforms to counteract the dissemination of counterfeit films in real-life situations.

5 Conclusion

In this study, we proposed a novel method for detecting deepfake videos produced by AI.

Our main contribution to science and development is to perform a literature review and pose problems. In addition, an original research structure was developed. Individual frames of an image can be reconstructed from a video if it is initially provided as input. A facial landmark detector can be used to determine the location of a person's lips, nose, and eyes. This data can be utilized to infer various facial expressions and characteristics, including eye blinks. Before providing this input to the model, preliminary preprocessing is required. In the first case, it resizes the input image frames to 224 by 224 and crops the facial region of interest. It is important to ensure that each image is in the RGB channel. By applying this specialized deep learning method based on the CNN model, the classification step can determine whether a particular video is fake or not. Finally, an original algorithm for identifying deep fakes of images in video collections was developed. The proposed algorithm consists of the following steps: Receiving a collection of video data, extracting individual frames from the video, identifying facial features using a landmark prediction model, performing initial processing on the frames, dividing the data into three sections, adjusting hyperparameters, employing a customised CNN model with extra training layers, conducting testing on a separate test set, computing performance metrics, and classifying the outcomes as either fake or real. The proposed approach exhibited a testing accuracy of 91.47%, a loss of 0.342, and an AUC score of 0.92. It surpassed two other methods, CNN and MLP-CNN, in terms of performance. Furthermore, our method achieved greater accuracy than contemporary models such as XceptionNet, Meso-4, EfficientNet-BO, MesoInception-4, VGG-16, and DST-Net.

The developed method can be used to determine 'deep fake' or 'deep images' in the different types of images, including social media services. Our results should influence the research of other scientists in the field of detecting 'deep fakes' or 'deep images'. However, future research should focus on increasing the diversity of persons who can be consistently recognized by the algorithm, such as people of colour. Additionally, combining more spatial and temporal facial data could further improve the algorithm's accuracy. To enhance the resilience of our suggested approach against malicious attacks, further investigation is needed to explore the use of Generative Adversarial Networks (GANs) for generating more authentic counterfeit films that may be employed for testing purposes.

The efficacy and accuracy of our models in detecting fake news are influenced by various challenges and limitations, including the datasets used, concerns about overfitting and underfitting, image-based attributes, the encoding of feature vectors, machine learning models, and data fusion. Our suggested approach is effective in combating the dissemination of counterfeit films, especially on social media platforms.

Acknowledgement: Not applicable.

Funding Statement: Bo Dong was supported by Science and Technology Funds from the Liaoning Education Department (Serial Number: LJKZ0104).

Author Contributions: The authors confirm their contribution to the paper as follows: Study conception and design: D. Gura; data collection: B. Dong; analysis and interpretation of results: D. Mehiar; draft manuscript preparation: N. Al Said. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data available on request from the authors. The data that support the findings of this study are available from the corresponding author, Bo Dong, upon reasonable request.

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

References

[1] A. M. Almars, "Deepfakes detection techniques using deep learning: A survey," *J. Comput. Commun.*, vol. 9, no. 5, pp. 20–35, 2021. doi: 10.4236/jcc.2021.95003.

[2] A. O. Beketaeva, A. Z. Naimanova, N. Shakhan, and A. Zadauly, "Simulation of the shock wave boundary layer interaction in flat channel with jet injection," *Z. Angew. Math. Mech.*, vol. 103, no. 8, pp. e202200375, 2023. doi: 10.1002/zamm.202200375.

[3] L. Nataraj *et al.*, "Detecting GAN generated fake images using co-occurrence matrices," 2019. doi: 10.48550/arXiv.1903.06836.

[4] H. Wang, Z. Li, Y. Li, B. B. Gupta, and C. Choi, "Visual saliency guided complex image retrieval," *Pattern Recognit. Lett.*, vol. 130, pp. 64–72, 2020. doi: 10.1016/j.patrec.2018.08.010.

