ISSN: 0067-2904

A Survey on Detecting Deep Fakes Using Advanced AI-Based Approaches

Mohamed Abdulrahman Abdulhamed[*1,2] , Asaad Noori Hashim [2]
[1] *Department of Computer Science, Computer Science and Information Technology college, University of Basra, Basrah, Iraq*
[2] *Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq*

Received: 3/3/2023 Accepted: 13/8/2023 Published: 30/9/2024

Abstract

Today, artificial intelligence is used to clone human faces, which leads to a new technology known as "deepfakes." Recently, machine learning (ML) approaches and the use of deep learning (DL) networks have captured researchers' competition to achieve the highest classification accuracy in building efficient models for digital content deepfake detection. Therefore, this review analyzes and compares existing deepfake detection methods based on advanced artificial intelligence algorithms. Thus, deepfake detection techniques were classified into three major categories based on the classifier model used (machine learning, deep learning, or hybrid) and then compared to show the aspects that influence the efficiency and accuracy of the algorithms. This research helps researchers develop efficient classification models for deepfake detection applications. Based on the survey information reviewed in this study, a discussion of open issues and future directions is presented. The most important challenges and research directions related to deepfake detection methods are discussed.

Keywords: Deep Learning, Deepfake Detection, Media Forensic, Deepfake Classification

دراسة استقصائية لطرق اكتشاف التزييف العميق القائمة على تعبيرات الذكاء الاصطناعي المتقدم

محمد عبدالرحمن عبدالحميد [*1,2], اسعد نوري هاشم[2].
[1]قسم علوم الحاسوب, كلية علوم الحاسوب وتكنولوجيا المعلومات, جامعة البصرة, البصرة, العراق
[2]قسم علوم الحاسوب, كلية علوم الحاسوب والرياضيات, جامعة الكوفة, النجف, العراق

الخلاصة

اليوم ، يتم استعمال الذكاء الاصطناعي لاستنساخ الوجوه البشرية ، مما يؤدي إلى ظهور تقنية جديدة تعرف باسم Deepfake. في الآونة الأخيرة ، استحوذت طرق التعلم الآلي (ML) واستعمال شبكات التعلم العميق (DL) على منافسة الباحثين لتحقيق أعلى دقة لإنشاء نماذج التصنيف الفعالة للكشف عن المحتوى الرقمي المزيف. لذلك ، تحلل وتقارن ورقة المراجعة هذه طرق الكشف عن التزييف العميق بالاعتماد على خوارزميات الذكاء الاصطناعي المتقدمة. بحيث ، تم تصنيف تقنيات الكشف عن Deepfake إلى ثلاث فئات رئيسية بناءً على نموذج المصنف المستعمل وهي (التعلم الآلي ، التعلم العميق ، أو الهجين) لمقارنتها ، ثم إظهار وتوضيح الجوانب التي تؤثر على كفاءة ودقة تلك الخوارزميات في نماذج التصنيف لهذه المشكلة.

*Email: mohammed@uobasrah.edu.iq

وبالتالي ، فإن الهدف من هذه الدراسة هو مساعدة الباحثين في اقتراح نماذج التصنيف الفعالة والمناسبة لتطبيقات اكتشاف Deepfake. وبناءً على معلومات المسح التي تمت مراجعتها في هذه الدراسة ، تمت مناقشة القضايا المفتوحة والاتجاهات المستقبلية للكشف عن المحتوى المزيف. بالإضافة إلى ذلك ، تمت مناقشة أهم التحديات واتجاهات البحث المستقبلية المتعلقة بالكشف عن Deepfake ومشاكل الطب الشرعي متعددة الوسائط.

1. Introduction

In recent years, video manipulation—and specifically facial manipulation—has drawn a lot of attention. This is especially true after the emergence of "deepfakes," which use deep learning tools to manipulate images and videos. Deep fake algorithms can use auto-encoders or generative adversarial networks to add faces from the source to the target video. With this technology, videos of manipulated faces can be easily created by drawing on large amounts of data for training. Several methods for detecting deepfake videos have been developed since an anonymous user posted on Reddit that turned the faces of a group of celebrities into pornographic videos in 2017. Using recurrence networks in return, some methods detect temporal discrepancies across frames of faces in videos, while others use convolutional networks to detect visual defects [1], [2]. Because of their high-quality videos and their accessibility to different users, deepfake techniques have gained growing popularity recently. Deep-Face-Lab, Fake-App, and Faceswap by GAN are some of the widespread face manipulation apps that are based on the generative adversarial network and auto-encoder-decoder architectures [3], [4]. GAN comprises two deep neural networks: a generator and a discriminator. As a result, synchronous training occurs during the learning process. A deep learning discrimination network is used to distinguish between the original and artificially manufactured image, which is generated by a deep learning neural network called a generator, where the latter relies on random samples and a training set to produce the fake contents [1]. In an auto-encoder architecture, the extractor extracts hidden features from face photos while the decoder reconstructs the photos. The encoder-decoder pairs need to be trained on distinct sets of faces to switch between target and source faces, with each pair of encoders and decoders having shared encoder weights. Because of a link between the first face's decoder and its features set, a reconstruction of the second face from the first original face can be achieved [5]. However, a huge dataset of real and fraudulent movies is required to train the model for deep fake detection, which is a binary classification issue that assesses the authenticity of videos [4]. So, there are several deep fake video datasets available. Table 1 shows the popular seven datasets for this problem [6], [7].

