

A systematic literature review on the effectiveness of deepfake detection techniques

Laura Stroebel, Mark Llewellyn, Tricia Hartley, Tsui Shan Ip & Mohiuddin Ahmed

To cite this article: Laura Stroebel, Mark Llewellyn, Tricia Hartley, Tsui Shan Ip & Mohiuddin Ahmed (2023) A systematic literature review on the effectiveness of deepfake detection techniques, *Journal of Cyber Security Technology*, 7:2, 83-113, DOI: [10.1080/23742917.2023.2192888](https://doi.org/10.1080/23742917.2023.2192888)

To link to this article: <https://doi.org/10.1080/23742917.2023.2192888>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 29 Mar 2023.



Submit your article to this journal



Article views: 8422



View related articles



CrossMark

View Crossmark data



Citing articles: 28

REVIEW ARTICLE

OPEN ACCESS



A systematic literature review on the effectiveness of deepfake detection techniques

Laura Stroebel, Mark Llewellyn, Tricia Hartley, Tsui Shan Ip
and Mohiuddin Ahmed 

School of Science, Edith Cowan University, Joondalup, Western Australia, Australia

ABSTRACT

With technological advances, the generation of deepfake material is now within reach of those operating consumer-grade hardware. As a result, much research has been undertaken on deepfake detection techniques. This work has analysed and measured the performance of various detection techniques using multiple metrics and discussed the effectiveness of these deepfake detection techniques. This has been undertaken by examining and analysing the current state of deepfake detection techniques. Unlike other existing surveys, this work produced a Systematic Literature Review (SLR) on research conducted from the beginning of 2021 to August 2022. This SLR includes tabulated data containing details of the techniques used and the accuracy of those techniques, performance metrics provided in each study, summaries of the datasets used, and challenges and future trends. This SLR has been undertaken with a focus on using a mixed methods approach. This SLR has determined that deep learning (DL) has surpassed machine learning (ML) as the preferred deepfake detection model. However, ML is still a primary focus method in medical imagery. It was also discovered that traditional artificial neural networks are no longer effective and require additional modules to produce ensembled and multi-attentional architectures.

ARTICLE HISTORY

Received 29 November 2022
Accepted 14 March 2023

KEYWORDS

Deepfake detection
techniques; machine
learning; deep learning;
methods; dataset

1. Introduction

The remarkable technological advancements, especially in artificial neural networks, coupled with high computing power, have led to the creation of technologies that can be used to tamper with digital content. These technologies, including FakeApp and FaceApp, have been used intensely in the recent past to generate real-looking, but fake, visual content. Most of these technologies enable a creator to alter various attributes in image and video content, like hairstyle, age, voice, and many others, or even to swap the entire face in the

CONTACT Mohiuddin Ahmed  m.ahmed.au@ieee.org  School of Science, Edith Cowan University Perth, Western, Australia

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

image with another. This has led to the concept of Deepfake, a name coined from Deep Learning and Fake, which exploits the powers of deep learning to create fake, realistic video and image content. Most of these deepfake technologies result from two core neural network technologies: generative network and discriminative network which when combined, give rise to Generative Adversarial Networks (GANs).

Deepfake content has a wide range of useful applications across many industries, that are yet to be fully harnessed. Abdulreda and Obaid (2022), in their study [2], discuss techniques of facial modification. They argue that Entire Face Synthesis is a facial modification technique used to produce non-existent faces. Methods such as StyleGAN utilise GANs and can produce high-quality face pictures that can be utilised in video games and 3D modelling. However, the technology has been widely used by malicious actors for malicious intent. For example, as early as 2017, deepfake technologies were used to share pornographic videos that had celebrity faces as actors. Also, these technologies have been used to spread misinformation in various avenues, mainly across social media platforms, conducting cybercrimes and creating political tension and instability.

This has prompted the need to detect and distinguish between what is real and what is fake. So far, much research has been done on deepfake detection techniques and the ability of these techniques to identify deepfake images and videos using artificial intelligence (AI) models trained on well-established datasets of deepfake and typical examples. However, there has been limited research undertaken on the validity of the training datasets used by these AI models which, due to a lack of annotation of the features and potential bias of the content, may cause the deepfake detection tool to fail [3].

Bias in the datasets used for AI and machine learning has been suspected for some time and studies suggest AI models trained on unbalanced datasets will be biased against certain groups and will perform poorly. Deepfake detection technologies using these AI models will sometimes fail to identify deepfake content due to the lack of diversity in the dataset they are trained on. Further to this, Xu et al. (2022) propose that the deepfake detection tools themselves have a bias that influences results and this is further exacerbated by the bias in the dataset used [4].

The limited number of audio datasets, particularly non-English based, is another area where detection of deepfake media is lacking. Adding to this, with the impact of accents in the current audio datasets yet to be investigated, and audio deepfakes increasing, the need for further study in this area is paramount.

Table 1 summarises the literature reviews/surveys that have already been undertaken on this subject – these will be discussed in more detail in [section 3.1](#).

The main contributions of this SLR are:

- A comprehensive survey of current literature relating to deepfake detection technologies – specifically, those produced between January 2021 and August 2022.
- A comprehensive analysis and discussion of new deepfake detection techniques. This study details the most recent techniques and provides performance metrics for them.
- Discussion of the challenges faced in deepfake detection techniques.
- Gives highlights of what the future of deepfake detection will look like based on the most recent advancement in the domain.

This paper has been organised in the following manner:

- **Section 2:** The systematic literature review (SLR) process, including the criteria used to identify viable source documentation and the research questions asked.
- **Section 3:** Discuss other SLR's undertaken on this subject and overview modern deepfake detection methods.
- **Section 4:** Discuss this SLR's findings and introduce some limitations, challenges and future trends identified.
- **Section 5:** Conclusion.

2. Systematic literature review process

This SLR aimed to create a body of knowledge of deepfake detection techniques and to conduct a systematic review of the currently available literature regarding these techniques. The main objective was to undertake an SLR that analyses the effectiveness of deepfake detection techniques [1]. Our specific objectives were to:

- Examine and analyse the current state of deepfake, providing an up-to-date overview of recent research work on deepfake detection.
- Analyse and measure the performance of various deepfake detection techniques using multiple metrics; and
- Identify and discuss significant advances, challenges, and future trends.

To achieve these objectives, it was important to conduct our study in an organised and systematic way. The section below illustrates our approach.

2.1. Search strategy

An important part of this literature review was to gather existing publications that proposed deepfake detection techniques. We did not cover deepfake generation techniques. The initial search strategy was based



Table 1. Summary of previous works in comparison to our study. (ref)* means references listed as reviewed document numbers are not noted in the paper and thus the number is inferred from a paper's reference section. The (-) under the generation and detection models means the study did not explore the technology under that column.

Prior Literature Reviews/Surveys	Deepfake Researched	Completed Date	Documents Reviewed	Generation Models	Detection Models	Datasets Identified	Challenges
Weerawardana and Fernando [5]	Detection	August 2021	48 (ref)*	-	17	7	No definitive deepfake detection method identified and lack of quality datasets
Zhang [6]	Generation/ Detection	September 2021	99 (ref)*	12	25	15	Lack of benchmark test methods and quality datasets
Malik et al. [7]	Generation/ Detection	January 2022	130 (ref)*	17	40	17	Include lack of datasets, unknown types of attacks, temporal aggregation and unlabelled data
Juefei-Xu et al. [8]	Generation	March 2022	318+	91	117	24	Lack of quality datasets, lack of competitive baselines, generalisation and robustness, evaluation platform and metrics
Rana et al. [9]	Detection	February 2022	112	-	91	18	Need a framework to reduce inconsistencies and to produce definitive and systematic outcomes
Celebi et al. [10]	Detection	February 2022	20 (ref)*	-	9	7	There are 100 times more people working on deepfake generation than on detection
Almutairi and Elgibreen (2022) [11]	Detection	May 2022	57 (ref)*	-	22	8	Limited non-English detection methods, lack of research on accents and "noise", excessive pre-processing required
Masod et al. (2022) [12]	Generation/ Detection Up to March 2022	May 2022	436	100	67	16	Generalisation, Temporal coherence, Datasets integrity, Temporal aggregation

upon the strategy presented in [9] and was conducted using the standard research databases. **Table 2** shows the search strategy used in this paper.

Further investigation on identified areas of interest was done via the data sources in **Table 2** as well as other scholarly literature databases such as arXiv and Research Gate. **Figure 1** shows the search results for this SLR – made up of 83 documents reviewed and considered relevant for this SLR, 21 artifacts collected for reference purposes only and 16 documents reviewed and found to not fit the scope of this SLR based on the documents creation date or content.

Figure 2 shows the source type of the articles included in the Reviewed Works section of this document.

2.2. Research questions

We developed research questions to be used to assess the current literature and to gather relevant information from these sources. These questions also served the purpose of focussing our review on the areas we believed to be most relevant and important to this area of research. The right questions needed to be asked, so that the research community could benefit from useful, targeted information. **Table 3** outlines the research questions, as well as their purpose.

2.3. Data extraction and recording

Due to the large amount of data to collect, sort through and categorise, it was important to do this in an organised and meaningful way. This enabled us to monitor the data, ensure information was not missed and be able to extract meaning from the data.

Table 2. SLR search strategy.

Type	Description
Aim	Gather existing literature that outlined deepfake detection techniques, not limited to video and audio.
Data Sources	The majority of articles were located in: IEEE Publications Database, ProQuest Central, ScienceDirect and SpringerLink.
Query	A keyword search was conducted using the following query: deepfake OR faceswap OR “video manipulation” OR “fake face” OR “fake image” OR “fake video” AND detection OR detect OR “facial manipulation” OR “digital media forensics”. Articles were limited to those published in 2021 and 2022. NB: A search of titles only was attempted however there were too few results.
Method	Using the search query above, we obtained 411 search results. The abstracts were reviewed to determine if the publication related to deepfake detection techniques. These were then entered into a spreadsheet after which they were reviewed in more detail.
Inclusions/exclusions	Books were excluded, as were articles that did not contain information on deepfake detection or metrics for the deepfake detection technique discussed. Also, excluded were documents published before 2021.

