Deepfake Detection: A Systematic Literature Review

Article *in* IEEE Access · January 2022

DOI: 10.1109/ACCESS.2022.3154404

4 authors, including:

Md Shohel Rana
Florida Gulf Coast University
24 PUBLICATIONS **300** CITATIONS

SEE PROFILE

Andrew H. Sung
University of Southern Mississippi
27 PUBLICATIONS **362** CITATIONS

SEE PROFILE

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 25, 2022, accepted February 16, 2022, date of publication February 24, 2022, date of current version March 10, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3154404

Deepfake Detection: A Systematic Literature Review

MD SHOHEL RANA[1,2], (Member, IEEE), MOHAMMAD NUR NOBI[3], (Member, IEEE), BEDDHU MURALI[2], AND ANDREW H. SUNG[2], (Member, IEEE)

[1]Department of Computer Science, Northern Kentucky University, Highland Heights, KY 41099, USA
[2]School of Computing Sciences and Computer Engineering, The University of Southern Mississippi, Hattiesburg, MS 39401, USA
[3]Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX 78249, USA

Corresponding author: Md Shohel Rana (ranam2@nku.edu)

This work was supported in part by Northern Kentucky University and the University of Southern Mississippi.

ABSTRACT Over the last few decades, rapid progress in AI, machine learning, and deep learning has resulted in new techniques and various tools for manipulating multimedia. Though the technology has been mostly used in legitimate applications such as for entertainment and education, etc., malicious users have also exploited them for unlawful or nefarious purposes. For example, high-quality and realistic fake videos, images, or audios have been created to spread misinformation and propaganda, foment political discord and hate, or even harass and blackmail people. The manipulated, high-quality and realistic videos have become known recently as Deepfake. Various approaches have since been described in the literature to deal with the problems raised by Deepfake. To provide an updated overview of the research works in Deepfake detection, we conduct a systematic literature review (SLR) in this paper, summarizing 112 relevant articles from 2018 to 2020 that presented a variety of methodologies. We analyze them by grouping them into four different categories: deep learning-based techniques, classical machine learning-based methods, statistical techniques, and blockchain-based techniques. We also evaluate the performance of the detection capability of the various methods with respect to different datasets and conclude that the deep learning-based methods outperform other methods in Deepfake detection.

INDEX TERMS Deepfake detection, video or image manipulation, digital media forensics, systematic literature review.

I. INTRODUCTION

The notable advances in artificial neural network (ANN) based technologies play an essential role in tampering with multimedia content. For example, AI-enabled software tools like FaceApp [1], and FakeApp [2] have been used for realistic-looking face swapping in images and videos. This swapping mechanism allows anyone to alter the front look, hairstyle, gender, age, and other personal attributes. The propagation of these fake videos causes many anxieties and has become famous under the hood, Deepfake.

The term ''Deepfake'' is derived from ''Deep Learning (DL)'' and ''Fake,'' and it describes specific photo-realistic video or image contents created with DL's support. This word was named after an anonymous Reddit user in late 2017, who applied deep learning methods for replacing a person's face in pornographic videos using another person's face and created photo-realistic fake videos. To generate such counterfeit videos, two neural networks: (i) a generative network and (ii) a discriminative network with a FaceSwap technique were used [3], [4]. The generative network creates fake images using an encoder and a decoder. The discriminative network defines the authenticity of the newly generated images. The combination of these two networks is called Generative Adversarial Networks (GANs), proposed by Ian Goodfellow [5].

Based on a yearly report [6] in Deepfake, DL researchers made several related breakthroughs in generative modeling. For example, computer vision researchers proposed a method known as Face2Face [7] for facial re-enactment. This method transfers facial expressions from one person to a real digital 'avatar' in real-time. In 2017, researchers from UC Berkeley presented CycleGAN [8] to transform images and videos into different styles. Another group of

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

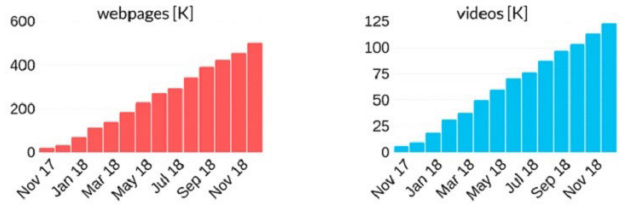

FIGURE 1. Left: Google search engine finds web pages containing "Deepfake" keyword (web pages count vs. month). Right: Google search engine finds web pages holding Deepfake related videos (web pages count vs. month).

