Journal of Cybersecurity and Privacy

MDPI

Article

Comparison of Deepfake Detection Techniques through Deep Learning

Maryam Taeb [1] and Hongmei Chi [2,*]

[1] Electrical and Computer Engineering, FAMU-FSU College of Engineering, Tallahassee, FL 32310, USA; mr21cg@my.fsu.edu
[2] Department of Computer Sciences, Florida A&M University, BBTA RM 309, 1333 Wahnish Way, Tallahassee, FL 32307, USA
* Correspondence: hongmei.chi@famu.edu; Tel.: +1-850-599-3050

Abstract: Deepfakes are realistic-looking fake media generated by deep-learning algorithms that iterate through large datasets until they have learned how to solve the given problem (i.e., swap faces or objects in video and digital content). The massive generation of such content and modification technologies is rapidly affecting the quality of public discourse and the safeguarding of human rights. Deepfakes are being widely used as a malicious source of misinformation in court that seek to sway a court's decision. Because digital evidence is critical to the outcome of many legal cases, detecting deepfake media is extremely important and in high demand in digital forensics. As such, it is important to identify and build a classifier that can accurately distinguish between authentic and disguised media, especially in facial-recognition systems as it can be used in identity protection too. In this work, we compare the most common, state-of-the-art face-detection classifiers such as Custom CNN, VGG19, and DenseNet-121 using an augmented real and fake face-detection dataset. Data augmentation is used to boost performance and reduce computational resources. Our preliminary results indicate that VGG19 has the best performance and highest accuracy of 95% when compared with other analyzed models.

Keywords: deepfake detection; digital forensics; media forensics; deep learning; VGG19; face-image manipulation

Citation: Taeb, M.; Chi, H. Comparison of Deepfake Detection Techniques through Deep Learning. *J. Cybersecur. Priv.* **2022**, *2*, 89–106. https://doi.org/10.3390/jcp2010007

Academic Editors: Phil Legg and Giorgio Giacinto

Received: 10 January 2022
Accepted: 21 February 2022
Published: 4 March 2022

1. Introduction

In the last few years, cybercrime, which accounts for a 67% increase in the incidents of security breaches, has been one of the most challenging problems that national security systems have had to deal with worldwide [1]. Deepfakes (i.e., realistic-looking fake media that has been generated by deep-learning algorithms) are being widely used to swap faces or objects in video and digital content. This artificial intelligence-synthesized content can have a significant impact on the determination of legitimacy due to its wide variety of applications and formats that deepfakes present online (i.e., audio, image and video).

Considering the quickness, ease of use, and impacts of social media, persuasive deepfakes can rapidly influence millions of people, destroy the lives of its victims and have a negative impact on society in general [1]. The generation of deepfake media can have a wide range of intentions and motivations, from revenge porn to political fake news. Rana Ayyub, an investigative journalist in India, became a target of this practice when a deepfake sex video showing her face on another woman's body was circulated on the Internet in April 2018 [2]. Deepfakes have also been published to falsify satellite images with non-existent landscape features for malicious purposes [3].

There are numerous captivating applications of deepfakery in video compositing and transfiguration in portraits, especially in identity protection as it can replace faces in photographs with ones from a collection of stock images. Cyber-attackers, using various

strategies other than deepfakery, are always aiming to penetrate identification or authentication systems to gain illegitimate access. Therefore, identifying deepfake media using forensic methods remains an immense challenge since cyber-attackers always leverage newly published detection methods to immediately incorporate them in the next generation of deepfake generation methods. With the massive usage of the Internet and social media, and billions of images available on the Internet, there has been an immense loss of trust from social media users. Deepfakes are a significant threat to our society and to digital evidence in courts. Therefore, it is highly important to obtain state-of-the-art techniques to identify deepfake media under criminal investigation.

As demonstrated in Table 1 (inspired by the figure presented in [1]), tampering of evidence, scams and frauds (i.e., fake news), digital kidnapping associated with ransomware blackmailing, revenge porn and political sabotage are among the vast majority of types of deepfake activities with the highest level of intention to mislead [1].

Table 1. Deepfake Information trust Table.

Type of Media	Examples	Intention to Mislead	Level of Truth
Hoax	Tampering evidence Scam and Fraud Harming Credibility	High	Low
Entertainment	Altering movies Editing Special effects Art Demonstration	Low	Low
Propaganda	Misdirection Political Warfare Corruption	High	High
Trusted	Authentic Content	Low	High

The first deepfake content published on the Internet was a celebrity pornographic video that was created by a Reddit user (named deepfake) in 2017. The Generative Adversarial Network (GAN) was first introduced in 2014 and used for image-enhancement purposes only [4]. However, since the first published deepfake media, it has been unavoidable for deepfake and GAN technology to be used for malicious uses. Therefore, in 2017, GANs were used to generate new facial images for malicious uses for the first time [5]. Following that, there has been a constant development of other deepfake-based applications such as FakeApp and FaceSwap. In 2019, Deepnude was developed and provided undressed videos of the input data [6]. The widespread strategies used to manipulate multimedia files can be broadly categorized into the following major categories: copy–move, splicing, deepfake, and resampling [7]. Copy–move, splicing and resampling involve repositioning the contents of a photo, overlapping different regions of multiple photos into a new one, and manipulating the scale and position of components of a photo. The final goal is to manipulate the user by conveying the deception of having a larger number of components in the photograph than those that were initially present. Deepfake media, however, leveraging powerful machine-learning (ML) techniques, have significantly improved the manipulation of the contents. Deepfake can be considered to be a type of splicing, where a person's face, sound, or actions in media is swiped by a fake target [8]. A wide set of cybercrime activities are usually associated with this type of manipulation technique, and while spreading them is easy, correcting the records and avoiding deepfakes are harder [9]. Consequently, it is becoming harder for machine-learning techniques to identify convolutional traces of deepfake generation algorithms, as there needs to be frequency-specific anomaly analysis. The most basic algorithms that were being used to train models for the task of deepfake detection such as Support Vector Machine (SVM), Convolution Neural Network (CNN), and Recurrent Neural Network (RNN) are now being coupled with multi-attentional [10] or ensemble [11] methods to

increase the performance and address weakness of other methods. As proposed by [12], by implementing an ensemble of standard and attention-based data-augmented detection networks, the generalization issue of the previous approaches can be avoided. As such, it is of high importance to identify the most suitable algorithms for the backbone layers in multi-attentional and ensembled architectures. As generation of deepfake media only started in 2017, academic writing on the problem is meager [13]. Most of the developed and published methods/techniques are focused on deepfake videos. The main difference between deepfake video- and image-detection methods is that video-detection methods can leverage spatial features [14], spatio-temporal anomalies [15] and supervised domain [16] to draw a conclusion on the whole video by aggregating the inferred output both in time and across multiple faces. However, deepfake image-detection techniques have access to one face image only and mostly leverage pixel- [17] and noise-level analysis [18] to identify the traces of the manipulation method.

Therefore, identifying the most reliable methods for face-image forgery detection that relies on convolutional neural networks (CNN) as the backbone for a binary classification task could provide valuable insight for the future direction in the development of deepfake-detection techniques. The overall approach taken in this work is illustrated in Figure 1.