Table 1: The top seven datasets used in deep fakes

Reference	Dataset Name	Subjects	No. of real videos	No. of fake videos
[8]	FaceForensics++ (FF++)	977	1000	5000
[9]	DeepFake-TIMIT	64	960	640
[10]	WildDeepfake	100		707
[11]	Deepfake Detection Challenge (DFDC)	66	5244	5244
[12]	Celeb-DeepFake (Celeb-DF)	72	6229	5639
[13]	DeeperForensics-1.0	100	60000	1000
[14]	Google/Jigsaw DeepFake Detection		3000	3000

Despite significant progress, many key issues with current deepfake detection methods remain unresolved. The videos made are appearing more realistic thanks to deepfakes constantly improving techniques. In this situation, it is likely that conventional techniques would not work to identify videos that have been altered using new deepfake algorithms [15]. Analyzing and projecting the future of deepfake-related research is important, as is developing the appropriate detection methods. In this study, we will concentrate on the current deepfake video detection scheme in an effort to encourage the creation of deepfake video detection techniques.

In this review paper, we highlight the latest developments in deep fake detection techniques in order to identify the latest and most effective methods to obtain the highest efficiency and precision in their classification process. Additionally, it highlights the difficulties and constraints that machine learning-based deepfakes present that cannot be resolved by existing techniques. The research has been categorized into three groups according to the algorithms' usage: machine learning (ML), deep learning (DL), and hybrid methods.

As a result, this article is organized as follows: Section 2 presents an overview of deepfakes, a series of algorithms for creating deepfake videos that have been proposed in recent years. Section 3 is dedicated to deepfake classification techniques, followed by a discussion on the state of deepfake video detection and the issues that remain unresolved in Section 4. Finally, Section 5 concludes with conclusions and recommendations for future directions.
Following is a summary of the contributions to our survey:
•ML, DL, and hybrid method detection methods for deepfake detection are classified into three categories in this review, as are their limitations.
•Moreover, this survey provides a prognosis for the future as well as a discussion of all the difficulties that researchers in this field face.
•In this review, we suggest and present some proposals for improving the feature extraction stage that is used in this field.

2. Related Works

There have been many survey studies conducted over the last three years in order to gain a deeper understanding of how deepfakes work, and many approaches based on machine learning and deep learning have been developed to detect deep fake videos and images. This section reviews the existing literature on deepfaking. Firstly, the survey [16] examines the reliability of deepfake detection studies. Transferability, interpretability, and robustness define deepfake detection research’s reliability challenges. While solutions have been frequently addressed for the three challenges, the general reliability of a detection model has rarely been considered, resulting in a lack of reliable evidence in real-life usage and court prosecutions of deepfake cases. Thus, they use statistical random sampling and publicly available benchmark datasets to evaluate the reliability of existing detection models on arbitrary Deepfake candidate suspects. This survey's reliable detection models are used to justify real-life fake cases involving different victim groups. Secondly, this paper [1] discusses deep learning-based methods for creating and detecting fakes. Using cutting-edge methods, the study detected deepfake videos and images in social media content. The authors examine deepfake detection technologies. The study covers cutting-edge methods for detecting deepfake videos and images in social media content to benefit researchers. The detailed description of this domain's latest methods and datasets will help compare existing works. Third, [17] this survey studies deepfake detection models that use deep learning

algorithms to combat deep fakes. The survey covers deepfake detection models and research methods. The paper discusses these models' pros and cons and makes research recommendations. Finally, an augmented dataset of real and fake faces is used to evaluate state-of-the-art face detection classifiers like CustomCNN, VGG19, and DenseNet-121 [18]. Data augmentation boosts performance and saves computational resources. VGG19 outperforms the other models in the study with 95% accuracy. The work sheds light on how different face detection classifiers detect fake faces, which could help combat deepfake proliferation. Furthermore, there exist several sources pertinent to our study that have provided us with a comprehensive perspective within the research domain, such as [19]–[23]. Conversely, our study provides a comparative study of the latest methods that employ advanced artificial intelligence, including machine learning, deep learning, and hybrid methods, to provide researchers with deeper insight into deepfake detection methods.

3. Deepfake Overview

Today, deep learning and computer graphics have revolutionized image and video processing. Towards this end, autoencoders and generative adversarial networks (GAN) were implemented to produce desirable results [24], in particular for facial synthesis, where photorealism is high [25], [26]. In addition, segmentation maps are also useful for generating synthetic images and videos [27]. The manipulation of faces has received a lot of attention because of their high semantic value and the variety of applications that can be found. Recently, several methods have been proposed for changing facial expressions [28], [29], for transferring expressions from one actor to the other [30], and for swapping faces [31]. Recent findings have shown that even without numerous training photos of the person you are trying to manipulate, it is possible to manipulate their face effectively [32]. Recently, several studies have been done on the transfer of movement from a source dancer to a target person's face and how the expression of a variety of emotions can be achieved [33]. However, there has also been research on transferring motions [34, 35]. As shown in Figure 1, a general description of the steps involved in a deep fake detection architecture can be found. Thus, we created this diagram from the sources above to track deepfake detection model construction.

Figure 1: The steps of the deep fake detection architecture

3.1. Deepfake Generation

Visual content can be manipulated in many ways, and new ways are proposed daily. We will briefly examine a few of the most popular and promising in this section. Splicing can

insert objects from different images, or copy-moving can insert objects from the same image. By extending the background to cover existing objects (inpainting), exemplar-based inpainting [36] can be used to delete existing objects. The task can be accomplished easily with widespread photo editing software. Besides improving visual appearance and guaranteeing coherent perspective and scale, some post-processing can also be applied, such as resizing, rotation, and color adjustment. The term "cheap fakes" is sometimes used to refer to data manipulation without the use of sophisticated artificial intelligence (AI) tools. So these tools will distort reality as much as possible. One could, for example, delete, insert, or clone groups of frames from a video to completely alter its meaning [37]. In contrast, artificial intelligence-based fake generation methods produce results that are almost accurate, such as generative adversarial networks (GAN). With enough data to fuel the algorithm, this technique enables the creation of nearly lifelike multimedia [38]. There are several applications of this technology, including photography, video games, and virtual reality, as well as movie production in the near future. Technology can be used for malicious purposes, such as extortion or spreading fake news, which can lead to privacy violations and a decrease in trust in journalism. In the long run, belief in journalism might decline, including that of reliable and reputable sources [39]. GANs are composed of two neural network components: an encoder and a decoder. To create fake data, the model first trains on a large data set using an encoder [17]. After that, a decoder is used to tell the difference between real and fake data. In order to make realistic faces, this proposed model needs a lot of input data, such as images or videos. Typically, Figure 2 shows the Generative Adversarial Networks (GNA) architecture.