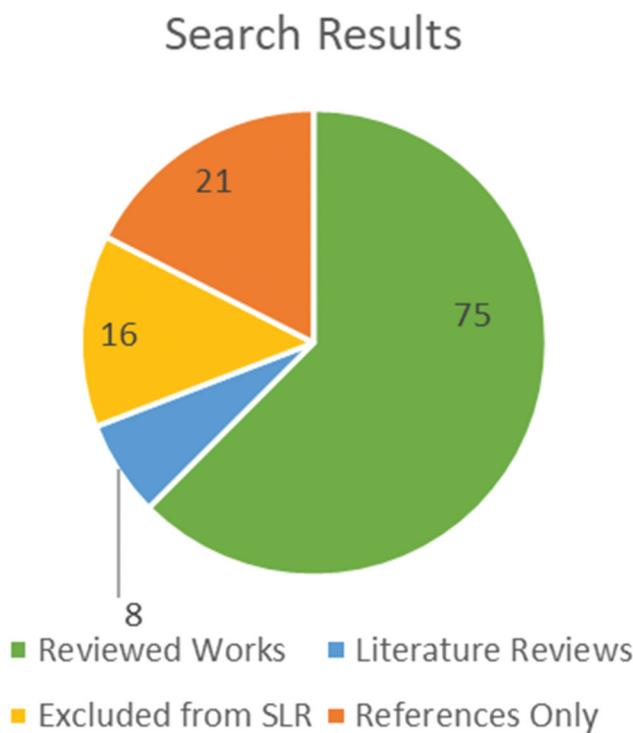


Figure 1. Documents returned by our search strategy and how they were categorised after review.

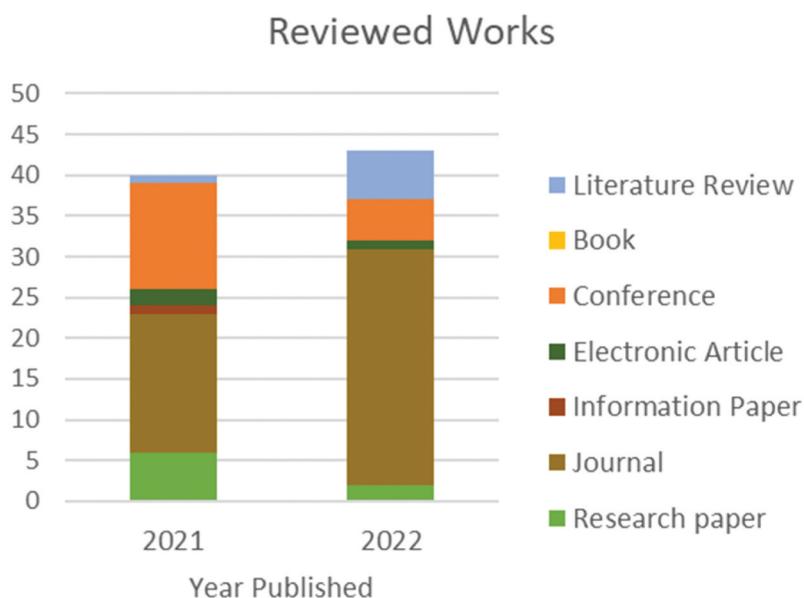


Figure 2. The source type of the documents (by year) included as reviewed works.

Table 3. Research questions.

Number	Question	Purpose
RQ1	What are the current deepfake detection techniques?	Determine the currently proposed detection techniques.
RQ1A	What is the effectiveness of these techniques?	Assess the effectiveness of the detection techniques using common metrics.
RQ2	What advances have occurred since the previous SLRs?	Describe the positive changes that have occurred.
RQ3	What are the current challenges involved in deepfake detection?	Identify areas that challenge successful deepfake detection.
RQ4	Can any future trends be identified, and if so, what are they?	Identify trends in deepfakes and their detection.

2.4. Data analysis and visualisation

The purpose of this phase was to review, compare and analyse the data obtained from all sources. This allowed us to generate insights and determine any trends that were present. The data was presented visually using methods such as pie charts and tables.

3. Reviewed works

For this SLR the works reviewed could be categorised in two ways – previous systematic literature reviews/surveys and modern deepfake detection methods.

3.1. Previous literature reviews/surveys

We reviewed 8 literature review/survey documents – 1 published in 2021 and 7 published in 2022. The majority included deepfake generation and detection techniques [6–8,12] and the remaining focused only on deepfake detection techniques [5,9–11]. However, while Celebi et al. (2022) discussed deepfake detection techniques, it was in the context of evidence for judicial proceedings, rather than specific models [10].

Masood et al. (2022) surveyed publications relating to audio/video manipulation, specifically generation and detection techniques available up to March 2022. The comprehensive study [12] covered works from 2017 to 2022, resulting in a total of 436 papers being reviewed. It [12] found some of the limitations of generation processes were generalisation across source datasets, pose variations, distance from the camera scenarios, and temporal coherence. While the challenges identified with deepfake detection technologies were the quality, fairness, and trust of deepfake datasets, temporal aggregation, and social media laundering, - to name a few.

Juefei-Xu et al. (2022) focused solely on facial deepfakes, thoroughly examining available generation and detection techniques, relating to this area, up to June 2021. This paper [8] noted that deepfake detection techniques were

proportionate to deepfake generation techniques and vice versa – that is, the creation of generation techniques was driven as much by the creation of detection techniques as the other way around (detection by generation). This survey [8] noted a need for competitive baselines to evaluate deepfake detection techniques so true performance could be assessed. Zhang (2022) agreed, indicating the deepfake detection techniques up to 2020, reviewed in this survey [6], were not robust and there were many challenges when assessing generation and detection models, including a lack of benchmark test methods and quality datasets.

Malik et al. (2022) took a similar approach, evaluating face image and video deepfake techniques (generation and detection) up to early 2021, and concluded there was a general inability in the detection models to transfer and generalise indicating further research was needed [7].

Rana et al. (2022) undertook a detailed SLR of 112 articles relating to deepfake detection technologies, published between 2018 and 2020, inclusive. They [9] classified these technologies into 4 categories (deep learning-based techniques, classical machine learning-based methods, statistical techniques, and blockchain-based techniques) and evaluated the performances of each when used with the datasets available at the time. At the conclusion of this SLR [9], deep learning-based methods were identified as outperforming all other categories. This conclusion was also noted by Weerawardana and Fernando (2021), who categorised the deepfake detection techniques they reviewed into traditional and deep learning methods only [5].

Almutairi and Elgibreen (2022) identified three types of audio deepfakes – synthetic-based, imitation-based and replay-based. This survey [11] examined machine and deep learning audio deepfake detection technologies and determined that machine learning methods proved more accurate than deep learning, but required excessive training and manual feature extraction, potentially making them unscalable. It [11] noted a lot more study in this area was needed to address existing gaps and challenges.

3.2. Modern deepfake detection methods

We reviewed 75 modern deepfake detection technology documents to assess developing trends and identify future research areas. Of these, the majority focused on video, image, or a combination of image/video deepfakes (90%). The remaining documents looked at audio or a combination of audio/video and all media.

There were several machine learning models proposed using a generative adversarial network (GAN) framework [13–15]. However, the majority were deep learning architectures based on artificial neural networks – such as convolutional neural network (CNN), deep neural network (DNN) and long short-term memory (LSTM). **Table 4** details the deepfake detection techniques discussed in

the reviewed documents. In total, 48 deepfake detection models were identified that met the objectives of this SLR. [Figure 3](#) is a cluster map of the noteworthy deepfake detection techniques and shows that most models were CNN based.

An overview of the identified techniques will be delivered in the following structure:

- Machine learning deepfake detection techniques.
- CNN, DNN and LSTM techniques used as standalone models and those extended to multifaceted architectures.
- Techniques to detect audio deepfakes.
- Datasets used to train, test and validate deepfake detection models - [Table 6](#) details the datasets used by the deepfake detection models reviewed in this SLR.

3.2.1. Machine learning

MSTA_Net, proposed by Yang et al. (2022) was a machine-learning deepfake detection model which examined the texture features of an image to identify anomalies. This model [13] used the whole image in the detection process, not just focused facial areas, to link the forged and non-forged areas of the image. That is, if inconsistency in the image texture was found the image was tagged fake, but if no inconsistency was found, the image was tagged as non-fake.

Other machine learning models included a self-supervising decoupling network model suggested by Zhang et al. (2021) that utilised dual feature learning and performed well with lower-quality images [86]. As well as a hybrid texture/noise model proposed by Fu et al. (2022) which used local binary pattern and subtractive pixel adjacency matrix features to detect and identify facial manipulation in deepfake images [17].

3.2.2. CNN, DNN, LSTM and RNN

EfficientNet-V2, discussed by Deng et al. (2022) was a CNN model evolved from the EfficientNet series first introduced in 2019. This more streamlined version [44] used fewer resources to produce higher test accuracy. It also could be used to identify multiple faces in an image and track them across multiple frames. However, it was noted [44] that the pre-processing required by the model caused some inconsistencies which needed manual intervention, and the model was not tested for generalisability. Wang et al. (2021) used EfficientNet as the testing model for the adversarial training framework they proposed [43].

Other CNN models discussed, were various versions of XceptionNet [18,22,47,87] and ResNet [20,21,23,49]. A unique 3-layer frequency CNN was proposed by Kohli et al. (2021) to work with a two-dimensional global discrete Cosine transform model [27]. As well as a lightweight 3D CNN [24] model which extracted features in the spatial-temporal dimension, proposed by Liu et al. (2021).