scholars from the University of Washington proposed a method to synchronize the lip movement in video with a speech from another source [9]. Finally, in November 2017, the term ''Deepfake'' emerged for sharing porn videos, in which celebrities' faces were swapped with the original ones. In January 2018, a Deepfake creation service was launched by various websites based on some private sponsors. After a month, several websites, including Gfycat [10], Pornhub, and Twitter, banned these services. However, considering the threats and potential risks in privacy vulnerabilities, the study of Deepfake emerged super fast. Rossler *et al.* introduced a vast video dataset to train the media forensic and Deepfake detection tools called FaceForensic [11] in March 2018. After a month, researchers at Stanford University published a method, ''Deep video portraits'' [12] that enables photo-realistic re-animation of portrait videos. UC Berkeley researchers developed another approach [13] for transferring a person's body movements to another person in the video. NVIDIA introduced a style-based generator architecture for GANs [14] for synthetic image generation. According to [6] report, Google search engine could find multiple web pages that contain Deepfake related videos (see Figure 1). We found the following additional information from this report [6]:

- The top 10 pornographic platforms posted 1,790+ Deepfake videos, without concerning pornhub.com, which has removed 'Deepfakes' searches.
- Adult pages post 6,174 Deepfake videos with fake video content.
- 3 New platforms were devoted to distributing Deepfake pornography.
- In 2018, 902 articles were published in arXiv, including the keyword GAN either in titles or abstracts.
- 25 Papers published on this subject, including non-peer reviews, are investigated, and DARPA funded 12 of them.

Apart from Deepfake pornography, there are many other malicious or illegal uses of Deepfake, such as spreading misinformation, creating political instability, or various cybercrimes. To address such threats, the field of Deepfake detection has attracted considerable attention from academics and experts during the last few years, resulting in many Deepfake detection techniques. There are also some efforts on surveying selected literature focusing on either detection methods or performance analysis. However, a more comprehensive overview of this research area will be beneficial in serving the community of researchers and practitioners by providing summarized information about Deepfake in all aspects, including available datasets, which are noticeably missing in previous surveys. Toward that end, we present a systematic literature review (SLR) on Deepfake detection in this paper. We aim to describe and analyze common grounds and the diversity of approaches in current practices on Deepfake detection. Our contributions are summarized as follows.

- We perform a comprehensive survey on existing literature in the Deepfake domain. We report current tools, techniques, and datasets for Deepfake detection-related research by posing some research questions.
- We introduce a taxonomy that classifies Deepfake detection techniques in four categories with an overview of different categories and related features, which is novel and the first of its kind.
- We conduct an in-depth analysis of the primary studies' experimental evidence. Also, we evaluate the performance of various Deepfake detection methods using different measurement metrics.
- We highlight a few observations and deliver some guidelines on Deepfake detection that might help future research and practices in this spectrum.

The remainder of the paper is organized as follows: Section II presents the review procedure by defining interest research questions. In Section III, we thoroughly discuss the findings from different studies. Section IV summarizes the overall observations of the study, and we present the challenges and limitations in Section V. Finally, Section VI concludes the paper.

II. PROCESS OF SLR

There are two landmark literature surveys proposed by Budgen *et al.* [15] and Zlatko Stapić *et al.* [16] in the field of software engineering. We adopt their approaches in our SLR and categorize the review process into three main stages as shown in Figure 2 in order to identify, evaluate, and understand various researches related to particular research questions.

Planning the Review. The purposes of this stage are to (a) identify the need, (b) develop criteria and procedures, and (c) evaluate the criteria and procedures related to this SLR.

Conducting the Review. Based on the guiding principles proposed in [17]–[19], this stage includes six obligatory phases.

A. Research Questions (RQs): The purpose of the RQ phase is to identify relevant studies that need to be considered in the current review. We determine a set of RQs (described later) in the context of the Deepfake domain.