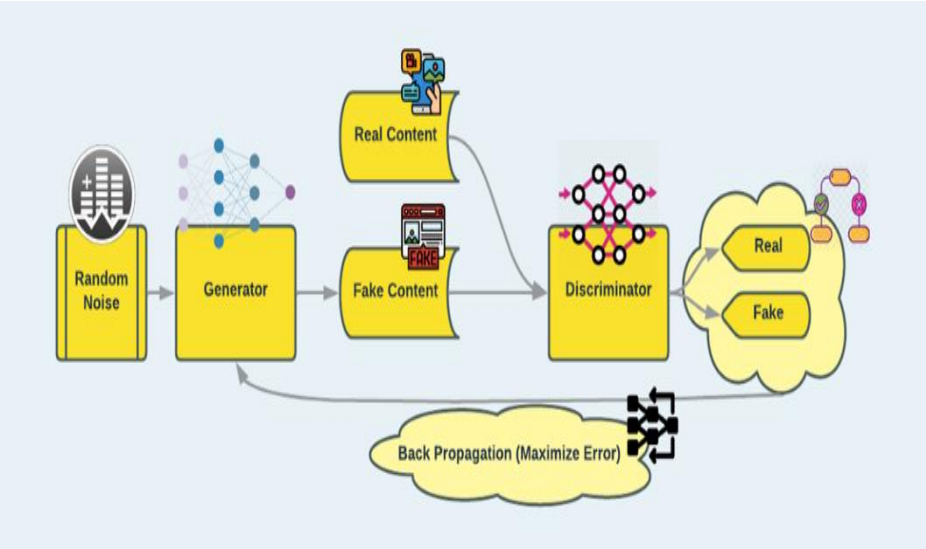

Figure 2: Describe the general architecture of GAN

A decoder is a binary classifier that compares real samples to fake samples and applies a SoftMax function to distinguish between them [17]. A substantial amount of data is required to train the deepfake model on these deepfake media. Data-trained models create fake images and videos. Social media's abundance of presidential and Hollywood celebrity videos can inspire rumors, which may harm society. [17], [40].

It is imperative to point out several significant applications that facilitate the creation of deep fake content [38], such as DeepFaceLab, DFaker, and Deep-Fake tf, which is based on the TensorFlow platform.

3.2. Deepfake Detection

The survey elucidates the prevalent techniques employed for the identification of deepfakes, encompassing machine learning (ML) and deep learning (DL)-based approaches as well as hybrid methodologies. A succinct summary of the research papers examined in this study has been presented, organized into three subsections as previously mentioned.

3.2.1. DEEPFAKE DETECTION TECHNIQUES BASED ON MACHINE LEARNING

This method generates a feature vector that is fed into a classifier to determine whether the videos or images have been altered by DeepFake using various cutting-edge feature selection algorithms. A variety of machine learning-based models are used, including Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), Random Forest (RF), Multiple Instance Learning (MIL), and Discriminant Analysis (DA). The summary of some articles in this category is as follows: In [41], the authors present a novel method that has been overlooked in previous studies despite its widespread use in real-life scenarios. Multiple-instance learning solves this problem by treating each video that is fed in as both a bag and an instance. Therefore, the S-MIL, which directly connects bag label prediction to instance embeddings, is necessary in order to alleviate gradient vanishment in traditional MIL. According to theoretical analysis, S-MIL alleviates gradient vanishment. Spatial-temporal encoding facilitates the accurate encoding of partially manipulated faces, while intraframe and interframe inconsistencies are fully modeled to promote detection performance. Based on the joint analysis of multiple temporal segments [32], the binary decision exploits both spatial and temporal textural dynamics, in contrast to previous approaches. A compact feature representation known to be extremely useful for detecting face spoofing attacks, Local Derivative Patterns on Three Orthogonal Planes (LDP-TOP), is used to achieve this. This paper presents a novel approach that utilizes an expectation-maximization algorithm for the purpose of detecting and extracting deepfake fingerprints from images. During image generation, GANs leave convolutional traces (CT) that are used to represent the fingerprint [42]. A new method to expose fake faces generated by deepfake is presented in the paper [3]. Deepfakes are created by connecting parts of the original image that were composited with face regions, causing errors when estimating the 3D head position from the face image. To prove this phenomenon, it is tested, and then a classification method based on it is developed. SVM classifiers are evaluated with real and fake face images using features based on this hint. The proposed approach [34] employs various deep learning techniques and is grounded in metric learning. It has shown a high level of efficacy in detecting deep fakes in scenarios involving image compression. By using a triplet network architecture, the metric learning method is helpful because it needs fewer frames per video to judge how realistic it is. Using this algorithm, real and fake embedding vector clusters are enhanced by enhancing the feature space distance. A method is proposed in [43] to identify deepfake videos with minimal computational power using visual artifacts found in the generated deepfakes. using a three-layer neural network to classify videos as real or fake, and then confirming the results by calculating the variance of Laplacian in different patches on a face and comparing them to identify deepfakes. An additional summary of this research based on machine learning techniques can be found in Table 2.