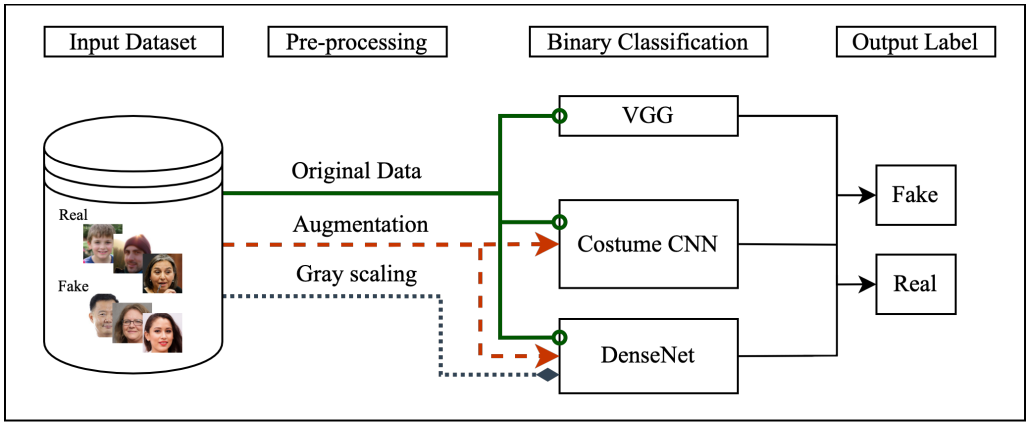

Figure 1. General overview of our proposed approach to detect deepfake media in a digital forensics scenario.

DenseNet has shown significant promise in the field of facial recognition. DenseNet as an extension of Residual CNN (ResNet) architecture has addressed the low-supervision problem of all its counterparts by initiating a between-layer connection using dense blocks. The dense blocks in the DenseNet architecture improve the learning process by leveraging a transition layer (essentially convolution, average pooling, and batch normalization between each dense block) that concatenates feature maps. As such, gradients from the initial input and loss function are shared by all the layers. The described implementation reduces the number of required parameters and feature maps, and consequently provides a less computationally expensive model. Therefore, we have decided to test DenseNet's capabilities and compare it with other neural network architectures.

VGG-19, as an algorithm that has been widely used to extract the features of the detected face frames [19], was chosen to be compared with the DenseNet architecture. VGG-19's architecture eases the face-annotation process by forming a large training dataset with the use of online knowledge sources that are then used to implement deep CNNs to perform the task of face recognition. The formed model is then evaluated on face recognition benchmarks to analyze model efficiency regarding the generation of facial features. During this process, VGG-19 is trained on classifiers with sigmoid activation function in the output layer which produces a vector representation of facial features (face embedding) to fine-tune the model. The fine-tuning process differentiates class similarities using Euclidean distance that is achieved using a triplet loss function that aims at comparing Euclidean spaces of similar and different faces using learning score vectors. The CNN architecture

implemented in VGG-19 implements fully connected classifiers that include kernels and ReLU activation followed by maxpooling layers.

Finally, we have implemented a Custom CNN architecture to evaluate the performance of previously described algorithms and analyze the effectiveness of dropout, padding, augmentation and grayscale analysis on model performance.

This study aims to provide an in-depth analysis on the described algorithms, structures and mechanisms that could be leveraged in the implementation of an ensembled multi-attentional network to identify deepfake media. The result of this work contributes to the nascent literature on deepfakery by providing a comparative study on effective algorithms for deepfake detection on facial images within the possible use of digital forensics in criminal investigations.

The rest of this paper is organized as follows. Section 2 provides a literature review of the algorithms and datasets that are widely used for deepfake detection. Section 3 provides details on the analysis methods and configurations of the compared algorithms as well as with the details on the tested dataset. Section 4 provides the results of the comparative analysis. Finally, Section 5 concludes with implications, limitations, and suggestions for future research.

2. Literature Review

Anti-deepfake technology can be divided into three categories: (1) detection of the deepfake; (2) authentication of the published content; and (3) prevention of the spread of contents that can be used for deepfake production. Technology towards detection and authentication of deepfakery is growing fast; however, the capacity to generate deepfakes is proceeding much faster than the ability to detect them. Twitter has reported attempts to publish misinformation and fake media by 8 million accounts per week [20]. There has been a wide variety of deepfake media, and the detection techniques that have been used to identify them is shown in Figure 2. This has created a massive challenge for researchers to provide a solution that can promptly analyze all the posted material on the Internet and social media platforms to identify deepfakes. Previous research has mostly aimed at improving previously developed technologies to train a new detection system.

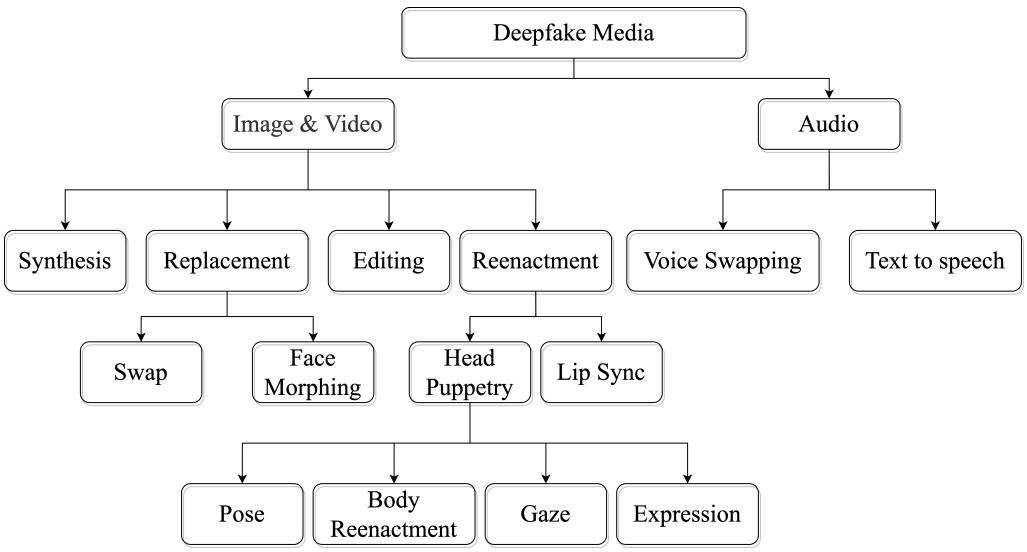

Figure 2. Current deepfake media types and detection techniques.

2.1. Deepfake Detection Datasets

Deepfake detection systems typically leverage binary classifiers to cluster information into real and fake classes. This method requires a great quantity of good-quality authentic and tampered data to train classification models. The first known datasets that had a great impact on the growth and improvement of deepfake detection technologies

were UADFV [21] and DFTIMIT [22]. FaceForensics++ dataset includes 977 downloaded videos from YouTube, provides 1000 sequences of original unobstructed faces, as well as their manipulated versions. The manipulated versions were generated by four methods: Deepfakes, Face2Face, FaceSwap and NeuralTextures [23]. The DeepFakeDetection dataset (DFD) released by Google in collaboration with Jigsaw contains over 363 original sequences from 28 paid actors in 16 different scenes as well as over 3000 manipulated videos using deepfakes [23]. The Deepfake Detection Challenge (DFDC) dataset [24] published by Facebook is another publicly available large dataset that includes over 100,000 total clips from 3426 actors, produced with deepfake, GAN-based and unsupervised models. Celeb-DF (v2) [25] dataset published by [25] is an extension to Celeb-DF (v1) that contains real and fake videos that are generated via deepfake algorithm by providing images with the same quality as the synthesized videos circulating online. This dataset provides 5639 videos with subjects of different ages, ethnic groups and genders, and their corresponding deepfake videos. The DeeperForensics-1.0 dataset is a large-scale benchmark for face forgery detection that represents the largest face forgery detection dataset by far. This benchmark includes 60,000 videos forming a total of 17.6 million frames generated by an end-to-end face-swapping framework. Furthermore, extensive real-world perturbations are applied to obtain a more challenging benchmark of larger scale and higher diversity [26].