**Table 4.** The deepfake detection techniques proposed in the reviewed documents.

DFD Techniques	DFD Models/Methods	DFD Dataset	DF Media Researched Documents	Source	Year	Metrics (NB: averaged if multiple results)	
						DF Type	AUC (%)
Machine Learning Models							
SSDN	(Self-supervised decoupling network	FF++	Image	[16]	2021	FM	98 (HQ) 91.8 (LQ)
GAN	MSTA-Net – multi-scale self-texture attention generative network	FF++, CDF, DFDC, DFOr-1.0 FFHQ	Image	[13]	2022	FS	FF++ 92 (HQ) FF++ 88 (LQ)
SVM	Hybrid Texture/Noise, SVM – using local binary pattern and subtractive pixel adjacency matrix features	FF++	Image	[17]	2022	FM	99.4 97.6
Deep Learning Models							
CNN	XceptionNet – comparison test	SYSU-OBFORG, SULFA	Video	[18]	2021	FM	FF++ 99.3
CNN	Yolo (V2)	CDF, FF++ [20]	Video	[19]	2021	VM	99
CNN	ResNet-50 - face recognition model pretrained on MS1M-ArcFace	RFID [21]	Image	[20,21]	2021	FS	97 [21]
	/WebFace2/2M [20]						
	Comparison test [21]						
CNN	Passive/Proactive method (Bilateral filtering/joint adversarial training – XceptionNet/ MesInceptionNet)	FF++	Video	[22]	2021	FM	99.7 [21] highest result 90 (over many techniques/ metrics)
CNN	Three-part architecture – using separate image noise analysis on manipulated and blending artifacts in an image	FF++, CDF	Image	[23]	2021	FM	FF++ 99 CDF 99 CDF 99
CNN	FF++, DFTIMIT, DFDC, CDF [24]	Video	[24,25]	2021	FM	FF++ 99 CDF 81 FF++/CDF	FF++ 99 (HQ/ LQ) [25]
	FF++, VidTIMIT [25]						
CNN	3D CNNs	FF++, DFTIMIT, DFDC, CDF [24]	Image	[26]	2021	FM	FF++ 99/94 (HQ/ LQ) [25]
CNN	Shallow-FakeFaceNet (SFFN) (Shallow CNN)	Own dataset HFM, RFFD, CelebA	Image	[27]	2021	FM	DFTIMIT 99 DFDC 93 DFDC 93 CDF 98 [24]
CNN	fCNN with two-dimensional global discrete Cosine transform (2D-GDCT)	FF++, CDF	Video	[28]	2021	FM	64×64 94 1024×1024 99.9
DNN	Dual path system – using transformer and node-compressed architecture	DFDC, CDF, FF++	Video				CDF 66.5 FF++ 96.7 DFDC 94.9 CDF 92.3 DFDC 88.4 CDF 74.3

(Continued)

**Table 4.** (Continued).

DFD Techniques	DFD Models/Methods	DFD Dataset	DF Media Researched	Source Documents	Year Pub	DF Type	Metrics (NB: averaged if multiple results)	
							AUC (%)	ACC (%)
CNN++	Bita-Net- utilising ResNet-50 and U-Net	Trained on DFDC FF++, CDF, DeepfakeMIT, UADFV	[29]	2021	FM		FF++ 99	CDF 99
CNN++	FST-Net (dual-stream architecture) - utilising ResNet-152 and S3D (separable 3D CNN)	Trained on ImageNet FF++	Video	[30]	2021	FM	97 (LQ)	DFTIMIT 99
CNN++	AMTENnet – a combination of AMTEN and CNN FeatureTransfer – two-stage adversarial training pipeline	HFF, FF++ Trained on ImageNet FF++, DFTIMIT, DFDC, CDF,DFD	Video Video	[31] [32]	2021 2021	FS FM	98 (HQ) 93 (LQ)	UADFV 99 (RAW) HFF 99
CNN++	AFIFN (Advanced Fake Image-Feature Network) using DCT-based processing and Y CrCb based pre-processing	CelebA	Image	[33]	2021	FM	DFDC 79	DFTIMIT 96
CNN++	CWSA-Net (Channel-Wise Spatiotemporal Aggregation) with EfficientNet-B0 backbone	FF++, CDF, DFDC	Video	[18]	2021	FM	CDF 98	DFDC 79
CNN++	PGT, SWDCT_IDCT	FF++, CDF, WDF	Image, video	[34]	2021	FS	FF++ 1.0	WDF 79.8
CNN++	LRNet	UADFV, FF++, CDF, Dfor-1.0	Video	[35]	2021	VM	WDF 87.8	Dfor-1.0 97.74
CNN++	DCWNet (dual-tree complex wavelet-based face forgery network) with ResNet-34 backbone	FF++	Image	[36]	2021	FM	UADFV 98.5	FF++ 99.7
CNN++	DeepfakeNet – combination of ResNet and Inception model	FF++, Kaggle, DFTIMIT	Video	[37]	2021	FM	CDF 56.9	FF++ 96.7

(Continued)

**Table 4.** (Continued).

DFD Techniques	DFD Models/Methods	DFD Dataset			DF Media Researcher Documents	Source Documents	Year Pub	DF Type	AUC (%)	Metrics (NB: averaged if multiple results)
		FF++, DFD, UADFV	Video	[38]						
CNN++	Multi-dimensional biological signals									FF++ 96 DFD 97 UADFV 95 CASIA 94.7 Own dataset 83.2
CNN++	SiteForge CNN with added Local Interpretable Model-agnostic Explanations (LIME)	CASIA 2.0 Own dataset Twitter Indian Dataset 2.0 FF++	Image	[39]	2021	FM/IM				
CNN++	ResNet-18 - pre-processed with Image saliency and guided filter processing	DFTIMIT, FF++, DFDCP	Image	[40]	2021	FM				FF++ 99
SCNN	Set Convolutional Neural Network (SCNN) using MesoNet/XceptionNet	For [42]	Audio	[41]	2021	FM	80/99 (HQ)	80/96(HQ)	80/94 (LQ)	
CNN	TGN – temporal convolution network	For [42]	Audio	[11,42]	2021	FM	79/95 (LQ)	92 [42]		
CNN	STN – spatial transformer network	For [42]	Audio	[11,42]	2021	FM				80 [42]
CNN	EfficientNet-B0 - adversarial trained on FF++ Comparison test	FF++ DFDC training	Image/ Video	[18,43]	2022	FM				69 [43] 96 (many datasets) [18]
CNN	EfficientNet-V2	FF++, FF1W-10K	Video	[44]	2022	FS				FF++ 98
DNN	MC-LCR – multimodal contrastive classification by locally correlated representations Xception (pre-trained on ImageNet), MLP-Mixer	FF++, CDF, DFDC, DF-1.0	Video	[45]	2022	FM	FF++ 99/90 FF++/DFDC 71	FF++ 10K 93 FF++ 98/88 FF++/CDFC 71 FF++/DFDC 70		
DNN	Face/background squares captured/Siamese noise trace extraction/ noise similarity analysis	CDF, UADFV	Video	[46]	2022	FS	CDF 99.92 UADFV 88.95	FF++ 1.0 / 0.75		
CNN + ViT	ViXNet	FF++, CDF, DFID	Image	[47]	2022	FM	FF++ 99/75 CDF 99/73 DFD 99/75	FF++ 97/69 CDF 94/67 DFD 95/68		

(Continued)

Table 4. (Continued).

DFD Techniques	DFD Models/Methods	DFD Dataset	DF Media Researched Documents	Source	Year Pub	DF Type	Metrics (NB: averaged if multiple results)	
							AUC (%)	ACC (%)
CNN++	ID-Reveal – 3D morphable model (3DMM) + Temporal ID Network + GAN	DFD	Video	[48]	2022	FM	87/96 (FR/ FS)[HQ] 90/94 (FR/ FS)[LQ]	76/85 (FR/FS)[HQ] 82/78 (FR/FS)[LQ]
CNN++	Dual-level collaborative framework (un-named) - utilising ResNet-50 and temporal learning models	Trained on ImageNet FF++, CDF, DFDC	Video	[49]	2022	FM	FF++ 99 CDF 98 DFDC 99	FF++ 95 CDF 96 DFDC 99
CNN++	DenseNetXX – extends ResNet by adding transition layer	RFFD	Image	[21,50]	2022	FM [50] FS [21]	DN121 97.1 [21] DN169 95 [21]	DN121 97 [21] DN169 95 [21]
CNN++	VGG-XX – fully connected classifiers followed by maxpooling layers	RFFD	Image	[21,50]	2022	FM [50] FS [21]	DN201 99.4 [21]	DN201 99.4 [21]
CNN++	CustomCNN – analyses dropout, padding, augmentation and grayscale on model performance	RFFD	Image	[50]	2022	FM	VGG-19 96 [50]	VGG-19 95 VGG-19 94 [21]
CNN++	LISiam (Localisation Invariance Siamese Networks) using Xception backbone	FF++, CDF	Video	[51]	2022	FM	VGG-19 98.7 [21] VGG-16 97.7 [21]	VGG-19 94 [21] VGG-16 92 [21] VGG-Face 99 [21]
CNN++ CNN++	Multilayer Perceptron Frame-temporality two-stream convolutional network – with MesoNet – for frame-level and ResNet-18 for time-dependent residual features	CDF, FF++	Video Video	[52] [53]	2022	FM FS	81 CDF 87 FF++ 94	87 CDF 80.74 FF++ 86.61

(Continued)

Table 4. (Continued).