3.2.2. DEEPFAKE DETECTION TECHNIQUES BASED ON DEEP LEARNING

Computer vision, machine vision, and natural language processing are just some of the many areas that have achieved advanced success by relying on deep learning. Feature extraction and selection mechanisms of deep learning models, which are capable of learning directly from data, have been widely used in computer vision [44], [45]. The following deep

learning-based models were used in deepfake detection studies, as shown in Figure 3. There are a number of recent sources that use deep learning techniques to tackle the problem of detecting deep fakes, which will be discussed in the next paragraph.

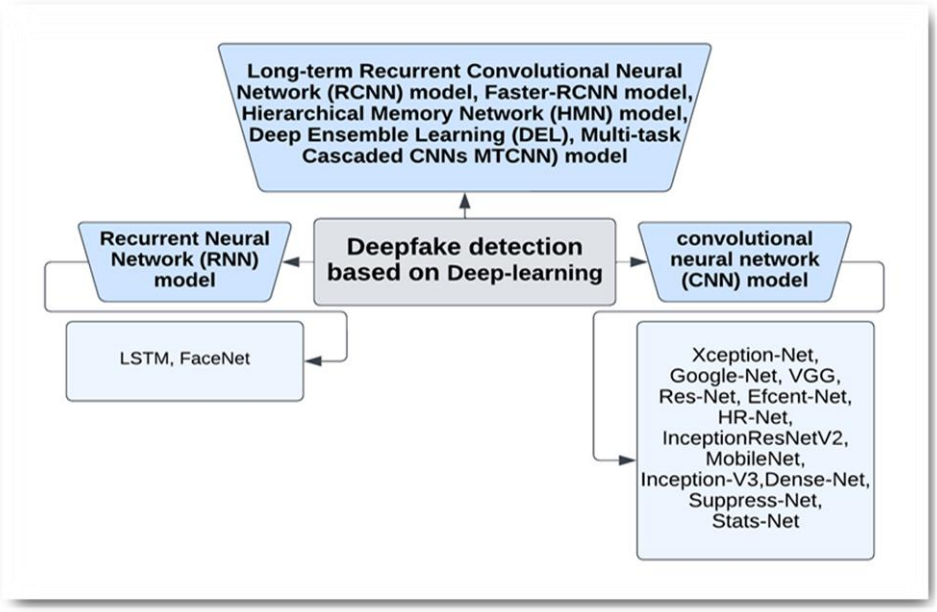

Figure 3: Deep learning-based deepfake

The authors in [46] proposed a DeepFake detection and classification model based on a five-layered convolutional neural network (CNN). After the face region has been extracted from video frames, the CNN, which has been improved with ReLU, is utilized to extract characteristics from those facial features. To ensure model accuracy while maintaining a suitable weight for the DeepFake-detection-influenced video, a CNN equipped with a ReLU model was used. A new method for detecting deepfakes is presented in [47]: YOLO-CNN-XGBoost. Taking advantage of the InceptionResNetV2 CNN, features from video frames are extracted by picking out the face areas from YOLO face detectors. XGBoost is utilized as a recognizer at the top level of CNNs to maximize these features. With YOLO face detection, facial features are extracted from video frames using the InceptionResNetV2 convolutional neural network. XGBoost is employed in the uppermost layer of convolutional neural networks (CNNs) as a classifier. As a method of detecting and quantifying deepfake videos, [48] uses ResNext, an algorithm based on convolutional neural networks (CNN) and long short-term memory (LSTM). A novel counterfeit feature extraction technique using deep learning and error level analysis (ELA) was proposed in [49] to improve distinguishing deepfake-generated images. The final layer includes a cross-entropy loss function. In the final SoftMax layer, there is a cross-entropy loss function, which is related to entropy and information theory. There are limitations to DeepFake's resolution generation. With this algorithm, the fake face area appears in the foreground and the original area in the background. Convolutional neural networks (CNNs) can be used to detect counterfeit features in images using the ELA method. As a result, [50] proposed a CNN network model that has higher performance and is lighter than others. DeepFake detection was improved by combining a manual method with an AI-based algorithm. They cleaned and processed the most important information, regions, and features using deep neural networks to achieve high accuracy. Using deep transfer learning, [51] proposed a method to detect face swapping, resulting in a 96% true positive rate and very few false alarms. Their approach differs from

existing methods that only provide detection accuracy; they also provide uncertainty for each prediction, which is critical for the credibility of such detection systems. Plus, a website was developed to collect pair-wise image comparisons from human subjects so that the performance of human recognition could be assessed. The [52] study proposes preserving and extracting the unique features of images based on cutting-edge deepfake generation techniques for this purpose. To do this, a new method of learning acting has been proposed, called pair-wise self-consistency learning (PCL), to extract these source features and detect deepfake images. To support PCL, the Inconsistency Image Generator (I2G) provides richly annotated training data. In addition to the foregoing, Table 3 provides a summary of the sources cited.