[5] S. Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot . . . for now," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, Piscataway, NJ, USA, IEEE, 2020, pp. 8695–8704. Accessed: Feb. 08, 2024. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_CNN-Generated_Images_Are_Surprisingly_Easy_to_Spot..._for_Now_CVPR_2020_paper.html

[6] C. C. Hsu, C. Y. Lee, and Y. X. Zhuang, "Learning to detect fake face images in the wild," in *2018 Int. Symp. Comput., Consum. and Control (IS3C)*, Piscataway, NJ, USA, IEEE, 2018, pp. 388–391. doi: 10.1109/IS3C.2018.00104.

[7] D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE Int. Conf. Adv. Video and Signal Based Surveillance (AVSS)*, Piscataway, NJ, USA, IEEE, 2019, pp. 1–6. doi: 10.1109/AVSS.2018.8639163.

[8] F. Sun, N. Zhang, P. Xu, and Z. Song, "Deepfake detection method based on cross-domain fusion," *Secur. Commun. Netw.*, vol. 2021, pp. 2482942, 2021. doi: 10.1155/2021/2482942.

[9] A. Bovet and H. A. Makse, "Influence of fake news in Twitter during the 2016 US presidential election," *Nat. Commun.*, vol. 10, pp. 7, 2019. doi: 10.1038/s41467-018-07761-2.

[10] B. Kropf, M. Wood, and K. Parsons, "Message matters: Correcting organizational fake news," *Comput. Hum. Behav.*, vol. 144, pp. 107732, 2023. doi: 10.1016/j.chb.2023.107732.

[11] H. Alcott and M. Gentzkow, "Social media and fake news in the 2016 election," *J. Econ. Perspect.*, vol. 32, no. 2, pp. 211–236, 2017. doi: 10.3386/w23089.

[12] M. V. Nutskova, E. Y. Rudiaeva, V. N. Kuchin, and A. A. Yakovlev, "Investigating of compositions for lost circulation control," in V. Litvinenko (Ed.), *Youth Technical Sessions Proceedings*, London: CRC Press, 2019, pp. 394–398. 10.1201/9780429327070

[13] H. T. Phan, N. T. Nguyen, and D. Hwang, "Fake news detection: A survey of graph neural network methods," *Appl. Soft Comput.*, vol. 139, pp. 110235, 2023. doi: 10.1016/j.asoc.2023.110235.

[14] A. Aggarwal, M. Mittal, and G. Battineni, "Generative adversarial network: An overview of theory and applications," *Int. J. Inf. Manage. Data Insight*, vol. 1, no. 1, pp. 100004, 2021. doi: 10.1016/j.jjimei.2020.100004.

[15] J. C. Dheeraj, K. Nandakumar, A. V. Aditya, B. S. Chethan, and G. C. R. Kartheek, "Detecting deepfakes using deep learning," in *2021 Int. Conf. Recent Trends on Electron., Inf., Commun. & Technol. (RTEICT)*, Piscataway, NJ, USA, IEEE, 2021, pp. 651–654. doi: 10.1109/RTEICT52294.2021.9573740.

[16] M. Li, B. Liu, Y. Hu, and Y. Wang, "Exposing deepfake videos by tracking eye movements," in *2020 25th Int. Conf. Pattern Recognit. (ICPR)*, Piscataway, NJ, USA, IEEE, 2020, pp. 5184–5189. doi: 10.1109/ICPR48806.2021.9413139.

[17] Y. Al-Dhabi and S. Zhang, "Deepfake video detection by combining convolutional neural network (CNN) and recurrent neural network (RNN)," in *2021 IEEE Int. Conf. Comput. Sci., Artificial Intell. and Electron. Eng. (CSAIEE)*, Piscataway, NJ, USA, IEEE, 2021, pp. 236–241. doi: 10.1109/CSAIEE54046.2021.9543264.

[18] A. Badale, L. Castelino, and J. Gomes, "Deep fake detection using neural networks," *NTASU—2020*, vol. 9, no. 3, pp. 349–354, 2021. doi: 10.17577/IJERTCONV9IS03075.