DFD Techniques	DFD Models/Methods	DFD Dataset	DF Media Researched	Source Documents	Year Pub	DF Type	Metrics (NB: averaged if multiple results)		
							AUC (%)	ACC (%)	
CNN++	Patch-DFD – using RESNet-50 and Inception-v3, pre-trained on ImageNet	DFTIMIT, CDF, FF++	Image	[54]	2022	FM	DFTIMIT 99.42/99.1	FF++ 96.23/87.36	
RNN (CNN +LSTM)	Xception-ConvLTSM – with spatiotemporal attention mechanism	FF++, CDF, DFDC	Video	[55]	2022	FS	CDF 98.88 FF++ 99	CDF 99	
RNN (CNN +LSTM)	EfficientNet-B3 [56] With vanilla LSTM [57]	FF++ [56] Celeb-DF [57]	Video [56] Image [57]	[56,57]	2022	FS [56,57] FM [56]	Celeb-DF 94 DFDC 92 FF++ 99 FF++ 99 (raw) [56] FF++ 99 (HQ) [56] FF++ 91 (LQ) [56]	[57]	
RNN (CNN +LSTM)	Meso-4 with vanilla LSTM	Celeb-DF	Image	[57]	2022	FM	Celeb-DF 96	Celeb-DF 93.9 [57]	
RNN (CNN +LSTM)	ResNet and LSTM	CDF	Video	[58]	2022	FM	CDF 88.8	CDF 91	
Jointly modelling video and audio modalities for deepfake detection		Other Models							
PSCC-Net [60] Progressive Spatio-Channel Correlation Network		FF++, DFDC		Video and Audio		2021 AM/VM		FF++ 95	
EfficientNet and Vision Transformers [62]		Own dataset		Image Video		[61] 2022 AM VM		99.6 95.1	

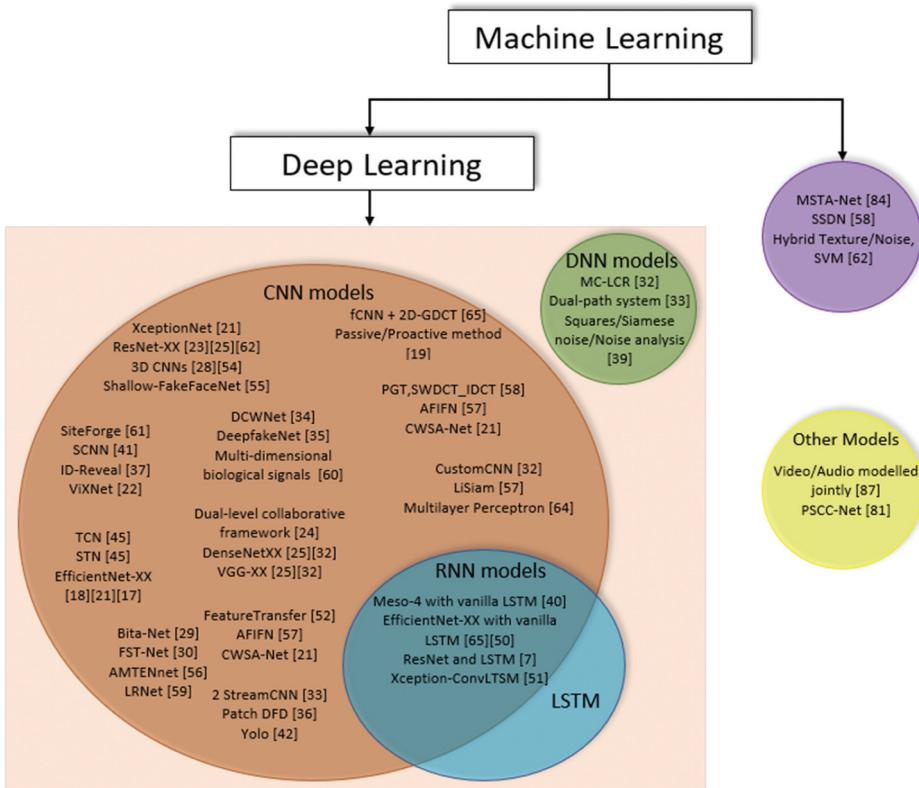


Figure 3. Cluster map of identified deepfake detection models.

Many papers combined the traditional techniques (CNN, DNN, LTSM) with other modules to produce ensembled and multi-attentional architectures. For example, we found a number of models that extended various versions of ResNetXX. Bita-Net, created by Ru et al. (2021), was based on human detection methods – using a temporal examination at high frame rate and frame by frame scrutiny of key frames with an additional attention branch- using ResNet-50 and U-Net. This model [29] produced excellent, consistent results across multiple datasets, even when trained by one dataset and then validated and tested on another.

Pu et al. (2022) examined videos with frame-level and video-level methods to identify fake content. ResNet-50 extracted facial features, and then fed them to a temporal learning model and a video and frame level classifier, in a three-step examination process to determine if a video was fake or not. This model [49] was touted to outperform all existing methods for frame and video level detection, as well as being robust to video quality and database variations. However, it [49] was not generalisable to other types of facially manipulated images or videos, such as those that were GAN-generated (completely made-up).

Table 5. The description of the acronyms used in Table 4.

Acronym	Meaning
AM	Audio Manipulation
CNN + Vit	Convolution neural network extended with a vision transformer
CNN++	Convolution neural network extended with other modules/components
FM	Facial Manipulation
FS	Face Swapping
IM	Image Manipulation
LSTM++	Long short-term memory extended with other modules/components
SCNN	Set convolutional neural network
SVM	Support vector machine
VM	Video Manipulation

Zhang et al. (2021) proposed a dual-stream architecture, called FST-Net, made up of frequency and spatial-temporal streams, enhanced with an attention mechanism on each, and using ResNet-152. This model [30] proved successful across high and low-quality videos but had limitations when used with some datasets, which reduced its generalisability. Future work [30] suggested for this model involved incorporating synthetic audio detection to improve accuracy. Along a similar line, Chen et al. (2022) proposed a spatiotemporal attention mechanism that worked with an LSTM to enhance its functionality. This model [55] had a 4 step architecture made up of a spatiotemporal attention mechanism, a CNN, an LSTM and a predictor to make the final decision.

Other models which used ResNetXX as their backbone were DenseNetXX – a model used for comparative testing in several papers [21,50]. A two-stream model created by Hu et al. (2022) analysed frame-temporality streams in videos and using ResNet-18 [53]. A dual-tree complex wavelet transform-based network [36] with ResNet-34 as the backbone of the two streams was proposed by Gao et al. (2021). DeepfakeNet, based on a combination of ResNet's stacking feature and Inception's split-transform-merge feature [37], was introduced by Gong et al. (2021). This technique was also discussed by Yu et al. (2022) using ResNet-50 and Inception-v3 [54].

ID-Reveal created by Cozzolino et al. (2022), combined a facial feature extractor, a temporal network for biometric anomaly detection and a GAN to predict expression-based motion. This model [48] successfully identified deep-fakes in high and low-quality videos, producing good results in detecting face-swapping instances but less so for other types of facial manipulation. This model [48] needed to be trained on pristine videos of the source identity, which was limiting, but meant it was generalisable and robust when used with other datasets.

MC-LCR, created by Wang et al. (2022) was a DNN using a multimodal contrastive classification by locally correlated representations [45]. It was a two-stream framework working in the spatial and frequency domains that, while

giving good results on high and low-quality images, seemed to struggle with varying light conditions and biometric challenges.

A dual-path system proposed by Luo et al, (2021) achieved good results using neural ordinary differential equations (NODE) and a facial feature transformer architecture [28]. However, the technique was suitable for facial manipulation deepfakes only as it focused too much on local facial features meaning it did not perform well on face-swapping deepfakes.

Vamsi et al. (2022) created an RNN model called ConvNets, consisting of ResNet and LSTM [58]. The paper [58] noted, due to the architecture of the model, less pre-processing was required and images were handled in a way that made the model fast to implement and easy to use. It [58] also claimed high performance when compared to other CNN models using the same datasets.

Xu et al. (2021) through extensive experimentation with multiple variations of detection techniques developed an SCNN which gave robust and verifiable results. The authors asserted that the detection technique [41] was more robust than previous models and potentially could be integrated with other backbone networks to enhance performance.

Raskar et al. (2021) proposed YOLO which claimed to be extremely accurate and efficient in terms of computational time in object-based video forgery detection. This technique [19] detected complex copy-move video deepfakes with 99% accuracy. This technique was best suited to copy-move attacks including scaling, rotation and flipping [19]. However, the researchers recognised that the technique cannot efficiently detect inter-frame video deepfakes.

Bakas et al. (2021) created a technique [88] that claimed to work well with compressed videos. However, the performance dropped when one or more complete group of pictures (GOP) was removed from the video sequence.

Sun et al. (2021) developed a technique [89] that was efficient when applied to data that had been forged using deepfake technologies. However, while performing well with data faked with deepfake techniques, the model performed poorly when given neutral data.

3.2.3. Audio deepfakes

Audio manipulation was discussed as one of the latest evolutions of deepfake exploitation in several papers [12,30,42,90,91], with synthetic voice generation (text-to-speech) or voice conversion algorithms producing almost undetectable deepfake media. Martin (2022) stated, '*Industry research firm Gartner predicts that within two years, 20% of all successful account takeover attacks will use deepfakes and synthetic voice augmentation*' (p 14) [91]. Zhou et al. (2021) suggested that incorporating audio and video streams in deepfake detection models would increase the performance of the model and potentially identify deepfake content in one or both streams [59]. Masood et al. (2022) noted audio deepfakes had received less attention than other detection areas, even though synthetic voices were a threat to

voice-controlled and voice-authenticated systems, presenting a unique challenge going forward [12]

Khochare et al. (2021) compared two approaches for audio deepfake detection [42]. A feature-based classification approach utilising machine learning algorithms, which converted an audio file sample into a collection of features that were used to calculate whether it was real or fake. Then an image-based classification approach utilising deep-learning algorithms, where melspectrograms produced from a sample and examined by several CNN models determined if the audio was real or fake. The outcome of this comparison [42] was that image-based classification (with deep learning algorithms) outperformed feature-based classification using machine learning. However, they noted [42] raw audio classifier models would improve accuracy and overhead, as there would be no pre-processing (manipulation) of the audio files needed.