B. Search strategy (SS): A predefined search strategy aims to find as many as primary studies related to our research questions. We try to establish an unbiased search

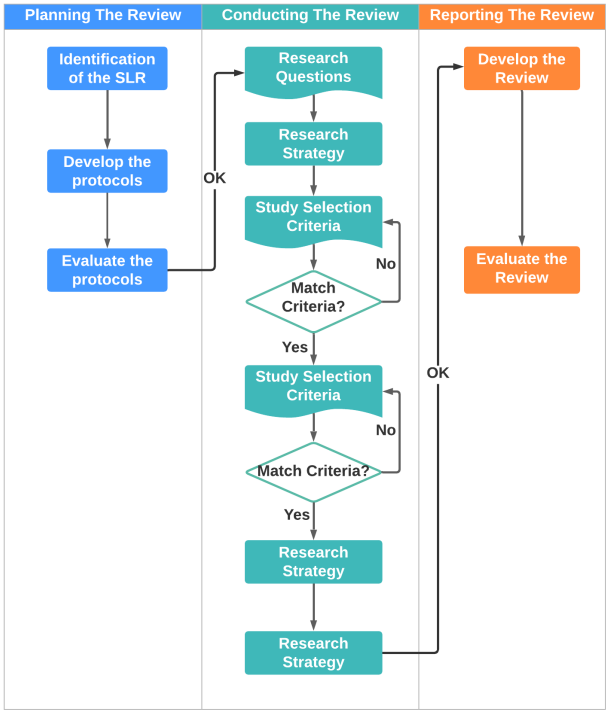

FIGURE 2. The process of the SLR.

strategy to detect as much of the relevant literature as possible.

C. Study Selection Criteria (SSC): There are challenges in the literature selection process, including the language of the study, knowledge of the authors, institutions, journals, or year of publication, etc [17]. Before ascertaining selection criteria, we follow careful consideration to ensure fairness in selecting primary studies that provide significant evidence about research questions.

D. Quality Assessment Criteria (QAC): The goal of assessing each primarily selected study's quality is to ensure that the study findings are relevant and unbiased. We develop a set of quality criteria for evaluating individual studies.

E. Data extraction and monitoring (DEM): We carefully determine how the information required from selected studies would be obtained and record their pieces of evidence.

F. Data Synthesis (DS): Data synthesis aims to organize and summarize outcomes obtained from the selected studies. We follow a set of procedures to synthesize information better.

Reporting the Review. After completing the review of all the studies, we report the outcomes in a suitable form to the distribution channel and target audience.

A. RESEARCH QUESTIONS (RQs)

Choosing research questions (RQs) is the first step in defining a particular study's overall purpose and expected outcomes. As such, we establish our RQs to make them meaningful to researchers because the right question leads to raising confidence in a domain [18]. Therefore, to recognize the recent exercise in the field of Deepfake detection, we define four such crucial questions (RQ 1-4) along with some supplementary questions (SRQs), shown in Table 1. As pointed out in the table, we first identify the different categories of Deepfake detection techniques. Next, we investigate the procedures of related empirical experiments. Under the same research question, RQ-2, we further deepen down by asking some supplemental questions (SRQ-2.1 to SRQ-2.4) to follow internal details that include:

- Describing datasets used to conduct experiments.
- Features that are commonly used by several methods.
- Models or algorithms used to detect Deepfake.
- Measurement metrics used to assess various method's performances in detecting such Deepfakes.

Then, we evaluate the overall performance of different methods using various measurement metrics in RQ-3. Finally, we compare models with respect to efficiency using the same dataset and same measurement metric.

B. SEARCH STRATEGY (SS)

We intended to collect as many works as possible that are relevant to our research questions. During collecting Deepfake detection studies, we tried to include all the combinations of related search phrases or keywords to avoid any bias. The key idea of using Boolean terminology for combining those searching terms with 'AND' or 'OR'. The search words can be outlined primarily (Deepfake OR FaceSwap OR Video manipulation OR Fake face/image/video) AND (detection OR detect) OR (Facial Manipulation OR Digital Media Forensics). Instead of relying on one or two sources, we looked into several repositories to ensure a proper search. However, there are many digital repositories are available for finding the research articles. We selected 10 popular repositories from them by considering their relevance and availability as listed below:

- Web of Science
- IEEE Xplore Digital Library
- ACM Digital Library
- ScienceDirect (ELSEVIER)
- SpringerLink
- Google Scholar
- Semantic Scholar
- Cornell University
- Computing Research Repository
- Database Systems and Logic Programming (DBLP)

The repositories include journals, conferences, and archives. We limit our search duration from January 2018 to December 2020.