For our research and analysis, we took the "Real and Fake Face-Detection" dataset from Yonsei University [27] that contains expert-generated high-quality PhotoShopped face images. The dataset includes 960 fake and 1081 real images that are composites of different faces, separated by eyes, nose, mouth, or whole face. The second dataset that has been used in this work is the "140K Real and Fake Faces" that consists of 70K real faces from the Flickr dataset collected by Nvidia, as well as 70K fake faces sampled from the 1 million fake faces (generated by StyleGAN) that were published by Bojan [28]. These two datasets were used to include both GAN-generated images along with expert/human-generated images to provide many good-quality data. All the above-mentioned datasets can be used for image and video classification, segmentation, generation and augmentation of new data. Table 2 represents a cumulative comparison of the mentioned datasets; please note that the rows with a "*" sign include images only (not videos). Deepfake datasets have been categorized into two generations based on several factors and elements. Considering release time and synthesis algorithms involved in the generation of the data, UADFV and DF-TIMIT are categorized as the first generation. Considering the quality and quantity of the generated data, DFD, DeeperForensics, DFDC, and the Celeb-DF datasets are categorized as the second generation [25].

Table 2. Comparison of publicly available deepfake datasets.

	Real		Fake				
Dataset	**Video**	**Frame**	**Video**	**Frame**	**Generation Method**	**Release Date**	**Generation Group**
UADFV	49	17.3K	49	17.3K	FakeAPP	11/2018	1st
DF-TIMIT	320	34K	320	34K	Faceswap-GAN	12/2018	1st
*Real & Fake Face Detection	1081	405.2K	960	399.8K	Expert-generated high-quality photoshopped	01/2019	2st
FaceForensics++	1000	509.9k	1000	509.9K	DeepFakes, Face2Face, FaceSwap, NeuralTextures	01/2019	2nd
DeepFakeDetection	363	315.4K	3068	2242.7K	Similar to FaceForensics++	09/2019	2nd
DFDC	1131	488.4K	4113	1783.3K	Deepfake, GAN-based, and non-learned methods	10/2019	2nd
Celeb-DF	590	225.4K	5639	2116.8K	Improved DeepFake synthesis algorithm	11/2019	2nd
*140K Real & Fake Faces	70K	15.8M	70K	15.8M	StyleGAN	12/2019	2nd
DeeperForensics	50,000	12.6M	10,000	2.3M	Newly proposed end-to-end face swapping framework	06/2020	2nd

2.2. Deepfake Detection Algorithms

Deepfake detection techniques aim to conceal revealing traces of deepfakes by extracting semantic and contextual understanding of the content. Research in the field of media forensics provides a wide range of imperfections as indicators of fake media: face wobble, shimmer and distortion; waviness in a person's movements; inconsistencies with speech

and mouth movements; abnormal movements of fixed objects such as a microphone stand; inconsistencies in lighting, reflections and shadows; blurred edges; angles and blurring of facial features; lack of breathing; unnatural eye direction; missing facial features such as a known mole on a cheek; softness and weight of clothing and hair; overly smooth skin; missing hair and teeth details; misalignment in face symmetry; inconsistencies in pixel levels; and strange behavior of an individual doing something implausible are all the indicators and features used by deepfake detection algorithms [13]. The use of deep-learning techniques and algorithms such as CNN and GAN has made deepfake detection more challenging for forensics models because deepfakes can preserve pose, facial expression and lighting of the photographs [29]. Frequency domain, JPEG Ghost and Error Level Analysis (ELA) are among the first methods that were used to identify manipulation traces on images. However, they are not successful in identifying manipulated images that are generated with deep-learning and GAN algorithms. Neural networks are one of the most widely used methods for deepfake detection. There are some proposals on the usage of X-rays [18], and spectrograms [30] to identify traces of blending and noise in deepfake media. However, such methods cannot detect random noise and suffer from a performance drop when encountering low-resolution images. Deepfakes are implemented mainly using a CNN that generates deepfake images and an encoder–decoder network structure (ED), or GAN [4] that synthesizes fake videos. Deepfake detection techniques focused on anomalies in the face region only can be categorized into holistic and feature-based matching techniques [31]. The holistic techniques, which are mostly used to identify deepfake face images and include Principal Component Analysis (PCA), Support Vector Machines (SVM), and CNN, mainly analyze the face as a whole. These techniques aim at reducing data dimensionality by forming a smaller set of linear combinations of the image pixels that are then fed to a binary classifier to identify authentic and fake images. Feature-based or attention-based matching techniques, however, are used for both deepfake video and image identification, and split the whole face into different regions of focus such as eye, nose, lips, skin, head position, color mismatches, etc. [32]. Holistic techniques are successful in detecting localized deepfake characteristics (i.e., anomalies in the face and jaw region) and can be leveraged to identify specific feature characteristics (eyes, nose, mouth) that could be significant in detection [12]. Convolutional Neural Network (CNN)-based image classification and recognition models have been proven to be trainable to classify manipulated images from authentic ones [33]. Luca et al. [34] aimed to extract and detect fingerprints that represent convolution traces left in the process of generating GAN images using the Expectation-Maximization algorithm. Wang et al. [35] demonstrated that with careful pre- and post-processing and data augmentation, a standard classifier trained on ProGAN, an unconditional CNN generator can be generalized surprisingly well to unseen architectures, datasets, and training methods. CNN have also been trained to detect manipulation techniques such as lack of eye-blinking [36], missing details in eyes from an image [37], and facial wrapping artifacts. Furthermore, CNNs have been shown to be able to capture distinctive traces of generation methods that have worked on further wrapping the faces with high-resolution sources [17].

VGG19 and VGG16 has significantly improved large-scale fake image recognition by increasing the layer depth (23/26 layers) of CNN-based models [38]. Chang et al. [39] presented an improved VGG network, namely NA-VGG, based on image augmentation and noise-level analysis to detect a deepfake face image. The experimental results using the Celeb-DF dataset shows that NA-VGG improved accuracy over other state-of-the-art fake image detectors. Kim et al. [40] demonstrated that VGG-16 has a better performance than the ShallowNet architecture to classify genuine facial images from disguised face images.

Furthermore, DenseNet architecture has also been demonstrated to be computationally more efficient with its feed-forward design network, which connects each layer to every other layer [41]. In DenseNet architecture, feature maps of all former layers are used as the input for each layer. DenseNet requires significantly fewer parameters and computation to achieve state-of-the-art performance [33]. Hsu, Chih-Chung, Yi-Xiu Zhuang, and Chia-Yen

Lee [42] in their work proposed a fake face-image detector based on the novel CFFN, consisting of an improved DenseNet backbone network and Siamese network architecture. Their comprehensive analysis demonstrated that deep features-based deepfake-detection systems such as DenseNet obtain significant accuracy when trained and tested on the same kind of manipulation technique.

Feature-based techniques have started identifying the deficiencies of deepfake generation methods such as unnatural eye-blinking patterns and temporal flickering, which gave rise to a more improved generation of deepfake models that were trained on datasets that have addressed the identified deficiencies. Yang et al. [43] demonstrated that facial landmarks could be used to provide an estimate of head posture direction. The work of [44,45] illustrated that eye pupils' inconsistencies are one of the indicators of fake media. Some studies [46] including audio into the training process have illustrated that the difference between lip movements and voice matching distinguishes real and fake media. There have been some efforts on domain-specific deepfake detection such as [47] that leveraged forensic techniques to model political leaders' facial expressions and speaking patterns; however, it would be a more difficult task to train and generalize such approach for the whole world. Even though feature-based techniques are more robust to deformations, they have been mainly designed to have the best performance on domain-specific datasets. Holistic techniques are competent in learning human faces and extracting higher-dimensional semantic features for classification.