3.2.4. Datasets

Hazirbas et al. (2021) suggested the top 5 deepfake detection tools perform poorly on some specific groups as they did not generalise to all people [3]. This paper [3] went on to introduce a new dataset, called the Casual Conversations dataset. It was made up of 45,000 videos using 3,011 subjects from various age, gender and skin tone groups – which were annotated with additional classifiers to enhance the accuracy of the AI models trained on it. The conclusion of this study [3] was the deepfake detection techniques trained on a traditional dataset had a strong biased towards lighter skin tones, and gender classification was more successful for older age groups (+45 years old).

Nadimpalli et al. (2022), after examining the gender labelling on current datasets, concluded when used to train deepfake detection tools they produced biased results, skewed to successfully identify male-based deepfakes more often than female-based ones [85]. That is, female-based content had a higher false match and higher non-match rate than male-based content.

Xu et al. (2022) went further, investigating 41 distinct classifiers – demographic (age, gender, ethnicity) and non-demographic (hair, skin, accessories, etc), across five deepfake datasets to determine if the datasets had diversity. The paper [4] also analysed whether the deepfake detection tools themselves had AI bias which could lead to generalizability, fairness and security issues. The analysis [4] concluded both, the investigated datasets and deepfake detection models, exhibited bias with regard to demographic and non-demographic classifiers. That is, the datasets showed a lack of diversity across the attributes of their content and the detection models themselves had a strong bias toward certain attributes.



Almutairi and Elgibreen (2022) identified a lack of datasets available for audio deepfake detection model training, particularly non-English-based datasets [11]. They also questioned the impact of accents (even in English-based datasets) on the accuracy of audio deepfake detection [11].

To conclude this section, we would like to mention several papers [15,92], published in 2022, which presented medical imagery deepfakes as an emerging threat. Solaiyappan et al. (2022) discussed the need for more work in detecting deepfakes generated by GANs rather than tampering methods, such as copy-move or image-splicing methods [15]. However, Arora et al. (2022) recognised the threat of deepfake images in medical imagery but also suggested that deepfake generation in this area was an opportunity to build realistic, effective training material that would not be impacted by restrictions placed on personal information management [92].

4. Observations

In this section, we have addressed the research questions mentioned above, which we believed were the most relevant and important in this area of research. The section discusses the deepfake detection techniques identified through this SLR, including metrics gathered on the effectiveness of these techniques and the datasets used by them. Advances in deepfake detection technologies were noted and current challenges and limitations were identified. Finally, potential trends in future deepfake detection techniques have been discussed.

4.1. Deepfake detection techniques

4.1.1. RQ1 what are the current deepfake detection techniques?

We identified 48 deepfake detection models which met the objectives of this SLR. These covered deepfake detection on all types of media – image, video, audio, text and a combination of these. Figure 4 shows the media types discussed in the documents reviewed. From Table 4, which details the deepfake detection models and when they were published, we can see there were 28 noteworthy deepfake detection techniques proposed in 2021 and 23 in 2022 – bearing in mind the search for this SLR was conducted up to and including August 2022.

Table 6 details the acronyms used in Table 4 – for example: CNN++ represents convolution neural network extended with other modules/components

4.1.2. RQ1A what is the effectiveness of these techniques?

The metrics shown in Table 4, against the identified deepfake detection techniques, were either Area Under the ROC Curve (AUC) and/or accuracy (ACC). All models reported high performances in these metrics when trained, validated, and

tested on the same dataset. However, many experienced a significant performance drop when trained on one dataset and validated and tested on another.

Another impact on the performance of these models was the quality of the data in the dataset used [57] – that is, if the dataset contained high-quality data, the performance metric was very high, but when the dataset consisted of low-quality data (for example compressed images), the same model experienced a drop in performance. Many models [32,41,51] showed comparison metrics when trained, validated, and tested on several datasets to evaluate performance across different quality data and this also contributed to the variations in the metrics shown in [Table 4](#). It was noted [8] that until standardised performance evaluation criteria for deepfake detection models had been established, the metrics recorded in the research of individual models had little value, as they could not be used as a viable comparison metric.

These factors indicated many of the models were not generalisable or robust when presented with challenging or unknown conditions, meaning they were not equipped to handle real-world applications [57].

4.1.3. Datasets

There were 30 datasets used across the deepfake detection techniques identified by this SLR, which have been detailed in [Table 5](#). However, many of the deepfake detection models used one or more of the following datasets:

FaceForensics++ - contains 1,000 real and 4000 manipulated videos sourced from 977 YouTube videos. The fake videos come in two compression versions – high quality (c23) and low quality (c40) [54]. This is a first-generation dataset that was produced in 2019. It was the most used dataset identified for this SLR, being used in 51% of the models listed.

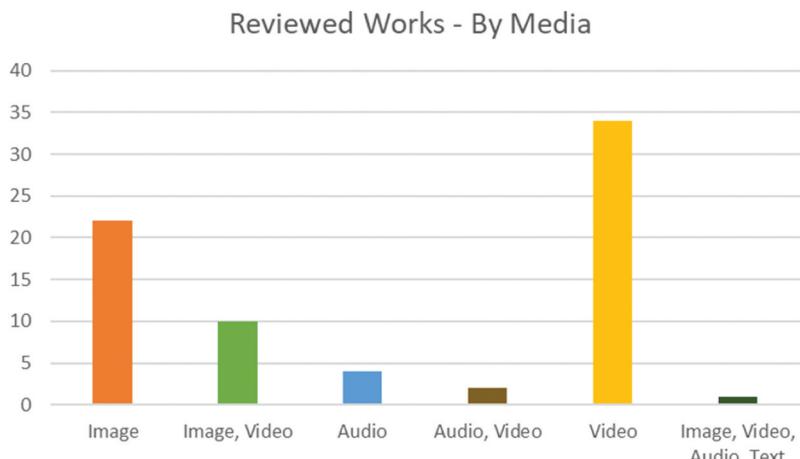


Figure 4. Reviewed documents classified by the discussed media.

**Table 6.** Datasets discussed in the reviewed documents

Name		Data Type	Approx Size	Ratio Real:Fake		Gen	Year	% Usage	Source
				Gen	Year				
VidTIMIT Audio-Video Dataset	VidTIMIT	Video/Audio	350	1:0	1	2001	2	[63,64]	
ImageNet		Image	14,000,000 + 244,527	1:0	1	2009	4	[65]	
Lung Image Database Consortium image collection	LIDC-IDRI	Medical imagery	244,527	1:0	1	2012	1	[66]	
CelebA		Image	202,599	1:0	1	2015	6	[67]	
Large-scale Scene Understanding	LSUN	Image	1,000,000	1:0	1	2015	1	[68]	
FaceScrub		Image	106,863	1:0	1	2016	1	[69]	
Deepfakes Image Dataset	DFID	Image	19,809	1:0.6	1	2018	1	[47,64]	
Unknown-attack Deepfake Video	UADFV	Video	98	1:1	1	2018	5	[70]	
DeepfakeTIMIT	DFTIMIT	Video	960	0:1	1	2018	11	[70]	
VGGFace2		Image	3,310,000	1:0	1	2018	1	[71]	
CASIA-WebFace	CASIA	Image	494,414		1	2018	4	[72]	
FaceForensics++	FF++	Video	5,000	1:4	1	2019	51	[70,73]	
100K-Face		Image	100,000	0:1	1	2019	1	[60]	
ASV spoof 2019 - LA subset	ASVS19-LA	Audio	121,461	1:8.7	1	2019	1	[74]	
UADFV	UADFV	Video	98	1:1	1	2018	5	[70]	
DeepFakeDetection	DFD	Video	3,431	1:8.5	2	2019	23	[12,70,73]	
Real and Fake Face-Detection	RFFD	Images	2,041	1:0.8	2	2019	4	[75]	
CT-GAN Dataset	CT-GAN	Medical imagery	100	0:1	2	2019	1	[76]	
Fake or Real Deepfake Detection Challenge Dataset	FoR DFDC	Audio/Video	195,000 / 128,154		2	2019 / 2020	6 / 20	[77] / [12,70]	
CelebDF	CDF	Video	6,229	1:1.95	2	2020	33	[34,70,78]	
DeeperForensics-1.0	DFor-1.0	Video	60,000	5:1	2	2020	2	[73]	
Diverse Fake Face Dataset	DFFD	Image/Video	2,600,000	1:1.8	2	2020	1	[79]	
Hybrid Fake Face	HFF	Image	155,000	1:1.5	2	2021	1	[80]	
Wild DeepFake	WDF	Video	7,314	1:0.9	2	2021	1	[34,81]	
FaceForensics in the Wild	FFIW-10K	Video	10,000	0:1	3	2021	1	[82]	
Casual Conversations	CC	Video	45,186	1:0	3	2021	1	[83]	
Korean Deepfake Detection Dataset	KoDF	Images/Video	237,942	1:2.6	3	2021	1	[84]	
Gender Balanced DeepFake Dataset	GBDF	Video	10,000	1:4	3	2022	1	[85]	

CelebDF – contains 590 real videos sourced from YouTube that have been manipulated into 5,639 deepfake videos. This dataset contains higher-quality videos than most, as an attempt has been made to remove visible source artifacts. It was used in 34% of the models listed.

Deepfake Detection – was created by Google and has 3,431 videos, with a ratio of 1 real to 8.5 fake videos in the dataset. It was used in 23% of the models listed.

Deepfake Detection Challenge Dataset - contains 128,154 videos sourced from 3,426 paid actors – the video breakdown is 104,500 fake and 23,654 real videos. This dataset consists of videos in different lighting conditions (indoor/outdoor) that were taken with high-resolution cameras. It was used in 19% of the models listed.