C. STUDY SELECTION CRITERIA (SSC)

We establish three inclusion criteria in our search procedure in order to select the relevant articles while searching in these 10 digital repositories.

TABLE 1. Define research questions for the SLR.

QID	Research Questions	Objectives
RQ-1	What are the well-known Deepfake detection techniques?	Describe various categories of Deepfake detection techniques.
RQ-2	What is the way to perform empirical tests to detect Deepfake using these studies?	Discover the process of empirical experiments.
SRQ-2.1	What datasets are typically utilized for detecting Deepfake?	Find datasets that are used in experiments.
SRQ-2.2	What are the features typically utilized in detecting Deepfake?	Distinguish the features that are widely used.
SRQ-2.3	What models are used to detect Deepfake manipulation?	Classify the commonly applied models.
SRQ-2.4	What measurement metrics are used for computing the performance of Deepfake detection methods?	Identify the measurement metrics that are mostly used to evaluate the performance of models.
RQ-3	What is the Classification Framework for Deepfake Detection Approaches?	Identify the Classification Framework for Deepfake Detection Approaches.
RQ-4	What is the general efficiency of a variety of Deepfake detection strategies based on experimental proof?	Evaluate the efficacy of many Deepfake detection methods.
RQ-5	Is the Efficiency of Deep Learning Models Better than Non-deep Learning Models in Deepfake Detection Based on Experimental Results?	Based on the experiment, evaluate each model's performance, and perform a comparative analysis between them.

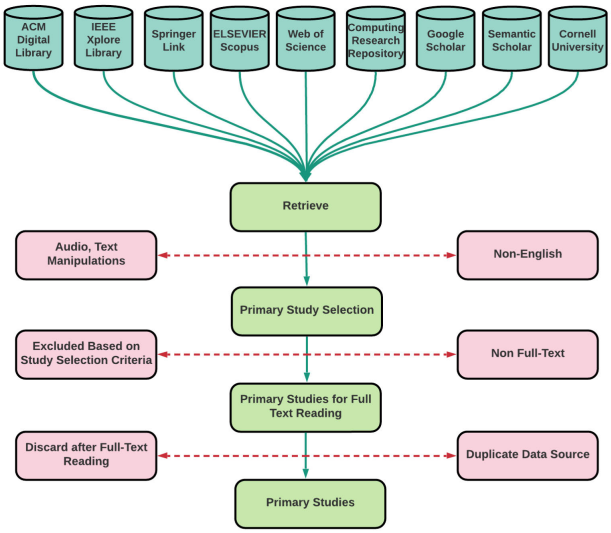

FIGURE 3. Study selection process.

- The search phrases are part of the title or abstract, or keywords.
- Some works mainly dealt with Deepfake without mentioning related keywords in the title, abstract, or keywords. In such a case, we look for the desired keywords in other parts of the literature. We include those works if we find any.
- Empirical evidence is explicitly presented in writing.

Besides, a series of exclusion criteria are also established to skip studies that may not be relevant from this review (see Figure 3):

- Studies that are not written in English.
- A few pieces of research are published concurrently in conferences and journals. In such a case, we considered the most comprehensive one to avoid duplicates.
- As our primary objective of this SLR is to study image or video manipulation, we omitted audio and text manipulation analysis.
- We filter out the research that focuses on specific *transformation techniques* in Deepfake detection.

D. QUALITY ASSESSMENT CRITERIA (QAC)

Assessing the quality of evidence contained within an SLR is as important as analyzing the data within. Results from a poorly conducted study can be skewed by biases from the research methodology and should be interpreted with caution. Such studies should be acknowledged as such in the systematic review or outright excluded. Selecting appropriate criterion to help analyze strength of evidence and imbedded biases within each paper is also essential. Based on the criterion defined in [20] we validate the selected studies using these criterion and review these studies by applying the requirements. Also, a cross-checking approach has been used for assessing these selected studies to ensure consistency among different findings. After this quality assessment phase, we finalized 91 research articles and 21 additional reviews (7 SLRs, 10 analyses, and 4 surveys) representing Deepfake detection.