3.2.3. DEEPFAKE DETECTION TECHNIQUES BASED ON HYBRID METHODS

As a result, many researchers are developing hybrid models based on combining artificial intelligence methods to detect deep fakes. Below, we summarize some of the approaches that have been reviewed in this regard. They extracted temporal features from the data using an optical flow-based feature extraction technique, which they then fed into a hybrid classification model in the paper [53]. In Table 2, we show the results of this hybrid model, which is a combination of CNNs and recurrent neural networks (RNNs). [54] proposes a hybrid transformer network to detect deepfake videos using early feature fusion. They used CNNs (XceptionNet and EfficientNet-B4) to extract features. This study presents a hybrid transformer network based on early fusion for detecting deepfake media. FaceForensics++ and DFDC benchmarks are employed to train the transformer end-to-end and the feature extractors. In addition to this, they also proposed novel techniques for augmentation of face cut-outs as well as random techniques for augmentation of face cut-outs. In [55], the authors suggest a way to find deepfakes called HCiT. It combines convolutional neural networks (CNN) and vision transformers (ViT). With the HCiT hybrid architecture and the self-attention feature of the ViT for features extracted process, this structure extracts local information and boosts detection accuracy by utilizing the benefits of CNN. Using the approach proposed in [47] as well as the simultaneous use of transfer learning in autoencoders, the objective is to develop a new framework that detects fake videos using a hybrid model that incorporates convolutional neural networks (CNN) and recurrent neural networks (RNN). To train the deep fake detection model, DFDC and FF++ datasets are used, along with various pre-trained architectures such as VGG16 [56], Inception ResNetV5 [57], Efficient Nets, and Efficient Nets with LSTMs, and classification metrics, such as accuracy and AUC, are evaluated. The authors [58] suggest an end-to-end visual forensic framework using various modes to classify genuine and fake content effectively. where the model makes use of both original content and frequency domain analysis to fully exploit the richness of image latent patterns. Pattern extraction is carried out by two separate EfficientNets, a neural network architecture designed for image classification that is light and efficient. Afterward, they design a late-fusion mechanism based on the importance of the underlying information to fuse the learned features in the original and frequency domains. The authors proposed a high-confidence manipulation localization architecture in [59] by utilizing resampling features, LSTMs, and encoder-decoder networks to distinguish between altered and unmanipulated regions. Artifacts such as downsampling, upsampling, rotation, and shearing are captured with resampling features. By combining an encoder and an LSTM network, their proposed network uses larger receptive fields to study the differences between manipulated and unmanipulated areas (spatial maps). To locate image tampering, the decoder network maps low-resolution features into pixel-wise predictions.

4. Discussion And The Open Issues:

A discussion of the findings presented in this paper will be discussed in this section, along with how further research can be proposed to improve AI-based deep fake detection. Additionally, deepfake classification and detection models are compared in terms of their performance in feature extraction methods in order to increase classifier accuracy and efficiency.

4.1. Comparative Study

Here, the performance results of various deepfake detection models are compared through feature extraction methods, which can help achieve high classification accuracy with greater efficiency. However, some researchers propose new neural networks or improve existing neural networks to develop a classifier. There is another category of researchers who used adaptive activation functions rather than common hidden layer functions in their research to enhance accuracy and shorten training time. In contrast, other researchers use weight initialization methods in which these values influence neural network convergence [60]. Alternatively, some researchers use an optimization algorithm to find the optimal parameters for their machine learning algorithms. This is done to enhance the deepfake classifiers' performance and accuracy. Several factors can improve the overall performance of a deep learning-based classification system. In neural networks, these factors help speed convergence by combining adaptive activation functions, optimizations, or initialization methods. Consequently, the time required for training will be significantly reduced. Numerous optimization techniques are used, such as Momentum, Adam, RMS prop, and mini-batch gradient descent, to hasten the convergence of the neural network. Therefore, a variety of deep learning approaches have been proposed for detecting deep fakes in images and videos, including long short-term memory (LSTM), recurrent neural networks (RNN), and even hybrid approaches. These approaches are explained in Table 2. The two main types of fake video detection methods can be classified according to whether they use visual artifacts within a video frame or temporal features across frames. Consequently, the implementation of machine learning is facilitated by a streamlined process consisting of three essential steps: pretreatment, feature extraction, and feature selection. In contrast, it has been observed that deep learning techniques exhibit superior accuracy in detecting deepfakes compared to other machine learning algorithms while also necessitating minimal preprocessing. As a result, a deep learning algorithm was used to overcome the computational complexity of these steps. Table 2 illustrates how some researchers combine machine learning and deep learning to produce a new and efficient hybrid classification model with promising results.

Table 2: Summary of the reviewed papers for Deepfake Detection Based on Advanced AI-Based Approaches

Ref.	Deepfake Dataset	Feature extraction methods	Classification Methods	Classification Accuracy
Machine learning techniques				
[31]	FFPMS and FaceForensics++(FF++) dataset	Multi-Instance Learning (MIL) model and 0.928575 for the Sharp Multi-Instance Learning (S-MIL) model.	multiple spatial-temporal encoded bags with different kernel sizes to represent a video. The video's final fake score is obtained by S-MIL processing this super bag.	The overall result is 0.884825 for the MIL model and 0.928575 for the S-MIL model.

[32]	FaceForensics++ dataset	Temporal information is captured by LDP	Linear Support Vector Machines (SVMs)	On the basis of testing on Cross-Dataset, the classification averages of AUCs computed for two versions of algorithms (F,B) for the single-manipulation scenario are 76,68% and 78,34%.
[33]	A real-case scenario using Deepfakes generated by FACEAPP	The Expectation-Maximization algorithm is trained to identify and extract a fingerprint from (GAN).	K-NN, Linear SVM, Linear Discriminant Analysis (LDA)	93%
[2]	UADFV and subset from the DARPA MediFor GAN Image/Video Challenge	68 facial landmarks.	SVM	ROC curves obtained from UADFV deepfake and DARPA GAN datasets were 0.890 and 0.843, respectively,
[34]	Celeb-DF	Use a Metric learning to find the distance between the feature spaces of the real and fake video embedding vector clusters.	Sequence Classification based on RNN; 3D – CNN model	AUC score of 99.2%. accuracy of 90.71%
[35]	UADF and latest DeepFakeDetection dataset.	A three-layer neural network calculates Laplacian variance for different face patches to confirm results.	three-layer neural network	The approach suggested by the authors produces superior results in terms of both computational efficiency and accuracy (exceeded 90%).
Deep learning techniques				
[37]	DeepFake and Face2Face datasets	five-layered convolutional neural networks (CNNs)	a network-in-network (NIN) with ReLU is used to classify task	98% for DeepFake, while 95% for Face2Face.
[38]	the CelebDF-FaceForencics++(c23) merged dataset	InceptionResNetV2 CNN	XGBoost that works as a recognizer	92.62 % accuracy
[39]	Celeb-Deep fake dataset	Resnext and LSTM	ResNext	91% accuracy
[40]	MUCT dataset	deep learning and error level analysis (ELA)	SoftMax layer	The AUC testing for this algorithm is 97.6%
[41]	DeepFake Detection Dataset (DFDC) and Celeb-DF v2	manual distillation extraction, target-specific regions extraction and multi-region ensemble feed to CNN-based model	Multi-Region Ensemble based on MTCNN for classification stage	0.978 of AUC for DFDC, while 0.978 of AUC for Celeb-DF v2
[42]	aggregated celebrity database and Chicago Face Dataset (CFD)	deep transfer learning	class probability produced by the SoftMax function	dataset performs better than 90.0%.