[19] Y. Li, M. C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking," in *2018 IEEE Int. Workshop on Inf. Forensics and Security (WIFS)*, Piscataway, NJ, USA, IEEE, 2019, pp. 1–7. doi: 10.1109/WIFS.2018.8630787.

[20] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano and H. Li, "Protecting world leaders against deep fakes," in *CVF Conf. Comput. Vis. and Pattern Recognit. (CVPR) Workshops*, Piscataway, NJ, USA, IEEE, 2019, pp. 38–45. Accessed: Feb. 8, 2024. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.pdf?source=post_page

[21] G. Lee and M. Kim, "Deepfake detection using the rate of change between frames based on computer vision," *Sens.*, vol. 21, no. 21, pp. 7367, 2021. doi: 10.3390/s21217367.

[22] M. Nagao, "Natural language processing and knowledge," in *2005 Int. Conf. Natural Lang. Process. and Knowl. Eng.*, Piscataway, NJ, USA, IEEE, 2005, pp. 1–10. doi: 10.1109/NLPKE.2005.1598694.

[23] D. Pan, L. Sun, R. Wang, X. Zhang, and R. O. Sinnott, "Deepfake detection through deep learning," in *2020 IEEE/ACM Int. Conf. Big Data Comput., Appl. and Technol. (BDCAT)*, Piscataway, NJ, USA, IEEE, 2020, pp. 134–143. doi: 10.1109/BDCAT50828.2020.00001.

[24] S. Kolagati, T. Priyadharshini, and V. Mary Anita Rajam, "Exposing deepfakes using a deep multilayer perceptron—Convolutional neural network model," *Int. J. Inf. Manag. Data Insights*, vol. 2, no. 1, pp. 100054, 2022. doi: 10.1016/j.jjimei.2021.100054.

[25] D. D. Billur, T. M. Manu, and V. Patil, "A comparative analysis of video summarization techniques," *Int. J. Manuf. Eng.*, vol. 13, no. 3, pp. 10–24, 2023. doi: 10.5815/ijem.2023.03.02.

[26] M. S. Rana and A. H. Sung, "DeepfakeStack: A deep ensemble-based learning technique for deepfake detection," in *2020 7th IEEE Int. Conf. Cyber Security and Cloud Comput. (CSCloud)/2020 6th IEEE Int. Conf. Edge Comput. and Scalable Cloud (EdgeCom)*, Piscataway, NJ, USA, IEEE, 2020, pp. 70–75. doi: 10.1109/CSCloud-EdgeCom49738.2020.00021.

[27] S. Albawi, T. A. Mohammed, and S. Al-Azawi, "Understanding of a convolutional neural network," in *2017 Int. Conf. Eng. and Technol. (ICET)*, Piscataway, NJ, USA, IEEE, 2018, pp. 1–6. doi: 10.1109/ICEngTechnol.2017.8308186.

[28] M. S. H. Mukta *et al.*, "An investigation of the effectiveness of deepfake models and tools," *J. Sens. Actuator Netw.*, vol. 12, no. 4, pp. 61, 2023. doi: 10.3390/jsan12040061.

[29] G. Hu, Y. Guo, and L. Abualigah, "Genghis Khan shark optimizer: A novel nature-inspired algorithm for engineering optimization," *Adv. Eng. Inform.*, vol. 58, pp. 102210, 2023. doi: 10.1016/j.aei.2023.102210.

[30] M. Ghasemi, M. Zare, A. Zahedi, M. Akbari, S. Mirjalili, and L. Abualigah, "Geyser inspired algorithm: A new geological-inspired meta-heuristic for real-parameter and constrained engineering optimization," *J. Bionic. Eng.*, vol. 21, pp. 374–408, 2023. doi: 10.1007/s42235-023-00437-8.