4.2. Deepfake detection advancement

4.2.1. RQ2 what advances have occurred since the previous SLRs?

In an effort to create robust and generalisable deepfake detection models [31,50], many papers proposed the combination of traditional techniques (CNN, DNN, LTSM) with other modules to produce ensembled and multi-attentional architectures. This evolution can be seen via the published dates of the documents reviewed for this SLR. Stand-alone CNN models and some multi-attentional architectures were proposed in papers published in 2021, but only multi-attentional architectures were discussed in the papers published in 2022.

Audio deepfakes were recognised in the literature reviews/surveys reviewed for this SLR, but mostly they focused on audio-visual deepfakes, rather than audio alone. Almutairi and Elgibreen (2022) [11] and Masood et al. (2022) [12] were the first to include audio deepfakes into their literature reviews and both of these papers were published towards the end of our search date (May and June 2022 respectively). This indicates audio deepfakes are a new area of interest. This was also the case with medical imagery deepfakes. The only papers [15,92] found talking about medical imagery deepfakes were published in June and July 2022, which again suggested it was an emerging area of research.

Traditionally, deepfake detection models focused on only one type of facial manipulation – that is, face swapping or feature manipulation of the primary face in the image or video. However, recent models [39,61] were used to assess the whole image for manipulation, including images with no faces in them, which suggested another emerging area of deepfake detection. One model [44] went so far as successfully identifying multiple deepfake faces in images over frame by frame examination of videos.

4.3. Challenges and limitations

4.3.1. RQ3 what are the current challenges involved in deepfake detection?

Some of the main challenges identified with deepfake detection technologies were:

- The quality, fairness, and trust of deepfake datasets (biased and imbalanced data) [6,7,9,12,90];
- Robustness of the deepfake detection techniques against unknown attack types [7];
- Temporal aggregation [7,12]; and
- Social media laundering [12]

This survey [8] noted a need for competitive baselines to evaluate deepfake detection techniques so the true performance could be assessed. Zhang (2022)



agreed, indicating the deepfake detection techniques up to 2020, reviewed in this survey [6], were not robust and there were many challenges when assessing generation and detection models, including a lack of benchmark test methods and quality datasets.

In addition, deepfake detection techniques, due to a lack of generalisation or their narrow field of specialisation, offered poor practical applications. An example of this was discussed by Guo et al. (2021) where an adaptive manipulated traces extraction network (AMTEN) was combined with a CNN to enhance the efficiency of the model, but was found to suffer from a lack of generalisability to make it useful in most practical settings [31]. Similarly, the 3D CNN model proposed by Liu et al. (2021) failed to be of any practical use due to the consumption of resources it required [24].

Concern over the viability of datasets was discussed in many papers [3,4,6–9,12,85,90] with several areas identified as needing further research and improvement.

Xu et al. (2022) proposed that most research in this area had focused on developing deepfake detection solutions utilising the available datasets, but generalizability issues had arisen from training these solutions on less diverse datasets [4]. Furthermore, they found that the detection solutions developed inappropriate bias due to the presence of certain attributes in the datasets (for example, being male or black or having a big nose). This was also noted by Masood et al. [12] who suggested there was an urgent need to address data fairness in detection techniques.

Hazirbas et al. (2021) produced a new, 'balanced' dataset using subjects from various age, gender and skin tone groups and annotated with additional classifiers to enhance the accuracy of deepfake detection techniques trained on it [3]. The age and gender classifiers in the dataset were annotated by the subjects themselves, which was believed would produce an unbiased view of these classifiers. The skin tone classifier was annotated using the Fitzpatrick skin phototype scale. This study [3] found the deepfake detection techniques examined had a strong biased towards lighter skin tones. This was concluded because they mostly failed on dark skin tones. Also, gender classifications were more successful for older age groups (+45 years old) across all skin tones.

As mentioned in section 4.1, the SLR was unable to identify any models that conclusively showed a sustained, high level of capability across multiple data qualities, that is, high and low quality data, particularly when presented with challenging or unknown conditions like light variations or biometric features (such as glasses or heavy eyebrows). Nor could any models detect deepfakes from multiple forms of manipulation (that is, face swapping and facial manipulation) especially when the manipulation was unknown to the model. This meant current deepfake detection models were not generalisable or robust when presented with challenging or unknown conditions, and were not equipped to handle real-world applications [57].

Temporal aggregation, the temporal consistency between frames in deepfake video content – that is, real and fake content appearing in sequential intervals that would require extra effort to identify and classify, was discussed in [7,12] as an area as yet to be investigated.

Images appearing on social media platforms have been heavily compressed during the upload process in order to limit the size and bandwidth needed. These manipulations, called social media laundering [12], remove a lot of the artifacts that have been added to an image or video during the deepfake generation process. This means the evidence that would normally aid deepfake detection is no longer available and makes detecting deepfake content from social media platforms challenging.

We had one final comment on deepfake detection challenges related to the potential for deepfake usage in legal cases. Media evidence is already prominent in the courtroom; and it would be surprising if a courtroom trial did not have any picture, video or audio evidence admitted either on behalf of the defence or prosecution. In 2016, the Bureau of Justice Assistance, U.S. Department of Justice estimated that 80% of crimes had video evidence associated with them. It seems logical then that deepfake videos, audio and imagery, would begin to be used by either side in a court of law, whether it be intentional or accidental. Pfefferkorn (2016) talks about a concept known as the ‘reverse CSI effect’ in which those involved in the legal process, jurors for example, begin to doubt it is ever possible to determine what is real [93]. Pfefferkorn believes that the courts already have the tools required to be able to weed out inauthentic evidence, although it may involve extra effort [93]. Venema (2020) believes that the courts are at risk of falling victim to the ‘deepfake defence’ if they haven’t already [94]. The future challenge lies in keeping up with the pace of deepfakes and being able to accurately detect them.

4.4. Looking ahead

4.4.1. RQ4 can any future trends be identified, and if so, what are they?

The majority of documents reviewed for this SLR related to image, video and image/video deepfake detection (90%) with the remaining discussing techniques for audio, audio/video or a combination of all media. This showed little research had been directed at deepfakes other than image and video – which suggests the main driver for deepfake detection has been social media conglomerates, which have a massive footprint in these media.

AI-generated voices have now become so ‘life-like’ it is difficult to distinguish them from real audio. The original need for audio deepfakes used in digital assistant devices has resulted in tools becoming publicly available that can produce undetectable audio deepfakes from a very small source sample. These are being touted as a growing threat to organisations. That is, real-time audio deepfake has the potential to be the new social engineering tool for



cyber-attacks. As email phishing attacks are mitigated with awareness training, spear phishing attacks using audio impersonation are becoming the new age of deepfake threats.

Identity compromise via voice activation systems is also a serious and growing cyber security threat. The use of voice biometrics is expected to increase in the next few years, as organisations move to passwordless authentication, and audio deepfakes present a challenge that needs to be overcome sooner rather than later.

There is a general lack of datasets for audio deepfake detection [11,12], particularly non-English speaking datasets, and this means a lot of work is needed in this area. Also, as yet, no research has been undertaken to evaluate the impact of accents on successful deepfake detection [11]. As already mentioned, audio deepfakes are a growing threat and immediate attention in these areas is required if deepfake detection models are to effectively manage the threat landscape.

More work also needs to be done on annotating datasets and ensuring detection models remain unbiased through and after training. This will go some way to reducing the bias found in deepfake detection techniques but work still needs to be done to ensure the techniques themselves are free from bias.

5. Conclusion

This paper comprehensively discussed methods used in the detection of deepfakes and determined current techniques. As part of the review process, this paper categorised papers into two categories, a) Previous Literature Reviews and Surveys and b) modern deepfake detection methods. The SLR reviewed 8 literature review/survey documents with 7 of those published in 2022. It also reviewed 75 documents from category b) modern deepfake detection methods, ranging from 2021 to August 2022, which identified developing trends and future research areas. The majority (90%) of the papers reviewed focused on video and image or a combination of both.

This paper identified 48 noteworthy detection models, across all types of media including image, video, audio, text, and a combination of all. Of these 48, 23 detection techniques were from research and proposals from 2022 leading up to August 2022, with a further 28 proposed in 2021.

Advancements in deepfake detection have proposed a combination of traditional techniques (CNN, DNN and LSTMs) and other modules to strengthen these techniques, producing ensembled and multi-attentional architectures. Recent papers written have adopted this strategy with many suggesting multi-attentional architectures in 2021 papers and papers written in 2022 only suggesting multi-attentional architectures.

Models discussed within this paper were judged against Area Under the ROC Curve (AUC) and/or accuracy (ACC); a growing method of reporting the effectiveness of deepfake techniques. It is worth noting that all models this report discovered, reported high performance in these metrics when trained, validated, and tested on the same dataset. However, when these models were trained and validated on differing datasets there were noticeable performance drops. As such this report concluded models were not robust or adaptable when presented with challenging or unknown conditions.

As part of this review, information was discovered that suggested deepfake detection tools perform poorly on various cultural and ethnic groups. It was concluded in studies reviewed that deepfake detection techniques trained on a standard dataset had a strong bias towards lighter skin tone and gender classification and had greater success with people aged above 45 years. As such, challenges identified within this paper include the quality, fairness, and trust of the deepfake datasets. Also, the robustness of these datasets when being used against unknown attack types was a challenge for detection. Many techniques reviewed have a large reliance on specific datasets, if these datasets do not represent a broad section of society, including cultural or ethnic representation, their effectiveness is severely reduced. As such, the viability of datasets is a large challenge moving forward.

A recommendation this paper suggests is the need to create a standardised approach to dataset usage and a uniform rating system where techniques can be validated against each other, reducing the ability to select the best or most desirable datasets. Currently, deepfake detection technique creators and owners can elevate their models using targeted datasets known to produce better results for that model type, making it appear superior.