Other techniques that leverage spatial features and spatio-temporal anomalies in the supervised domain such as Xception [48] and EfficientNet [49] have been shown to be more efficient than CNNs. Xception architecture claims to gain a more efficient use of model parameters due to depthwise separable convolutions that can understand as an inception module. Kumar and Bhavsar [16] demonstrated that Xception combined with metric learning can enhance the classification in high-compression scenarios. They were able to achieve an AUC score of 99.2% and accuracy of 90.71% for deepfake video identification on the Celeb-DF dataset. Ismail et al. [14] in their experimental analysis demonstrated that XceptionNet combined with an additional Bi-LSTM and LSTM layer can achieve a 79% ROC-AUC score. Li et al. [50] demonstrated that Xception does not have a good performance on face-image datasets (AUC of 73.2) and, furthermore, it has a high true-negative rate while having the lowest true-positive rate. To summarize, Xception may provide better performance for fake video detection; however, it does not address the generalizability issue across different datasets and does not perform well when fed with images only. EfficientNET proposes a new scaling method that uniformly scales all dimensions of depth/width/resolution using compound coefficient. Coccomini et al. [15] were able to achieve an AUC of 0.95% and F1-score of 88% on the DFDC dataset. Pokroy and Egorov [51] demonstrated that an increased scale in all dimensions may not always lead to higher accuracy due to the fact that CNNs will have to deal with more complex patterns that are difficult to transfer to a different task. Mitra et al. [52] were able to achieve a 96% accuracy on the FaseForensics++ dataset by making the complexity of detecting forged videos low using the depthwise separable convulsion of EfficientNet. In conclusion, Xception and EfficientNet, by uniformly scaling all dimensions, can gain a more efficient use of model parameters. Furthermore, they can extract spatial features and spatio-temporal anomalies by aggregating the inferred output both in time and across multiple faces due to their depthwise separable convolutions. These methods have illustrated that they can draw an improved conclusion on the whole video; however, they have not demonstrated any improvements to deepfake classification on a single image (i.e., deepfake image-detection).

Recent scholarly work has been focused on implementing an ensemble of holistic and feature-based detection networks by addressing the drawbacks of both methods. Dolescki et al. [53], in their work implementing a classification method, which involves a collection of classifiers with a certain utility function regarded as an aggregation operator, were able to achieve accuracy of 87%. Silva et al. [12] were able to achieve a 92% accuracy on the DFDC dataset by implementing a hierarchical explainable forensics algorithm that

incorporates humans in the detection loop. Hanqing et al. [10] proposed a multi-attentional deepfake detection network that can achieve a 97% accuracy by implementing multiple spatial attention heads, textural feature enhancement blocks and aggregating low-level textural features and high-level semantic features. Bonettini et al. [11] were able to achieve AUC of 87% on DFDC by assembling different trained Convolutional Neural Network (CNN) models that combined EfficientNetB4 with attention layers and Siamese training. Du et al. [54] demonstrated that a good balance between accuracy and efficiency can be achieved with two separated EfficientNet architectures that simultaneously analyze raw content and its frequency-domain representation.

Given that the most successful approaches to identifying and preventing deepfakes are deep-learning methods that rely on CNNs as the backbone for a binary classification task [12], and a large 2D CNN model can prove to be better than EfficientNet model if deepfake classification is the only desired result [55], we have evaluated the most common backbone architecture of existing developed frameworks (CNN, VGG-19 and DenseNet) that are demonstrated to have the best performance on the task of deepfake image classification.

3. Approach

Our proposed method for deepfake detection on images is shown in Figure 1. We have taken two different classification procedures in this work. As shown in both Figures 1 and 3, input data goes through the same procedure with the same architecture; however, Figure 3 demonstrates a second round of analysis with an additional post-processing classification step that has been added to the last output layer of the analyzed models. The second round of analysis with additional post-processing was performed to analyze the effects of principal component analysis on the task of deepfake classification. Further details about the post-processing step are described in the final paragraphs of the evaluation subsection of this section.

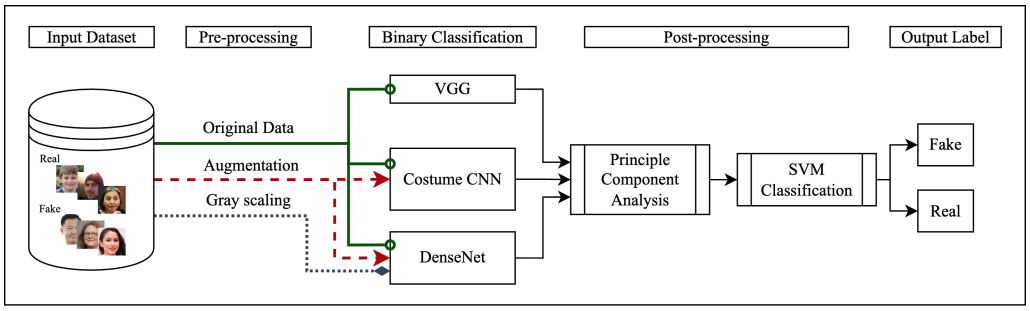

Figure 3. Detailed steps of post-processing in our proposed approach for deepfake detection.

3.1. Implementation

Input data are a dataset that is labeled and clustered into two categories of real and fake. They are augmented for training purposes using the following specifications:

- Rotation range of 20 for DenseNET and no rotation on Custom CNN
- Scaling factor of 1/255 was used for coefficient reduction
- Shear range of 0.2 to randomly apply shearing transformations
- Zoom range of 0.2 to randomly zoom inside pictures
- Randomized images using horizontal and vertical flipping

After augmentation, the face images are classified as either fake or real using three different models: Custom CNN, VGG, and DenseNET. We defined two classes for our binary classification task: 0 to denote the real (e.g., normal, validation, and disguised face images) and 1 to denote fake (e.g., impersonator face images) groups, respectively.

The "Real and Fake Face-Detection" dataset was used to train the three models at a learning rate of 0.001 and for 10 epochs. The test accuracy was then calculated using the test set. We applied data augmentation to flip all original images horizontally and vertically,

hence a three-fold increase of the dataset size (original image + horizontally flipped image + vertically flipped image).

The Custom CNN architecture included six convolution layers (Conv2D) each paired with batch normalization, max pooling and dropout layers. Rectified Linear Unit (ReLU) and sigmoid activation functions were applied for the input and output layers respectively. Dropout was applied to each layer to minimize over-fitting and padding was also applied to the kernel to allow for a more accurate analysis of images. The Custom CNN architectures have been trained and validated on the original and augmented datasets with a 1/255 scaling factor. Data augmentation was performed to observe effects of data aggregation on model performance and promote the generalizability of the findings. Details on augmentation process includes horizontal flip along with a 0.2 zoom range, shear range of 0.2 along with rescaling factor to avoid image quality to factor in model behavior during classification since not all the images had the same pixel-level quality.

Following a similar approach to [56], **the VGG-19** model that was used is a 16-layer CNN architecture paired with three fully connected layers, five maxpooling layers and one SoftMax layer that is modeled from architectures in [56]. VGG-19 has been pretrained on a wide variety of object categories, which leads to its ability to learn rich feature representations. VGG-19 has demonstrated that it can provide a high accuracy level when classifying partial faces. This architecture demonstrated that its highest accuracy is accessible when its size is increased [57]; therefore, we have applied a high-end configuration to it by adding a dense layer after the last layer block that provides the facial features and added a dense layer as the output layer with sigmoid activation function to fine-tune the model for the task of deepfake detection.