[43]	cross-dataset evaluations for seven popular datasets (FF++, CD2, DFDC, DF,F2F,FS and NT).	pair-wise self-consistency learning (PCL), for training ConvNets	Based on ConvNets CNN for classifier	improve averaged AUC by 96.45% to 98.05% and by 86.03% to 92.18%
Hybrid techniques				
[44]	DFDC, FF++, and Celeb-DF	optical flow to extract temporal features	Combination of CNN and (RNN) architectures.	Accuracy of 66.26 %, 91.21% and 79.49%, respectively
[45]	DFDC dataset	XceptionNet and EfficientNet-B4	Fully connected layers	98.24%
[46]	Faceforensics++ and DeepFake Detection Challenge datasets	Convolutional Neural Network (CNN)	Vision Transformer (ViT) self-attention mechanism	97.7% for DFDC, while 99.0% for FF++ dataset
[47]	DFDC and Face Forensics++	Transfer learning in autoencoders and a hybrid model of CNN and RNN.	EfficientNet and LSTM, ResNetV2, and VGG16 Inception for final inference.	Achieves an AUC score of 94% and 98%, respectively
[48]	DF-in-the-wild and DFDC	two separated EfficientNet	A late-fusion mechanism to fuse the learnt features	Achieved around 0.8 accuracy on two challenging datasets
[49]	NIST'16, IEEE Forensics, and COVERAGE datasets	Resampling features are used to capture artifacts	SoftMax layer	AUCs of 0.7936, 0.7577, and 0.7124, respectively

4.2. Discussion

As part of our study, we present a comprehensive study of deepfake detection methods based on the most recent research. Based on advanced artificial intelligence, the techniques used can be divided into three categories: machine learning, deep learning, and a combination of both, as described in Section 2.2. The survey findings reveal that the identification of deepfake content poses certain challenges and limitations. Additionally, the study delves into various distinctive areas of inquiry that could be recommended for future research. In instances where AI-based techniques can be rendered more palatable, they exhibit a high degree of precision, albeit with a dearth of generalizability.

Firstly, in the total methods based on ML and DL, a model's efficiency is greatly affected by the choice of features and classifiers. The selection of features and classes was not a high priority in previous studies. In the training phase of ML-based approaches, researchers should identify the classifier that is most suitable for eliciting reliable artifacts based on specific traits. In order to determine the mechanisms for extracting features appropriate for machine learning-based feature extraction models, the extracted features will need to be analyzed accurately based on the pseudo-sections of fake contents, since complete features are required for these models. Given that the majority of prior research on deepfake detection models only examined data from specific databases (CelebDF, FaceForencics++, etc.), the testing procedures for the proposed models should focus on fake data that is not included in the datasets [51, 61].

To build powerful deepfake detection learning models, we recommend using real data (videos and images) or cross-database mechanisms. There is also potential for enhancing detection with fake content detection models that learn from newly emerging content. We can also train an advanced AI-based system with online multimedia content to detect deep fakes by using a transfer-learning approach. In order to improve the classification accuracy of deep fake content, more training data should be released. One of the most important problems with deepfake detection is the lack of sufficient data to feed artificial intelligence techniques.

On the other hand, the proposed hybrid models are not easy. There are several considerations that may be costly at the expense of computing speed and efficiency, so it can be costly in these aspects. In Section 3.2.3, for hybrid studies, which are the group of methods that combined machine learning and deep learning algorithms, good results were obtained in feature extraction processes but did not outperform the single methods. The results of sampling testing in the research paper [62] showed the highest percentage in the data set (FF++), where the accuracy rate was approximately 99.20%, but it was done by using pre-trained based on transfer-learning techniques of CNN types with a set of machine learning algorithms to get the highest results when using the EfficientNet type. Using hybrid approaches, it is possible to solve the generalization problem in methods for detecting deepfakes from an analytical perspective, particularly if methods are selected precisely and used carefully. For the extraction of features, deep learning methods such as CNN, RNN, and LSTM are used, whereas machine learning algorithms such as SVM, LR, and XGBoost are used for the classification process to design efficient models. Therefore, it makes the proposed model strong against many attacks relating to deep fake content creation.

Subsequent studies could potentially concentrate on advanced artificial intelligence models. According to the present survey, there is a lack of machine learning-based research that has verified newly generated content; thus, the outcomes do not provide insight into the efficacy of the models, especially in generalization. Alternatively, enhancing the hyperparameters of the network, employing a less complex design model, and incorporating additional layers were frequently employed techniques for enhancing deep learning algorithms.