E. DATA EXTRACTION AND MONITORING (DEM)

This phase describes designing systems for the actual extraction of data from the studies. To find possibly relevant articles, we thoroughly searched nine popular libraries (see Figure 3). We chose studies that matched the following requirements: 1) The methods or results section stated what entities were or needed to be extracted, and 2) at least one entity was automatically extracted, with assessment findings for this kind of entity given. The answers to the RQs are determined based on the knowledge gained from the data extraction process (see Figure 4).

- *Author(s), publication sources, and publication times:* In this part, we obtain the author's information,

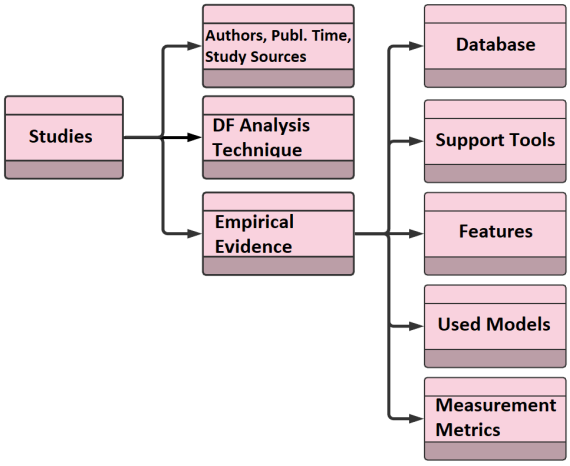

FIGURE 4. The information of the extracted data.

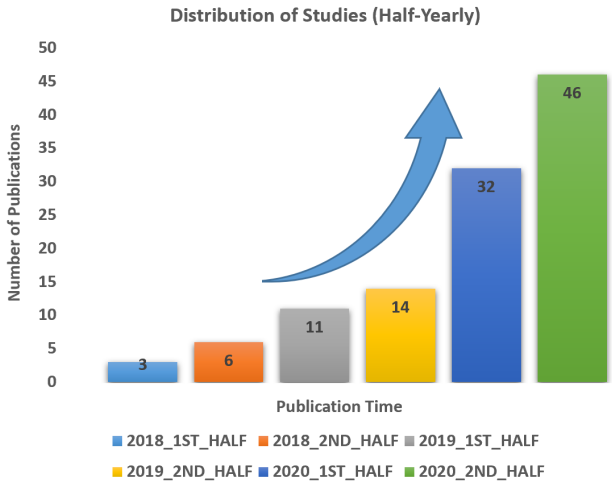

FIGURE 5. Distribution of studies (half-yearly).

publication period, and the origin of publications: conference, workshop, or journal.

- *Analysis techniques*: Based on this study, we identified the various methods based on the feature analysis that are applied for detecting Deepfakes.
- *Empirical evidence:* In this part, we focused on the following four components: (i) datasets that are used by the study authors, (ii) features used for analysis, (iii) models or methodologies applied, and (iv) measurement metrics that are used by study authors to evaluate their results.

F. DATA SYNTHESIS (DS)

The data synthesis phase specifically reviews the associated and comparative findings from the data extraction process, which can be presented as indications to support definitive responses to the RQs. After accumulating the data, we analyze them for further information extraction, and visualize the collected data through various data visualization tools and techniques, such as histograms, pie maps, tables, etc.

III. OUTCOMES

A. DESCRIPTION OF STUDIES

We accumulate a total of 112 studies from our determined sources within three years of the publication period.

1) PUBLICATION PERIOD

The Deepfake related research primarily emerged in 2018. Therefore, we considered the publication period from the beginning of 2018 until 2020. As presented in Figure 5, over the span, the number of publications increased exponentially. In the figure, we report half-yearly publications to count. As shown, there are only three publications in the first half of 2018, which becomes double in the second half. A similar trend continued in 2019. However, this trend broke in 2020, with a surge of 32 publications in only the first six months. This rising trend continues over the year, with almost 1.5 times more publications in the last part of the year than in the first half, indicating the research thirst in the Deepfake sphere.