The DenseNET architecture used in this work is Keras's DenseNet-264 architecture with an additional dense layer as the last output layer. This architecture starts with a 7×7 stride 2 convolutional layer followed by a 3×3 stride-2 MaxPooling layer. It also includes four dense blocks paired with batch normalization and ReLU activation function for the input layers and sigmoid activation function for the output layer. Furthermore, there are transition layers between each denseblock that include a 2 by 2 average pooling layer along with a 1 by 1 convolutional layer. The last dense block is followed by a classification layer that leverages the feature maps of all layers of the network for the task of classification which we have coupled with a denseblock with the sigmoid activation function as the output layer. This model was trained on 100,000 images and validated on 20,000 images. This model has been trained and validated on the original, grayscale and augmented datasets with a 1/255 scaling factor too. We aimed to add to the diversity of the training data by performing **augmentation to the DenseNet architecture** by applying a horizontal flip, a 20 range rotation along with the same rescaling procedure that was applied in the Custom CNN architecture. Because pixel-level resolution of grayscale and color images are different, we have also measured the importance of color on model behavior towards classifying data into the fake and real categories by training the DenseNet architecture on grayscale only data too. The VGG architecture, however, was only trained and tested on the original dataset. All the analyzed models in this work are used as they were designed with an additional custom dense layer with sigmoid activation function. The rationale behind adding this layer to all models was to add a useful rectifier activation function layer for the task of binary classification to produce a probability output in the range of 0 to 1 that can easily and automatically be converted to crisp class values.

3.2. Evaluation

The performance of the described models is assessed with accuracy, precision, recall, F1-score, average precision (AP) and area under the ROC curve.

Accuracy, simply put, indicates how close the model prediction is to the target or actual value (fake vs. real), meaning how many times the model was able to make a correct predication among all the predictions it has made. Equation (1) indicates the overall

formula used to calculate prediction, where *TPR* stands for true prediction and *TOPR* stands for total predictions made by the model.

$$Accuracy = \frac{TPR}{TOPR} \tag{1}$$

Precision, on the other hand, refers to how consistent results are regardless of how close to the true value they are using the target label. Equation (2) demonstrates the ratio that indicates the proportion of positive identifications by model that were actually correct. *TP* in Equation (2) stands for the number of true positives and *FP* stands for the number of false positives.

$$Percision = \frac{TP}{TP + FP} \tag{2}$$

The recall is the proportion of actual positives that were identified by the model that were correct. Equation (3) demonstrates this ratio where *TP* is the number of true positives and *FN* the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The F1-score, by taking into account both precision and recall, balances the precision and recall and indicates model ability to accurately predict both true-positive and true-negative classes. The F1 score can be interpreted as a harmonic mean of the precision and recall. For the task of deepfake classification, F1-score is a better measure to assess model performance, since both classes are of importance and the relative contribution of precision and recall to the F1 score are better than equal. Equation (4) demonstrates how F1-score is calculated.

$$\mathrm{F1} = \frac{2 * (Percision * Recall)}{Percision + Recall} \tag{4}$$

Average Precision (AP) was used as an aggregation function for the task of object detection to summarize the precision–recall curve as the weighted mean of precision achieved at each threshold, with the increase in recall from the previous threshold used as the weight based on Equation (5), where R_n and P_n are the precision and recall at the nth threshold [58].

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \tag{5}$$

Finally, as shown in Figure 3, the output vectors of the final hidden layer of the analyzed architectures were extracted and treated as a representation of the images. Dimensions of the vectors for the Custom CNN architecture, VGG-19 and DenseNet architectures were 512, 2048 and 1024, respectively. Principal Component Analysis (PCA) was performed to keep the most dominant variable vector points and preserved 50 principal components. The resulting vectors from the PCA were fed into a support vector machine (SVM) to classify them into the two classes of real and fake.

4. Preliminary Results

This section provides the results obtained from the three different neural network architectures that have been tested in this work. The dataset section provides an overview of the advantages, drawbacks, and improvements of the datasets described in the literature review.

4.1. Dataset

Deepfake datasets should have careful consideration of quality, scale, and diversity. UADF and DFTMIT provide a baseline dataset for preliminary analysis in deepfake de-

tection; however, they lack the quantity and diversity elements. The DeepFakeDetection dataset extends the preliminary FaceForensics dataset; however, it contains relatively few videos with few subjects and limited size and number of methods that are represented. The DFDC dataset addresses the drawbacks of the previously published datasets by providing a large number of clips, of varying quality, and with a good representation of the current state-of-the-art face-swap methods. However, it still has various visual artifacts that make them easily distinguishable from the real videos. The DFDC dataset resolves the limited availability of source footage, few videos and fewer subjects; however, the Celeb-DF dataset provides more relevant data to evaluate and support the future development of deepfake detection methods by fixing color mismatch, inaccurate face masks, and temporal flickering of previously discussed datasets. Finally, deeper forensics, by addressing the drawbacks of all mentioned datasets, provides a benchmark of larger scale and higher diversity that can be leveraged to achieve the best performance of deepfake detection algorithms. Table 3 summarizes the drawbacks and improvements of the described datasets.

Table 3. Dataset analysis summary.

Dataset	Drawbacks	Improvements
UADF DFTMIT	Lack of quantity and Diversity	Suitable baseline
DFD	Limited size and methods	Extension to FaceForensics dataset
DFDC	Distinguishable visual artifacts	Large number of clips of varying quality
Celeb-DF	Low realness score Biased: impractical for face Forgery detection	Fixed color mismatch Accurate face masks
Deeper Forensics-1.0	Challenging as a test database	High realness score

The mentioned datasets include videos that could be used for face detection in images; however, the "Real and Fake Face-Detection" dataset combined with the "140K Real and Fake Faces" includes both GAN-generated images as well as expert/human-generated images, and is considered by far one of the largest available face-image datasets. The two described datasets together include 70,960 fake and 71,081 real images. As shown in Table 4, 70K of the fake images are GAN-generated and 960 of them are human expert-generated. Similarly for the real images, 70K of them are GAN-generated and 1081 of them are human expert-generated. The distribution of the human-generated fake images is not balanced with the GAN-generated photos, but this is the largest human-generated image dataset available currently.

Table 4. Distribution of the used datasets.

Generation Method	Fake	Real
GAN	70K	70K
Human Expert	960	1081

4.2. Algorithms

The accuracy, precision and recall rates of analyzed models demonstrated in Table 5, the ROC curve demonstrated in Figure 4, the area under the ROC curve (AUC), F1-scores and AP results demonstrated in Table 6 were used to evaluate model performance in terms of separability and their ability to differentiate between classes. The algorithm comparison results revealed that the VGG-19 model had the best performance among all 3 other algorithms, with an accuracy level of 95%.