[31] A. E. Ezugwu, J. O. Agushaka, L. Abualigah, S. Mirjalili, and A. H. Gandomi, "Prairie dog optimization algorithm," *Neural. Comput. Appl.*, vol. 34, no. 22, pp. 20017–20065, 2022. doi: 10.1007/s00521-022-07530-9.

[32] J. Indhumathi, M. Balasubramanian, and B. Balasaigayathri, "Real-time video based human suspicious activity recognition with transfer learning for deep learning," *Int. J. Image Graph. Signal Process.*, vol. 13, no. 1, pp. 47–62, 2023. doi: 10.5815/ijigsp.2023.01.05.

[33] S. Salunkhe, S. Bhosal, and S. V. Narkhede, "An efficient video steganography for pixel location optimization using Fr-WEWO algorithm based deep CNN model," *Int. J. Image Graph. Signal Process.*, vol. 15, no. 3, pp. 14–30, 2023. doi: 10.5815/ijigsp.2023.03.02.

[34] S. Fung, X. Lu, C. Zhang, and C. T. Li, "DeepfakeUCL: Deepfake detection via unsupervised contrastive learning," in *2021 Int. Joint Conf. on Neural Netw. (IJCNN)*, Piscataway, NJ, USA, IEEE, 2021, pp. 1–8. doi: 10.1109/IJCNN52387.2021.9534089.

[35] R. Rafique, M. Nawaz, H. Kibriya, and M. Masood, "DeepFake detection using error level analysis and deep learning," in *2021 4th Int. Conf. Comput. Inf. Sci. (ICCIS)*, Piscataway, NJ, USA, IEEE, 2021, pp. 1–4. doi: 10.1109/ICCIS54243.2021.9676375.

[36] G. Jaiswal, "Hybrid recurrent deep learning model for DeepFake video detection," in *2021 IEEE 8th Uttar Pradesh Sec. Int. Conf. Electrical, Electron. and Comput. Eng. (UPCON)*, Piscataway, NJ, USA, IEEE, 2021, pp. 1–5. doi: 10.1109/UPCON52273.2021.9667632.

[37] S. Suratkar, E. Johnson, K. Variyambat, M. Panchal, and F. Kazi, "Employing transfer-learning based CNN architectures to enhance the generalizability of deepfake detection," in *2020 11th Int. Conf. Comput., Commun. and Netw. Technol. (ICCCNT)*, Piscataway, NJ, USA, IEEE, 2020, pp. 1–9. doi: 10.1109/ICCCNT49239.2020.9225400.

[38] M. C. El Rai, H. Al Ahmad, O. Gouda, D. Jamal, M. A. Talib and Q. Nasir, "Fighting Deepfake by residual noise using convolutional neural networks," in *2020 3rd Int. Conf. Signal Process. and Inf. Security (ICSPIS)*, Piscataway, NJ, USA, IEEE, 2020, pp. 1–4. doi: 10.1109/ICSPIS51252.2020.9340138.

[39] S. K. Hamed, M. J. Ab Aziz, and M. R. Yaakub, "A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion," *Heliyon*, vol. 9, no. 10, pp. e20382, 2023. doi: 10.1016/j.heliyon.2023.e20382.

[40] Meta, "Deepfake detection challenge datase," *Meta*, 2020. Accessed: Feb. 8, 2024. [Online]. Available: https://ai.facebook.com/datasets/dfdc/

[41] A. Zheng, "Evaluating machine learning models," *O'Reilly*, 2015. Accessed: Feb. 8, 2024. [Online]. Available: https://www.oreilly.com/library/view/evaluating-machine-learning/9781492048756/

[42] D. Godoy, "Understanding binary cross-entropy/log loss: A visual explanation," *Towardsdatascience*, 2018. Accessed: Feb. 8, 2024. [Online]. Available: https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a

[43] G. B. de Souza, D. F. da Silva Santos, R. Goncalves Pires, J. P. Papa, and A. N. Marana, "Efficient width-extended convolutional neural network for robust face spoofing detection," in *2018 7th Braz. Conf. Intell. Syst. (BRACIS)*, Piscataway, NJ, USA, IEEE, 2018, pp. 230–235. doi: 10.1109/BRACIS.2018.00047.