4.3. Challenges and limitations

The paper reviews machine and deep learning methods for detecting deepfakes. This paper discusses current methods' limitations and how data are available. Deepfakes cannot be detected accurately and automatically, according to the literature. First off, the results of this survey show that no reliable models have yet been developed to identify deepfakes; however, deepfake videos are simple to produce and exhibit amazing results that are very realistic. Despite this, deep learning-based methods perform better than other methods at addressing this phenomenon. Deepfake detection can also be relied upon to detect visual artifacts and inconsistencies within video frames, or it can be relied upon to detect discrepancies between frames, taking into account different temporal correlations. Secondly, the lack of high-quality datasets poses a major challenge to researchers. There is also a scaling issue with the current deep learning methods. In order to produce good results from deep learning models, large datasets are required for training, which are not freely accessible or require social media providers' permission. As a result of all these challenges, deep learning models for detecting fakes are in high demand.

5. Conclusion:

The importance of developing methods for detecting deepfake photos and videos has recently escalated significantly. The availability of many photos and videos on social media has led to the rise of “deepfakes.” In recent years, there has been a discernible increase in the accessibility of tools and algorithms for creating deep fakes. This has resulted in the infringement of individuals' privacy through the dissemination of fabricated content on social media platforms, among others. The primary objective of this research paper is to obtain comprehensive knowledge regarding the classification of deepfakes utilizing advanced artificial intelligence techniques. To achieve this goal, a comparative analysis was conducted

on the most recent research publications aimed at addressing the issue of deepfakes. A variety of key topics were discussed at the present time, such as data sets, feature extraction methods, and classification methods using machine learning and deep learning. In terms of feature extraction, hybrid methods provide superior results. At the same time, it was found in this study that deep learning algorithms are more suitable for speed-sensitive applications, as they can train faster and require less computational complexity. The following are some future perspectives in this promising field that can be used for further research studies on deep fake detection techniques using AI-based algorithms:

- Future research should focus on more resilient, scalable, and broadly applicable detection methods.
- Proposing standard preprocessing methods to be used in training neural network models that are based on AI.
- To ensure that neural networks converge as quickly as possible, weight initialization must be based on an entirely new mathematical model.
- For feature extraction methods, advanced AI-based models may detect more generalizable methods. where hybrid adaptation methods can be proposed to achieve this.

References

[1] M. R. Oraibi and A. M. Radhi, "Enhancement Digital Forensic Approach for Inter-Frame Video Forgery Detection Using a Deep Learning Technique," *Iraqi Journal of Science*, vol. 63, no. 6, pp. 2686–2701, Jun. 2022.

[2] H. K. Dhahir and N. H. Salman, "A Review on Face Detection Based on Convolution Neural Network Techniques," *Iraqi Journal of Science*, vol. 63, no. 4, pp. 1823–1835, Apr. 2022.

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

[4] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, Nov. pp. 1–6, 2018.

[5] Z. Wang, Q. She, and T. E. Ward, "Generative adversarial networks in computer vision: A survey and taxonomy," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[6] I. Castillo Camacho and K. Wang, "A comprehensive review of deep-learning-based methods for image forensics," *J Imaging*, vol. 7, no. 4, p. 69, 2021.

[7] C. Li *et al.*, "A continual deepfake detection benchmark: Dataset, methods, and essentials," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1339–1349, 2023.

[8] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, pp. 1–11, Oct. 2019.

[9] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," *arXiv preprint arXiv:1812.08685*, 2018.

[10] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *Proceedings of the 28th ACM international conference on multimedia*, pp. 2382–2390, 2020.

[11] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (dfdc) preview dataset," *arXiv preprint arXiv:1910.08854*, 2019.

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

[13] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2889–2898, 2020.

[14] L. Guarnera *et al.*, "The Face Deepfake Detection Challenge," *J Imaging*, vol. 8, no. 10, p. 263, Sep. 2022.

[15] R. Chesney and D. Citron, "Deepfakes and the new disinformation war: The coming age of post-truth geopolitics," *Foreign Aff.*, vol. 98, p. 147, 2019.

[16] T. Wang, K. P. Chow, X. Chang, and Y. Wang, "Deepfake Detection: A Comprehensive Study from the Reliability Perspective," *arXiv preprint arXiv:2211.10881*, 2022.

[17] A. M. Almars, "Deepfakes detection techniques using deep learning: a survey," *Journal of Computer and Communications*, vol. 9, no. 5, pp. 20–35, 2021.

[18] M. Taeb and H. Chi, "Comparison of Deepfake Detection Techniques through Deep Learning," *Journal of Cybersecurity and Privacy*, vol. 2, no. 1, pp. 89–106, Mar. 2022.

[19] I. Castillo Camacho and K. Wang, "A comprehensive review of deep-learning-based methods for image forensics," *J Imaging*, vol. 7, no. 4, p. 69, 2021.

[20] J. Mallet, R. Dave, N. Seliya, and M. Vanamala, "Using Deep Learning to Detecting Deepfakes," *arXiv preprint arXiv:2207.13644*, 2022.

[21] H. F. Shahzad, F. Rustam, E. S. Flores, J. Luís Vidal Mazón, I. de la Torre Diez, and I. Ashraf, "A Review of Image Processing Techniques for Deepfakes," *Sensors*, vol. 22, no. 12, p. 4556, 2022.

[22] A. Rahman *et al.*, "A qualitative survey on deep learning based deep fake video creation and detection method," *Aust. J. Eng. Innov. Technol*, vol. 4, no. 1, pp. 13–26, 2022.

[23] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake detection for human face images and videos: A survey," *IEEE Access*, vol. 10, pp. 18757–18775, 2022.

[24] H. Huang, P. S. Yu, and C. Wang, "An introduction to image synthesis with generative adversarial nets," *arXiv preprint arXiv:1803.04469*, 2018.

[25] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[26] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 43, no. 12, pp. 4217–4228, Dec. 2021.

[27] T.-C. Wang *et al.*, "Video-to-video synthesis," *arXiv preprint arXiv:1808.06601*, 2018.

[28] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.