The results of this study demonstrate that VGG-19 can be a suitable choice not only for partial face images, but also for full-face images confirming the findings of [57]. The better performance of VGG-19 is because it is pretrained on a wide variety of objects. AP was used as an aggregation function to summarize the precision–recall curve into a single value that represents the average of all precisions. VGG-19, even though it had the highest accuracy, had the lowest AP of 95% in comparison to all other analyzed models. The DenseNet architecture on the original dataset and grayscale dataset had a closer performance to VGG-19, with 94% accuracy. Results from DenseNET architecture demonstrates that gray channel-based analysis does not have a huge impact on model accuracy level in classifying images into the two categories of real and fake. The DenseNet architecture, even though was second best in terms of performance, achieved an AP of 99% on both augmented and grayscale datasets, which is slightly in contrast to the results found in [59] in terms of precision rate; however, it aligns with claims regarding detection time. Custom CNN architecture had the lowest accuracy level (89%). The second-highest AP score after DenseNet was the Custom CNN model. Augmented input reduced model performance and accuracy level on both DenseNET and Custom CNN by 5–22%. However, the Custom CNN had a better performance on augmented data in comparison to the DenseNet architecture. Precision and recall rates from DenseNet architecture trained on augmented data suggest that the final dense block that we have coupled with the DenseNet classification layer did not have a positive impact on model behavior. The issue with reduced performance on augmented data might be resolved by training the model for a larger number of epochs, since augmentation results in harder training samples. VGG-19, even though it was great in terms of performance, aligns with results from [60]; it was computationally very expensive, especially if fed with augmented data. DenseNET was computationally more efficient in comparison to VGG-19 and Custom CNN, which aligns with the results from [40]. The F1-score of the DenseNet architecture on grayscale was the highest, reaching 97% suggesting it could be a suitable backbone when dealing with unbalanced class distribution in their dataset. The second-highest F1-score was achieved by VGG-19, as it achieved a 95% F1-score. The lowest F1-score was achieved by the Custom CNN on augmented data, as the F1-score was only 85%. Taking F1-score as a measurement to balance precision and recall, DenseNet on grayscale data might seem to be a better solution, however, since the dataset used for training in this analysis had a balanced class distribution accuracy level and is a better judge in this analysis. The results from the PCA-SVM classification demonstrated that VGG-19 was able to form a distinctive cluster of fake and real images using the PCA vectors as a representation of the image (demonstrated in Figure 5). Custom CNN architectures and DenseNet trained on the original and augmented datasets showed decent classification. However, DenseNet trained on grayscale images presented very poor performance (Table 5).

Table 5. Algorithm comparison results. OD stands for Original Dataset, AD stands for Augmented Dataset and GS stands for Grayscale Dataset.

	Model Performance			**PCA-SVM Performance**		
Architecture	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
VGG-19	95	93	97	99	99	99
DenseNet OD	94	92	96	98	98	98
DenseNet AD	73	66	95	86	86	86
DenseNet GS	94	91	99	50	50	47
Custom CNN OD	89	91	87	97	97	97
Custom CNN AD	84	87	79	91	90	91

Overall analysis of the results reveal that all the architectures had a higher efficiency in detection and classification of GAN-generated images due to the traces that GAN

generators left on the generated media. Considering VGG-19's performance and behavior, even though it may not be the most computationally efficient model, it had a competitively better performance than the other analyzed model and it showed a promising improvement when coupled with PCA-SVM classification layers. This suggests that VGG-19 could be a more suitable backbone architecture for the task of deepfake detection related to the essential technical and legal requirements that determine evidence admissibility. Deepfakes are a threat to the admissibility of digital evidence in courts. Quick and effective detection of authentic media is critical in any criminal investigations. VGG-19 could be a fast solution for detecting deepfakes in courts. We must test more datasets from digital evidence and conduct further experiments.

Table 6. F-1, ROC-AUC, and AP scores.

Architecture	F-1	ROC-AUC	AP
VGG-19	95	96	93
DenseNet OD	92	99	99
DenseNet AD	92	97	97
DenseNet GS	97	99	99
Custom CNN OD	91	98	98
Custom CNN AD	85	95	95

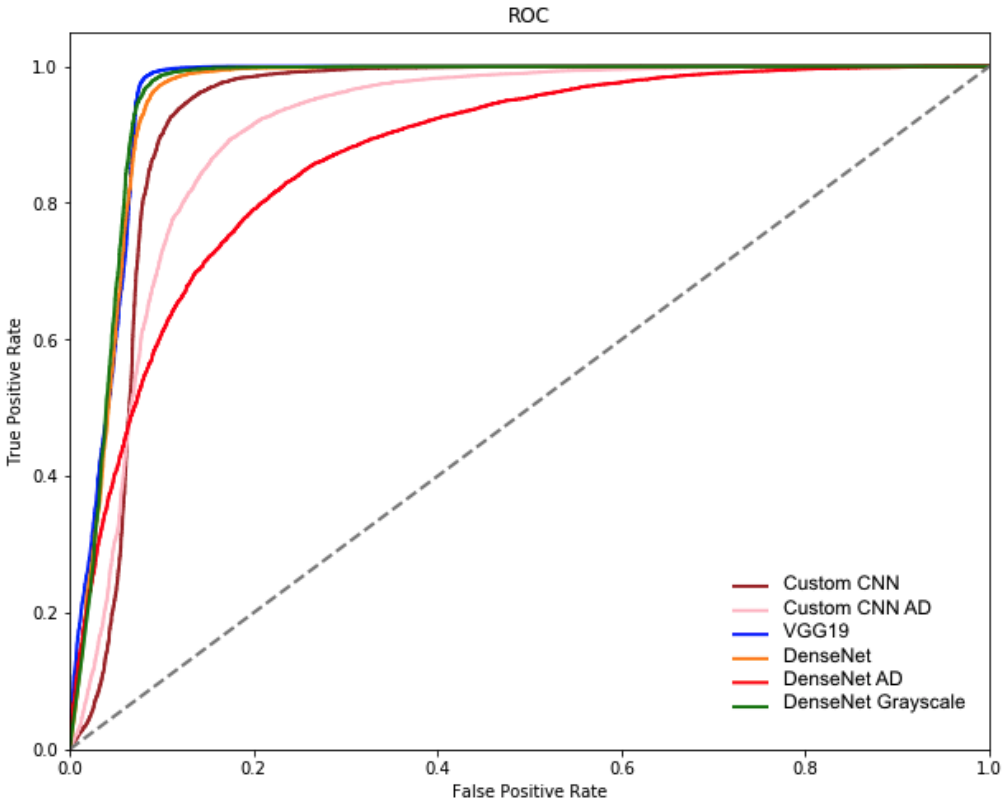

Figure 4. ROC curve representation.

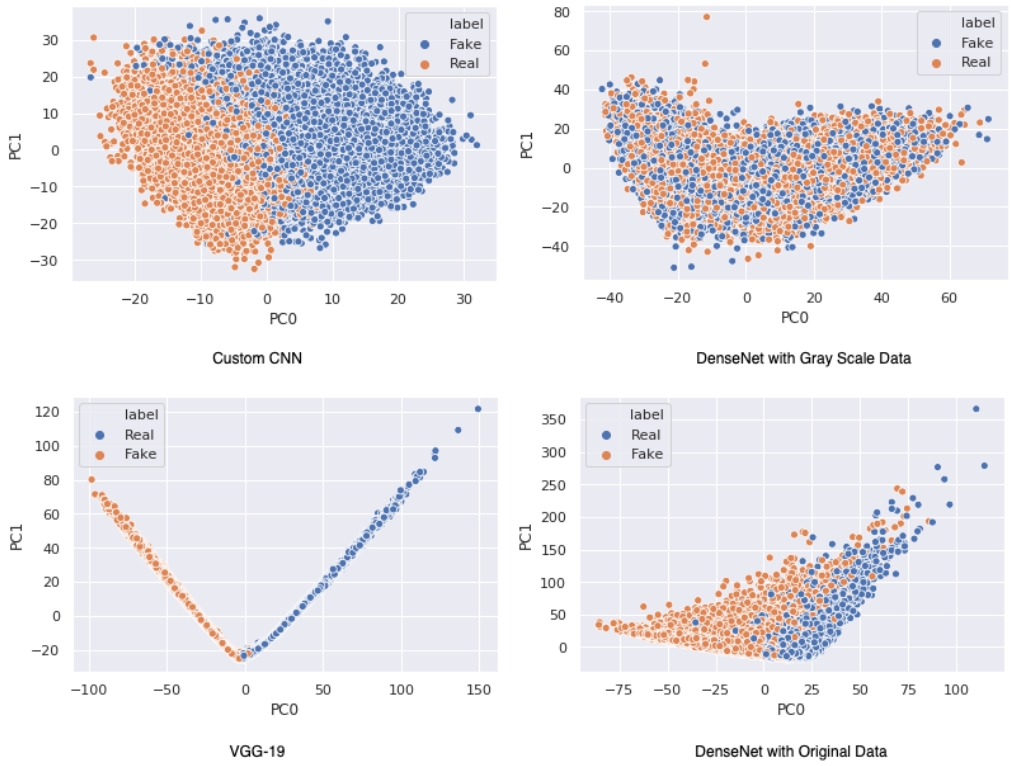

Figure 5. PCA-SVM clustering comparison.