[44] Infobae, "With these 5 applications you can create "deep fakes of photos and videos," *Infobae*, 2022. Accessed: Feb. 8, 2024. [Online]. Available: https://www.infobae.com/en/2022/03/30/with-these-5-applications-you-can-create-deep-fakes-of-photos-and-videos/

[45] R. K. Kaliyar, A. Goswami, and P. Narang, "DeepFakE: Improving fake news detection using tensor decomposition-based deep neural network," *J. Supercomput.*, vol. 77, pp. 1015–1037, 2021. doi: 10.1007/s11227-020-03294-y.

[46] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, "Understanding deep neural networks with rectified linear units," 2018. doi: 10.48550/arXiv.1611.01491.

[47] D. Guera, S. Baireddy, P. Bestagini, S. Tubaro, and E. J. Delp, "We need no pixels: Video manipulation detection using stream descriptions," 2019. doi: 10.48550/arXiv.1906.08743.

[48] T. Jung, S. Kim, and K. Kim, "Deep vision: Deep fakes detection using human eye blinking pattern," *IEEE Access*, vol. 8, pp. 83144–83154, 2020. doi: 10.1109/ACCESS.2020.2988660.

[49] U. A. Ciftci and I. Demir, "FakeCatcher: Detection of synthetic portrait videos using biological signals," in *IEEE Trans. on Pattern Anal. and Mach. Intell.*, IEEE, 2020. doi: 10.1109/TPAMI.2020.3009287.

[50] H. Qi *et al.*, "Deeprhythm: Exposing deepfakes with attentionsl visual heartbeat rhythms," in *Proc. 28th ACM Int. Conf. on Multimedia*, New York, NY, USA, Association for Computing Machinery, 2020, pp. 4318–4327. doi: 10.1145/3394171.3413707.

[51] S. Fernandes *et al.*, "Predicting heart rate variations of deep fake videos using neural ode," in *2019 IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Piscataway, NJ, USA, IEEE, 2019, pp. 1–9. doi: 10.1109/ICCVW.2019.00213.

[52] A. Chintha *et al.*, "Recurrent convolutional structures for audio spoof and video deep fake detection," *IEEE J. Sel. Top. Signal Process.*, vol. 15, no. 5, pp. 1024–1037, 2020. doi: 10.1109/JSTSP.2020.2999185.

[53] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, "Deep fake detection using spatiotemporal convolutional networks," 2020. doi: 10.48550/arXiv.2006.14749.

[54] C. Yu, J. Li, X. Li, X. Ren, and B. B. Gupta, "Four-image encryption scheme based on quaternion Fresnel transform, chaos and computer-generated hologram," *Multimed. Tools. Appl.*, vol. 77, pp. 4585–4608, 2018. doi: 10.1007/s11042-017-4637-6.

[55] N. T. U. Nhi and T. M. Le, "A model of semantic-based image retrieval using C-tree and neighbor graph," *Int. J. Semant. Web Inf. Syst.*, vol. 18, no. 1, pp. 1–23, 2022. doi: 10.4018/IJSWIS.

[56] H. Khalid, M. Kim, S. Tariq, and S. S. Woo, "Evolution of an audio-video multimodal deep fake dataset using unimodal and multimodal detectors," in *Proc. 1st Workshop on Synthetic Multimedia-Audiovisual Deepfake Generation and Detection*, New York, NY, USA, Association for Computing Machinery, 2021, pp. 7–15. doi: 10.1145/3476099.3484315.

[57] H. Ilyas, A. Javed, and K. M. Malik, "AVFakeNet: A unified end-to-end dense swin transformer deep learning model for audio-visual deepfakes detection," *Appl. Soft Comput.*, vol. 136, pp. 110124, 2023. doi: 10.1016/j.asoc.2023.110124.