[29] S. Qian *et al.*, "Make a face: Towards arbitrary high fidelity face manipulation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10033–10042, 2019.

[30] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2387–2395, 2016.

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

[32] Y. Nirkin, Y. Keller, and T. Hassner, "Fsgan: Subject agnostic face swapping and reenactment," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7184–7193, 2019.

[33] H. Averbuch-Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen, "Bringing portraits to life," *ACM Trans Graph*, vol. 36, no. 6, pp. 1–13, Dec. 2017.

[34] T. Wang, K. P. Chow, X. Chang, and Y. Wang, "Deepfake Detection: A Comprehensive Study from the Reliability Perspective," *arXiv preprint arXiv:2211.10881*, 2022.

[35] H. Kim *et al.*, "Deep video portraits," *ACM Trans Graph*, vol. 37, no. 4, pp. 1–14, Aug. 2018.

[36] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.

[37] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," *IEEE J Sel Top Signal Process*, vol. 14, no. 5, pp. 910–932, Aug. 2020.

[38] T. T. Nguyen *et al.*, "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding*, vol. 223, p. 103525, 2022.

[39] Y. He *et al.*, "ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis (Supplementary Material)" *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.

[40] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[41] X. Li *et al.*, "Sharp Multiple Instance Learning for DeepFake Video Detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, New York, NY, USA: ACM, pp. 1864–1872, Oct. 2020.

[42] L. Guarnera, O. Giudice, and S. Battiato, "Fighting Deepfake by Exposing the Convolutional Traces on Images," *IEEE Access*, vol. 8, pp. 165085–165098, 2020.

[43] M. A. Sahla Habeeba, A. Lijiya, and A. M. Chacko, "Detection of Deepfakes Using Visual Artifacts and Neural Network Classifier," in *Innovations in Electrical and Electronic Engineering*, M. N. Favorskaya, S. Mekhilef, R. K. Pandey, and N. Singh, Eds., Singapore: Springer Singapore, pp. 411–422, 2021.

[44] K. Huang, A. Hussain, Q.-F. Wang, and R. Zhang, *Deep learning: fundamentals, theory and applications*, vol. 2. Springer, 2019.

[45] H. A. Ahmed and E. A. Mohammed, "Detection and Classification of The Osteoarthritis in Knee Joint Using Transfer Learning with Convolutional Neural Networks (CNNs)," *Iraqi Journal of Science*, vol. 63, no. 11, pp. 5058–5071, Nov. 2022.

[46] J. B. Awotunde, R. G. Jimoh, A. L. Imoize, A. T. Abdulrazaq, C.-T. Li, and C.-C. Lee, "An Enhanced Deep Learning-Based DeepFake Video Detection and Classification System," *Electronics (Basel)*, vol. 12, no. 1, p. 87, Dec. 2022.

[47] A. Ismail, M. Elpeltagy, M. S. Zaki, and K. Eldahshan, "A New Deep Learning-Based Methodology for Video Deepfake Detection Using XGBoost," *Sensors*, vol. 21, no. 16, p. 5413, Aug. 2021.

[48] V. V. V. N. S. Vamsi *et al.*, "Deepfake detection in digital media forensics," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 74–79, Jun. 2022.

[49] W. Zhang, C. Zhao, and Y. Li, "A Novel Counterfeit Feature Extraction Technique for Exposing Face-Swap Images Based on Deep Learning and Error Level Analysis," *Entropy*, vol. 22, no. 2, p. 249, Feb. 2020.

[50] V.-N. Tran, S.-H. Lee, H.-S. Le, and K.-R. Kwon, "High Performance DeepFake Video Detection on CNN-Based with Attention Target-Specific Regions and Manual Distillation Extraction," *Applied Sciences*, vol. 11, no. 16, p. 7678, Aug. 2021.

[51] X. Ding, Z. Raziei, E. C. Larson, E. V Olinick, P. Krueger, and M. Hahsler, "Swapped face detection using deep learning and subjective assessment," *EURASIP J Inf Secur*, vol. 2020, no. 1, pp. 1–12, 2020.

[52] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15023–15033, 2021.

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

[54] S. A. Khan and D.-T. Dang-Nguyen, "Hybrid Transformer Network for Deepfake Detection," in *Proceedings of the 19th International Conference on Content-based Multimedia Indexing*, pp. 8–14, 2022.

[55] B. Kaddar, S. A. Fezza, W. Hamidouche, Z. Akhtar, and A. Hadid, "HCiT: Deepfake Video Detection Using a Hybrid Model of CNN features and Vision Transformer," in *2021 International Conference on Visual Communications and Image Processing (VCIP)*, IEEE, pp. 1–5, Dec. 2021.

[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[58] C. X. T. Du, L. H. Duong, H. T. Trung, P. M. Tam, N. Q. V. Hung, and J. Jo, "Efficient-Frequency: a hybrid visual forensic framework for facial forgery detection," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, pp. 707–712, Dec. 2020.

[59] J. H. Bappy, C. Simons, L. Nataraj, B. S. Manjunath, and A. K. Roy-Chowdhury, "Hybrid lstm and encoder–decoder architecture for detection of image forgeries," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3286–3300, 2019.

[60] S. S. Khalil, S. M. Youssef, and S. N. Saleh, "iCaps-Dfake: An integrated capsule-based model for deepfake image and video detection," *Future Internet*, vol. 13, no. 4, p. 93, 2021.

[61] M. Bonomi, C. Pasquini, and G. Boato, "Dynamic texture analysis for detecting fake faces in video sequences," *J Vis Commun Image Represent*, vol. 79, p. 103239, 2021.

[62] S. Suratkar and F. Kazi, "Deep Fake Video Detection Using Transfer Learning Approach," *Arab J Sci Eng*, vol. 48, pp. 9727–9737, Oct. 2022.