5. Conclusions and Future Work

The results of our work demonstrated that deep-learning architectures are reliable and accurate at distinguishing fake vs. real images; however, detection of the minimal inaccuracies and misclassifications remain a critical area of research. Recent efforts have focused on improving the algorithms that create deepfakes by adding especially designed noise to digital photographs or videos that are not visible to human eyes and can fool the face-detection algorithms [61]. The results of our work indicate that VGG-19 performed best, taking accuracy, F1-score, precision, AUC-ROC and PCA-SVM measures into the account. DenseNet had a slightly better performance in terms of AP, and the results from the Custom CNN trained on original data were satisfactory too. This suggests that aggregation of the results from multiple models, i.e., ensemble or multi-attention approaches, can be more robust in distinguishing deepfake media.

Future work could also leverage unsupervised clustering methods such as auto-encoders to analyze its effectiveness on the task of deepfake classification and provide a better interpretation of the CNN algorithms designed in this work. There could be classification methods developed that would examine and flag social media users who uploaded images/videos before being posted on the Internet to avoid the spread of misinformation [62]. We plan to further improve performance with deep-learning algorithms as well as exploring the application of stenography, steganalysis and cryptography in the identification and classification of the genuine and disguised face images [63]. Future work not only has to include collecting and experimenting with different disguised classifiers, but also must work on the development of training data that can improve the performance of implemented architectures as suggested by [33]. The authors of the paper plan to discover the use of information pellets on the development of an ensemble framework. As suggested in [64] using a patch-based fuzzy rough set feature-selection strategy can preserve the discrimination ability of original patches. Such implementation can assist in anomaly detection for the task of deepfake detection. By integrating the local-to-global feature-learning method with multi-attention and ensemble-modeling (holistic, feature-based, noise-level, steganographic) approach, we believe we can achieve a superior performance than the cur-

rent state-of-the-art methods. Considering the limitations of Eff-YNet network developed by [55], which has an advantage in examining visual differences within individual frames, analyzing EfficientNet performance on deepfake image datasets used in this work can be another direction for future work, as it may identify another suitable baseline model for ensembled approaches.

Author Contributions: Conceptualization, M.T. and H.C.; methodology, M.T.; software, M.T.; validation, M.T. and H.C.; formal analysis, M.T.; developed the theory and performed the computations, M.T.; resources, M.T.; data curation, M.T.; writing—original draft preparation, M.T.; writing—review and editing, H.C.; funding acquisition, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly funded by the National Centers of Academic Excellence in Cybersecurity Grant (H98230-21-1-0326), which is part of the National Security Agency. Research was partly sponsored by the Army Research Office and was accomplished under Grant Number W911NF-21-1-0264. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets that have been leveraged in this work are publicly available on Kaggle for both challenges of "140K Real and Fake Faces" and "Real and Fake Face Detection". The "140K Real and Fake Faces" dataset available at https://www.kaggle.com/xhlulu/140k-real-and-fake-faces published on February 2020, accessed on October 2021, includes 70K real faces collected from Flickr and 70K fake faces that are generated by GANs. The "Real and Fake Face-Detection" dataset available at https://www.kaggle.com/ciplab/real-and-fake-face-detection published on January 2019, accessed on October 2021 includes 960 fake and 1081 real face images that are generated by human expert in high-quality via Photoshop.

Acknowledgments: The authors would like to show their gratitude to Shonda Bernadin and the MDPI journal reviewers. This paper and the research behind it would not have been possible without the exceptional support of them. Their insight and expertise and exacting attention to detail has greatly assisted this research

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ferreira, S.; Antunes, M.; Correia, M.E. Exposing Manipulated Photos and Videos in Digital Forensics Analysis. *J. Imaging* **2021**, *7*, 102. [CrossRef]
2. Harwell, D. Fake-Porn Videos are Being Weaponized to Harass and Humiliate Women: 'Everybody is a Potential Target'. 2018. Available online: https://www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944/ (accessed on 28 November 2021).
3. Tucker, P. The Newest AI-Enabled Weapon: 'Deep-Faking' Photos of the Earth. 2021. Available online: https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/ (accessed on 28 November 2021).
4. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *2*. Available online: https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html (accessed on 28 November 2021).
5. Sajjadi, M.S.; Scholkopf, B.; Hirsch, M. Enhancenet: Single image super-resolution through automated texture synthesis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4491–4500.
6. Yu, P.; Xia, Z.; Fei, J.; Lu, Y. A Survey on Deepfake Video Detection. *IET Biom.* **2021**, *10*, 607–624. [CrossRef]
7. Ferreira, S.; Antunes, M.; Correia, M.E. A Dataset of Photos and Videos for Digital Forensics Analysis Using Machine Learning Processing. *Data* **2021** , *6*, 87. [CrossRef]
8. Durall, R.; Keuper, M.; Pfreundt, F.J.; Keuper, J. Unmasking deepfakes with simple features. *arXiv* **2019**, arXiv:1911.00686 .
9. De keersmaecker, J.; Roets, A. 'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence* **2017**, *65*, 107–110. [CrossRef]

10. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2185–2194.
11. Bonettini, N.; Cannas, E.D.; Mandelli, S.; Bondi, L.; Bestagini, P.; Tubaro, S. Video face manipulation detection through ensemble of cnns. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 5012–5019.
12. Silva, S.H.; Bethany, M.; Votto, A.M.; Scarff, I.H.; Beebe, N.; Najafirad, P. Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic. Sci. Int. Synerg.* **2022**, *4*, 100217. [CrossRef]
13. Westerlund, M. The emergence of deepfake technology: A review. *Technol. Innov. Manag. Rev.* **2019**, *9*, 40–45. [CrossRef]
14. Ismail, A.; Elpeltagy, M.; Zaki, M.; ElDahshan, K.A. Deepfake video detection: YOLO-Face convolution recurrent approach. *Peerj Comput. Sci.* **2021**, *7*, e730. [CrossRef]
15. Coccomini, D.; Messina, N.; Gennaro, C.; Falchi, F. Combining efficientnet and vision transformers for video deepfake detection. *arXiv* **2021**, arXiv:2107.02612.
16. Kumar, A.; Bhavsar, A.; Verma, R. Detecting deepfakes with metric learning. In Proceedings of the 2020 8th International Workshop on Biometrics and Forensics (IWBF), Porto, Portugal, 29–30 April 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
17. Li, Y.; Lyu, S. Exposing deepfake videos by detecting face warping artifacts. *arXiv* **2018**, arXiv:1811.00656.
18. Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; Guo, B. Face X-ray for more general face forgery detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020. [CrossRef]
19. Nguyen, H.H.; Yamagishi, J.; Echizen, I. Capsule-forensics: Using capsule networks to detect forged images and videos. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 2307–2311.
20. Albanesius, C. Deepfake Videos Are Here, and We're Not Ready. 2019. Available online: https://www.pcmag.com/news/deepfake-videos-are-here-and-were-not-ready (accessed on 5 December 2021).
21. Yang, X.; Li, Y.; Lyu, S. Exposing deep fakes using inconsistent head poses. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 8261–8265.
22. Korshunov, P.; Marcel, S. Deepfakes: A new threat to face recognition? assessment and detection. *arXiv* **2018**, arXiv:1812.08685.
23. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
24. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Ferrer, C.C. The DeepFake Detection Challenge Dataset. *arXiv* **2020**, arXiv:2006.07397.
25. Li, Y.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In Proceedings of the IEEE Conference on Computer Vision and Patten Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 3207–3216.
26. Jiang, L.; Li, R.; Wu, W.; Qian, C.; Loy, C.C. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 2889–2898.
27. Yonsei University. Real and Fake Face Detection. 2019. Available online: https://archive.org/details/real-and-fake-face-detection (accessed on 30 August 2021).
28. NVlabs. NVlabs/ffhq-Dataset: Flickr-Faces-HQ Dataset (FFHQ). 2019. Available online: https://archive.org/details/ffhq-dataset (accessed on 29 August 2021).
29. Nguyen, T.T.; Nguyen, C.M.; Nguyen, D.T.; Nguyen, D.T.; Nahavandi, S. Deep learning for deepfakes creation and detection: A survey. *arXiv* **2019**, arXiv:1909.11573.
30. Huang, Y.; Juefei-Xu, F.; Guo, Q.; Xie, X.; Ma, L.; Miao, W.; Liu, Y.; Pu, G. FakeRetouch: Evading deepfakes detection via the guidance of deliberate noise. *arXiv* **2020**, arXiv:2009.09213.
31. Zhao, W.; Chellappa, R.; Phillips, P.J.; Rosenfeld, A. Face recognition: A literature survey. *Acm Comput. Surv. (CSUR)* **2003**, *35*, 399–458. [CrossRef]
32. Maksutov, A.A.; Morozov, V.O.; Lavrenov, A.A.; Smirnov, A.S. Methods of deepfake detection based on machine learning. In Proceedings of the 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), St. Petersburg, Russia, 27–30 January 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 408–411.
33. Tariq, S.; Lee, S.; Kim, H.; Shin, Y.; Woo, S.S. Gan is a friend or foe? a framework to detect various fake face images. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, Limassol, Cyprus, 8–12 April 2019; pp. 1296–1303.
34. Cozzolino, D.; Thies, J.; Rössler, A.; Riess, C.; Nießner, M.; Verdoliva, L. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv* **2018** arXiv:1812.02510.
35. Wang, S.Y.; Wang, O.; Zhang, R.; Owens, A.; Efros, A.A. CNN-generated images are surprisingly easy to spot . . . for now. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8695–8704.
36. Li, Y.; Chang, M.; Lyu, S. Exposing AI Created Fake Videos by Detecting Eye Blinking. In Proceedings of the 2018 IEEE InterG National Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018.

37. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. Mesonet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS 2018), Hong Kong, China, 11–13 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–7.
38. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
39. Chang, X.; Wu, J.; Yang, T.; Feng, G. Deepfake face image detection based on improved VGG convolutional neural network. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–30 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 7252–7256.
40. Kim, J.; Han, S.; Woo, S.S. Classifying Genuine Face images from Disguised Face Images. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angelas, CA, USA, 9–12 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 6248–6250.
41. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 4700–4708.
42. Hsu, C.C.; Zhuang, Y.X.; Lee, C.Y. Deep fake image detection based on pairwise learning. *Appl. Sci.* **2020**, *10*, 370. [CrossRef]
43. Matern, F.; Riess, C.; Stamminger, M. Exploiting visual artifacts to expose deepfakes and face manipulations. In Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 1–7 January 2019. [CrossRef]
44. Jung, T.; Kim, S.; Kim, K. DeepVision: Deepfakes detection using human eye blinking pattern. *IEEE Access* **2020**, *8*, 83144–83154. [CrossRef]
45. Li, Y.; Chang, M.C.; Lyu, S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS 2018), Hong Kong, China, 11–13 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–7.
46. Korshunov, P.; Marcel, S. Speaker inconsistency detection in tampered video. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 2375–2379.
47. Agarwal, S.; Farid, H.; Gu, Y.; He, M.; Nagano, K.; Li, H. Protecting World Leaders Against Deep Fakes. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–20 June 2019; Volume 1.
48. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
49. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
50. Li, X.; Yu, K.; Ji, S.; Wang, Y.; Wu, C.; Xue, H. Fighting against deepfake: Patch&pair convolutional neural networks (PPCNN). In Proceedings of the Companion Proceedings of the Web Conference, Taipei, Taiwan, 20–24 April 2020; pp. 88–89.
51. Pokroy, A.A.; Egorov, A.D. EfficientNets for deepfake detection: Comparison of pretrained models. In Proceedings of the 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), St. Petersburg, Russia, 26–29 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 598–600.
52. Mitra, A.; Mohanty, S.P.; Corcoran, P.; Kougianos, E. A novel machine learning based method for deepfake video detection in social media. In Proceedings of the 2020 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS), Chennai, India, 14–16 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 91–96.
53. Dolecki, M.; Karczmarek, P.; Kiersztyn, A.; Pedrycz, W. Utility functions as aggregation functions in face recognition. In Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, Greece, 6–9 December 2016. [CrossRef]
54. Du, C.X.T.; Duong, L.H.; Trung, H.T.; Tam, P.M.; Hung, N.Q.V.; Jo, J.; Efficient-frequency: A hybrid visual forensic framework for facial forgery detection. In Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (IEEE SSCI), Canberra, Australia, 1–4 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 707–712.
55. Tjon, E.; Moh, M.; Moh, T.S. Eff-YNet: A Dual Task Network for DeepFake Detection and Segmentation. In Proceedings of the 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM), Seoul, Korea, 4–6 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8.
56. Do, N.T.; Na, I.S.; Kim, S.H. Forensics face detection from GANs using convolutional neural network. In Proceedings of the 2018 International Symposium on Information Technology Convergence (ISITC 2018), Jeonju, Korea, 24–27 October 2018.
57. Goel, R.; Mehmood, I.; Ugail, H. A Study of Deep Learning-Based Face Recognition Models for Sibling Identification. *Sensors* **2021**, *21*, 5068. [CrossRef]
58. Varoquaux, G.; Buitinck, L.; Louppe, G.; Grisel, O.; Pedregosa, F.; Mueller, A. Scikit-learn: Machine learning without learning the machinery. *Getmobile: Mob. Comput. Commun.* **2015**, *19*, 29–33. [CrossRef]
59. Son, S.B.; Park, S.H.; Lee, Y.K. A Measurement Study on Gray Channel-based Deepfake Detection. In Proceedings of the 2021 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, 20–22 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 428–430.
60. Amerini, I.; Galteri, L.; Caldelli, R.; Del Bimbo, A. Deepfake video detection through optical flow based cnn. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Montreal, BC, Canada, 11–17 October 2021; p. 2.

61. Li, Y.; Yang, X.; Wu, B.; Lyu, S. Hiding faces in plain sight: Disrupting ai face synthesis with adversarial perturbations. *arXiv* **2019**, arXiv:1906.09288.
62. Tolosana, R.; Romero-Tapiador, S.; Fierrez, J.; Vera-Rodriguez, R. Deepfakes evolution: Analysis of facial regions and fake detection performance. In Proceedings of the International Conference on Pattern Recognition (ICPR), Virtual Event, 10–15 January 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 442–456.
63. Corcoran, K.; Ressler, J.; Zhu, Y. Countermeasure against Deepfake Using Steganography and Facial Detection. *J. Comput. Commun.* **2021**, *9*, 120–131. [CrossRef]
64. Guo, Y.; Jiao, L.; Wang, S.; Wang, S.; Liu, F. Fuzzy sparse autoencoder framework for single image per person face recognition. *IEEE Trans. Cybern.* **2017**, *48*, 2402–2415. [CrossRef]