There are current shortfalls within the research field of deepfake detection outside of the current popular image and video types. As mentioned, this paper found the majority of research reviewed related to one or both types. As a result, more research is required in additional fields including audio-only deepfake detection.

Future trends identified in this paper include audio deepfakes as the latest evolution of deepfake exploitation, utilising synthetic voice generation or voice conversation algorithms that output undetectable deepfake audio. As previously mentioned, Gartner has predicted, within two years 20% of all successful account takeovers will be deepfakes and synthetic voice augmentation [91]. Leading to audio deepfakes being used in deception methods including spear phishing scams. This prediction highlights a future trend in deepfake generation and the requirement for additional research in audio deepfake detection. In addition to audio deepfakes, several papers also mentioned concerns surrounding medical imagery deepfakes. As such, additional research is required in detection methods of these deepfakes types.



This research has highlighted a positive aspect of deep fake material. Such as deep fake imagery can be used for good during medical training providing realistic material.

With the evolution of technology and the low-cost barriers to entry, the advancement of deep fakes will progress with a rapid trajectory. As this evolves, future challenges in detection techniques will be required to adapt at the same level or greater to those technological advancements.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Mohiuddin Ahmed  <http://orcid.org/0000-0002-4559-4768>

References

- [1] Curry L, Nembhard I, Bradley E. Qualitative and mixed methods provide unique contributions to outcomes research. *Circulation*. 2009;119(10):1442–1452.
- [2] Abdulreda AS, Obaid AJ. A landscape view of deepfake techniques and detection methods. *Int J Nonlinear Anal Appl*. 2022;13(1):745–755.
- [3] Hazirbas C, Bitton J, Dolhansky B. Towards Measuring Fairness in AI: the Casual Conversations Database. *IEEE Transactions on Biometrics, Behavior, and Identity Science*. 2021.
- [4] Xu Y, Terhörst P, Raja K, et al. A Comprehensive Analysis of AI Biases in DeepFake Detection with Massively Annotated Databases. arXiv preprint arXiv:2208.05845, 2022.
- [5] Weerawardana M, Fernando T, Deepfakes detection methods: a literature survey, in 2021 10th International Conference on Information and Automation for Sustainability (ICIAfs). 2021, IEEE Access: Negombo, Sri Lanka. p. 76–81.
- [6] Zhang T. Deepfake generation and detection, a survey. *Multimedia Tools Appl*. 2022;81(5):6259–6276.
- [7] Malik A, Kuribayashi M, Abdullahi SM, et al. DeepFake detection for human face images and videos: a survey. *IEEE Access*. 2022;10:18757–18775.
- [8] Juefei-Xu F, Wang R, Huang Y, et al. Counteracting malicious deepfakes: survey, battleground, and horizon. *Int J Comput Vis*. 2022;130(7):1678–1734. DOI:[10.1007/s11263-022-01606-8](https://doi.org/10.1007/s11263-022-01606-8)
- [9] Rana MS, Nobi MN, Murali B, et al. Deepfake detection: a systematic literature review. *IEEE Access*. 2022;10:25494–25513.
- [10] Celebi N, Liu Q, Karatoprak M, A survey of deep fake detection for trial courts, in 9th International Conference on Artificial Intelligence and Applications (AIAPP 2022). 2022, ResearchGate: Vancouver, Canada. p. 227–238.
- [11] Almutairi Z, Elgibreem H. A review of modern audio deepfake detection methods: challenges and future directions. *Algorithms*. 2022;15(5):155.
- [12] Masood M, Nawaz M, Malik KM, et al. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Appl Intell*. 2022;53(4):3974–4026. DOI:[10.1007/s10489-022-03766-z](https://doi.org/10.1007/s10489-022-03766-z)

- [13] Yang J, Xiao S, Li A, et al. MSTA-Net: forgery detection by generating manipulation trace based on multi-scale self-texture attention. *IEEE Trans Circuits Syst Video Technol*. **2022**;32(7):4854–4866. DOI:[10.1109/TCSVT.2021.3133859](https://doi.org/10.1109/TCSVT.2021.3133859)
- [14] Mikhaltsov VE, Semenova ZV, Stepanova EA, Effectiveness analysing of modification of training data set for voice forgery detection system. *Journal of Physics: Conference Series*. **2022**;2182(1):012096.
- [15] Solaiyappan S, Wen Y. Machine learning based medical image deepfake detection: a comparative study. *Mach Learn Appl*. **2022**;8:100298.
- [16] Fang S, Wang S, Ye R. DeepFake video detection through facial sparse optical flow based light cnN. *Journal of Physics: Conference Series*. **2022**;2224(1):012014.
- [17] Fu T, Xia M, Yang G. Detecting GAN-generated face images via hybrid texture and sensor noise based features. *Multimedia Tools Appl*. **2022**;81(18):26345–26359.
- [18] Lu Y, Liu Y, Fei J, et al. Channel-wise spatiotemporal aggregation technology for face video forensics. *Secur Commun Networks*. **2021**;2021:1–13.
- [19] Raskar PS, Shah SK. Real time object-based video forgery detection using YOLO (V2). *Forensic Science International*. **2021**;327.
- [20] Ramachandran S, Nadimpalli AV, Rattani A, *An experimental evaluation on deepfake detection using deep face recognition*, in *2021 International Carnahan Conference on Security Technology (ICCST)*. **2021**, IEEE: Hatfield, United Kingdom. p. 1–6.
- [21] Shad HS, Rizvee MM, Roza NT, et al. Comparative analysis of deepfake image detection method using convolutional neural network. *Comput Intell Neurosci*. **2021**;2021:3111676.
- [22] Luo Y, Ye F, Weng B, et al. A novel defensive strategy for facial manipulation detection combining bilateral filtering and joint adversarial training. *Secur Commun Networks*. **2021**;2021:1–10.
- [23] Hsu H-W, Ding J-J, *Deepfake algorithm using multiple noise modalities with two-branch prediction network*, in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. **2021**, APSIPA: Tokyo, Japan. p. 1662–1669.
- [24] Liu J, Zhu K, Lu W, et al. A lightweight 3D convolutional neural network for deepfake detection. *Int J Intell Syst*. **2021**;36(9):4990–5004. DOI:[10.1002/int.22499](https://doi.org/10.1002/int.22499)
- [25] Nguyen XH, Tran TS, Le VT, et al. Learning spatio-temporal features to detect manipulated facial videos created by the deepfake techniques. *Forensic Sci Int: Digital Invest*. **2021**;36:301108.
- [26] Lee S, Tariq S, Shin Y, et al. Detecting handcrafted facial image manipulations and GAN-generated facial images using shallow-FakeFaceNet. *Appl Soft Comput*. **2021**;105:107256.
- [27] Kohli A, Gupta A. *Detecting DeepFake, FaceSwap and face2face facial forgeries using frequency CNN*. multimedia tools and applications. *Int J*. **2021**;80(12):18461–18478.
- [28] Luo Z, Kamata S-I, Sun Z, *Transformer and node-compressed dnn based dual-path system for manipulated face detection*, in *2021 IEEE International Conference on Image Processing (ICIP)*. **2021**, Anchorage, AK, USA. p. 3882–3886.
- [29] Ru Y, Zhou W, Liu Y, Sun J, Li Q. Bita-net: bi-temporal attention network for facial video forgery detection. *2021 IEEE International Joint Conference on Biometrics (IJCB)*, Shenzhen, China. **2021**;1–8. doi:[10.1109/IJCB52358.2021.9484408](https://doi.org/10.1109/IJCB52358.2021.9484408)
- [30] Zhang M. *FST-Net: exploiting Frequency Spatial Temporal Information for Low-Quality Fake Video Detection*. in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*. **2021**. Washington, DC, USA: IEEE.
- [31] Guo Z. Fake face detection via adaptive manipulation traces extraction network. *Computer Vision and Image Understanding*. **2021**;204.

- [32] Chen B, Tan S, Alazab M. FeatureTransfer: unsupervised domain adaptation for cross-domain deepfake detection. *Secur Commun Networks*. 2021;2021:1–8.
- [33] Ananthi M. A secure model on Advanced Fake Image-Feature Network (AFIFN) based on deep learning for image forgery detection. *Pattern Recognit Lett*. 2021;152:260–266.
- [34] Zhou Y. *Face Forgery Detection Based on Segmentation Network*, in 2021 IEEE International Conference on Image Processing (ICIP). 2021, IEEE: Anchorage, AK, USA. p. 3597–3601.
- [35] Sun Z. *Improving the efficiency and robustness of deepfakes detection through precise geometric features*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, Nashville, TN, USA. p. 3609–3618.
- [36] Shichao G, Ming X, Gaobo Y. *Gaobodual-tree complex wavelet transform-based direction correlation for face forgery detection*. *Secur Commun Networks*. 2021;2021:1–10.
- [37] Gong D, Kumar YJ, Goh OS, et al. DeepfakeNet, an efficient deepfake detection method. *Int J Adv Comput Sci Appl*. 2021;12(6).
- [38] Jin X, Ye D, Chen C. Countering spoof: towards detecting deepfake with multidimensional biological signals. *Secur Commun Networks*. 2021;2021:1–8.
- [39] Singh B, Sharma DK. SiteForge: detecting and localizing forged images on microblogging platforms using deep convolutional neural network. *Comput Ind Eng*. 2021;162:107733.
- [40] Yang J, Xiao S, Li A, et al. Detecting fake images by identifying potential texture difference. *Future Gener Comput Syst*. 2021;125:127–135.
- [41] Xu Z, Liu J, Lu W, et al. Detecting facial manipulated videos based on set convolutional neural networks. *J Vis Commun Image Represent*. 2021;77:103119.
- [42] Khochare J, Joshi C, Yenarkar B, et al. A deep learning framework for audio deepfake detection. *Arab J Sci Eng*. 2021;47(3):3447–3458. DOI:10.1007/s13369-021-06297-w
- [43] Wang Z, Yiwen G, Zuo W. Deepfake forensics via an adversarial game. *IEEE Trans Image Process*. 2022;31:3541–3552.
- [44] Deng L, Suo H, Li D. Deepfake video detection based on efficientnet-v2 network. *Comput Intell Neurosci*. 2022;2022:3441549.
- [45] Wang G, Jiang Q, Jin X, et al. MC-LCR: multimodal contrastive classification by locally correlated representations for effective face forgery detection. *Knowledge-Based Syst*. 2022;250:109114.
- [46] Wang T, Liu M, Cao W, et al. *Deepfake noise investigation and detection*. forensic science international: digital investigation. *Forensic Sci Int: Digital Invest*. 2022;42:301395.
- [47] Ganguly S, Ganguly A, Mohiuddin S, et al. ViXNet: vision transformer with xception network for deepfakes based video and image forgery detection. *Expert Syst Appl*. 2022;210:118423.
- [48] Cozzolino D. *ID-Reveal: identity-aware DeepFake Video Detection*, in 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021, IEEE: Montreal, QC, Canada. p. 15088–15097.
- [49] Pu W, Hu J, Wang X, et al. Learning a deep dual-level network for robust DeepFake detection. *Pattern Recogn*. 2022;130:108832.
- [50] Taeb M, Chi H. Comparison of deepfake detection techniques through deep learning. *J Cybersecur Privacy*. 2022;2(1):89–106.
- [51] Wang J, Sun Y, Tang J. LiSiam: localization invariance siamese network for deepfake detection. *IEEE Trans Inf Forensics Secur*. 2022;17:2425–2436.
- [52] Kolagati S, Priyadarshini T, Mary Anita Rajam V. Exposing deepfakes using a deep multilayer perceptron – convolutional neural network model. *Int J Inf Manage Data Insights*. 2022;2(1):100054.

- [53] Hu J, Liao X, Wang W, et al. Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. *IEEE Trans Circuits Syst Video Technol.* **2022**;32(3):1089–1102. DOI:10.1109/TCSVT.2021.3074259
- [54] Yu M, Ju S, Zhang J, et al. Patch-DFD: patch-based end-to-end DeepFake discriminator. *Neurocomputing.* **2022**;501:583–595.
- [55] Chen B, Li T, Ding W. Detecting deepfake videos based on spatiotemporal attention and convolutional LSTM. *Inf Sci.* **2022**;601:58–70.
- [56] Saif S. Generalized Deepfake Video Detection Through Time-Distribution and Metric Learning. *IT Prof.* **2022**;24(2):38–44.
- [57] Chamot F, Geraerts Z, Haasdijk E. Deepfake forensics: cross-manipulation robustness of feedforward- and recurrent convolutional forgery detection methods. *Forensic Sci Int: Digital Invest.* **2022**;40:301374.
- [58] Vamsi VVNS. Deepfake detection in digital media forensics. *Global Transitions Proceedings.* **2022**;3(1):74–79.
- [59] Zhou Y, Lim S-N, Q.C.C.O.O. Ieee/Cvf International Conference on Computer Vision Montreal, *Joint Audio-Visual Deepfake Detection*, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. **2021**, Montreal, Canada: IEEE. p. 14780–14789.
- [60] Prototypr. *AI-Generated Faces: Free Resource of 100K Faces Without Copyright*. **2019** 18 September, 2019 [cited 08/10/2022]; Available from: <https://blog.prototypr.io/generated-photos-free-resource-of-100k-diverse-faces-generated-by-ai-2144a8615d1f>.
- [61] Liu X. PSCC-Net: progressive Spatio-Channel Correlation Network for Image Manipulation Detection and Localization. *IEEE Transactions on Circuits and Systems for Video Technology.* **2022**;1.
- [62] Coccominini DA, Messina N, Gennaro C, et al. Combining EfficientNet and vision transformers for video deepfake detection. In: Sclaroff S, Distante C, Leo M, Farinella GM Tombari F, editors. *Image analysis and processing – iciap 2022*. iciap 2022. lecture notes in computer science. Vol. 13233. Cham: Springer; **2022**. doi:10.1007/978-3-031-06433-3_19.
- [63] Babbar A *The VidTIMIT Audio-Video Dataset*. **2016** [cited 2022 18 September]; Available from: <https://www.kaggle.com/datasets/akshay4/speakerrecognition> .
- [64] Afchar D. *MesoNet: a Compact Facial Video Forgery Detection Network*, in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. **2018**, Hong Kong: IEEE. p. 1–7.
- [65] University S *ImageNet*. **2021** [cited 2022 18 September]; Available from: <https://www.image-net.org/>.
- [66] Bilello E *LIDC-IDRI*. **2022** 16 September, 2022 [cited 2022 20 September]; Available from: <https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI>.
- [67] Li J *CelebFaces Attributes (CelebA) Dataset*. **2018** [cited 2022 18 September]; Available from: <https://www.kaggle.com/datasets/jessicali9530/celeba-dataset> .
- [68] Yu F. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. Vol. 9. **2015**. 10.48550/arXiv.1506.03365.
- [69] Group VI *FaceScrub*. **2021** 23 March 2021 [cited 2022 19 September]; Available from: <http://vintage.winklerbros.net/facescrub.html>.
- [70] Dolhansky B. *The DeepFake Detection Challenge (DFDC) Dataset*. arXiv preprint arXiv:2006.07397, **2020**.
- [71] Xie W *VGGFace2 dataset for face recognition*. **2020** 18 February, 2020 [cited 2022 19 September]; Available from: https://github.com/ox-vgg/vgg_face2.
- [72] Code PW *CASIA-Webface*. [cited 2022 19 September]; Available from: <https://papers-withcode.com/dataset/casia-webface>.



- [73] Rathgeb C. Handbook of Digital Face Manipulation and Detection. In: Kang SB, editor. *Advances in Computer Vision and Pattern Recognition*. 1. Springer International Publishing. IX; 2022. p. 487.
- [74] Zhang Y, Jiang F, Duan Z. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Process Lett*. 2021;28:937–941.
- [75] University Y *Real and Fake Face Detection*. 2019 [cited 2022 19 September]; Available from: <https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection>.
- [76] Mirsky Y. {CT-GAN}: malicious Tampering of 3D Medical Imagery using Deep Learning. in *28th USENIX Security Symposium (USENIX Security 19)*. 2019, Santa Clara, CA.
- [77] Reimao R. *FoR: a Dataset for Synthetic Speech Detection*, in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. 2019, Timisoara, Romania: IEEE. p. 1–10.
- [78] GitHub I *Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics*. 2021 3 June 2021 [cited 2022 08 October]; Available from: <https://github.com/yuezunli/celeb-deepfakeforensics>.
- [79] Hao D. *DFFD: Diverse Fake Face Dataset*. 2020 June 2020 [cited 2022 8 October]; Available from: <http://cvlab.cse.msu.edu/dffd-dataset.html>.
- [80] GitHub I *Hybrid Fake Face Dataset*. 2021 17 March, 2021 [cited 2022 08 October]; Available from: <https://github.com/EricGzq/Hybrid-Fake-Face-Dataset>.
- [81] Zi B. *WildDeepfake: a challenging real-world dataset for deepfake detection*. in *Proceedings of the 28th ACM international conference on multimedia*. 2020, Seattle, WA, USA.
- [82] Zhou T *Face Forensics in the Wild*. 2021 4 October, 2021 [cited 2022 19 September]; Available from: <https://github.com/tfzhou/FFIW>.
- [83] MetaAI. *Casual Conversations Dataset*. 2021 [cited 2022 20 September]; Available from: <https://ai.facebook.com/datasets/casual-conversations-dataset/>.
- [84] Kwon P. *KoDF: a Large-scale Korean DeepFake Detection Dataset*, in *International Conference on Computer Vision (ICCV)*. 2021, IEEE: Montreal, Canada. p. 10724–10733.
- [85] Nadimpalli AV, Rattani A, *GBDF: gender balanced deepfake dataset towards fair deepfake detection*, in *International Conference on Pattern Recognition 2022*. 2022, ResearchGate: Montreal, Canada.
- [86] Zhang J, Ni J, Xie X, *DeepFake videos detection using self-supervised decoupling network*, in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. 2021 : Shenzhen, China. p. 1–6.
- [87] Chen B, Liu X, Zheng Y, et al. A robust gan-generated face detection method based on dual-color spaces and an improved xception. *IEEE Trans Circuits Syst Video Technol*. 2022;32(6):3527–3538. DOI:[10.1109/TCSVT.2021.3116679](https://doi.org/10.1109/TCSVT.2021.3116679)
- [88] Jamimamul Bakas RN, Bakshi S. Detection and localization of inter-frame forgeries in videos based on macroblock variation and motion vector analysis. *Comp Elec Eng*. 2021;89:106929.
- [89] Fang Sun NZ, Pan X, Song Z. Deepfake detection method based on cross-domain fusion. *Secur Commun Networks*. 2021;2021:11.
- [90] Muneef Z, Elgibreen H. A review of modern audio deepfake detection methods: challenges and future directions. *Algorithms*. 2022;15(5):19.
- [91] Martin EJ. Deepfakes: the latest trick of the tongue. *Speech Technology*. 2022;27(2):12–16.
- [92] Arora A, Arora A. Generative adversarial networks and synthetic patient data: current challenges and future perspectives. *Future Healthc J*. 2022;9(2):190–193.
- [93] Pfefferkorn R. “Deepfakes” in the courtroom. *Public Interest Law J*. 2020;29:245–275.
- [94] Venema AE, Geraadts ZJP. Digital forensics, deepfakes, and the legal process. *Sci Tech Lawyer*. 2020;16(4):14–17,23.