TABLE 2. Source of publications.

No	Name of the Publication Source	Type	QTY
1	Cornell University and Computing Research Repository	Archive	33
2	IEEE/CVF Conference on Computer Vision and Pattern Recognition	Conference	9
		Workshop	5
3	IEEE International Workshop on Information Forensics and Security	Workshop	3
4	IEEE International Conference on Acoustics, Speech and Signal Processing	Conference	2
5	International Conference on Machine Learning (ICML)	Conference	2
6	European Conference on Computer Vision	Conference	2
7	International Conference of the Biometrics Special Interest Group	Conference	2
8	IEEE Access	Journal	2

2) SOURCE OF PUBLICATIONS

We mainly consider eight different publication sources from recognized conferences, workshops, journals, and archives. We observe that more articles were published as archived papers in the domain of Deepfake, whereas a few papers were issued in the journal. We present source wise publications number in the Table 2. We didn't include the source in the table if the publication count is below two.

B. RQ-1: WHAT ARE THE POPULAR DEEPFAKE DETECTION TECHNIQUES?

As discussed in Section II-A, we explore the overall survey in the form of some research questions. As part of the discussion, we first determine the Deepfake detection techniques widely used in the literature. Though Deepfake

mainly manipulates images or video using deep learning (DL) based technique, other methods along with DL obtain Deepfake. We categorize different researches according to the applied techniques and describe them in the following sections.

1) MACHINE LEARNING BASED METHODS

Traditional machine learning (ML) algorithms are instrumental in comprehending the logic for any decision that could be expressed in human terms. Such methods are suitable for the Deepfake domain as there is a better grasp of the data and processes. In addition, tuning hyper-parameters and changing model designs are much more manageable. The tree-based ML approaches, for example, Decision Tree, Random Forest, Extremely Randomized Trees, etc., show the decision process in the form of a tree. Therefore, a tree-based method does not have any explainability issues.

GANs are used to automatically train a generative model by treating the unsupervised issue as supervised and creating photo-realistic fake faces in images or videos. Some ML-based methods aspire to show certain irregularities found in such GANs generated fake videos or images.

A very fundamental approach of Deepfake is to manipulate the human face to confuse its audiences. There are different approaches to do that. However, to fool the users, most techniques modify certain regions of the face, such as shade of the eyes, ear with a ring, etc. Such methods using a single part (a.k.a. feature) are limited to identifying or detecting the manipulated area. To overcome these, the authors in [21] proposed a Deepfake technique by combining a set of such features.

In [22], the consistency of the biological signs are measured along with the spatial and temporal [23]–[25] directions to use various landmark [26] points of the face (e.g., eyes, nose, mouth, etc.) as unique features for authenticating the legitimacy of GANs generated videos or images. Similar characteristics are also visible in Deepfake videos, which can be discovered by approximating the 3D head pose [27]. In most cases, facial expressions are associated initially with the head's movements. Habeeba *et al.* [88] applied MLP to detect Deepfake video with very little computing power by exploiting visual artifacts in the face region.

As far as the performance concern in machine learning-based Deepfake methods, it is observed that these approaches can achieve up to 98% accuracy in detecting Deepfakes. However, the performance entirely relies on the type of dataset, the selected features, and the alignment between the train and test sets. The study can obtain a higher result when the experiment uses a similar dataset by splitting it into a certain level of ratio, for example, 80% for a train set and 20% for a test set. The unrelated dataset drops the performance close to 50%, which is an arbitrary assumption.

2) DEEP LEARNING BASED METHODS

In the case of Deepfake detection in images, there are plenty of works where deep learning-based methods are applied to detect specific artifacts generated by their generation pipeline. Zhang *et al.* [33] introduced a GAN simulator that replicates collective GAN-image artifacts and feeds them as input to a classifier to identify them as Deepfake. Zhou *et al.* [34] proposed a network for extracting the standard features from RGB data, while [35] proposed a similar but generic resolution. Besides, in [36]–[38], researchers proposed a new detection framework based on physiological measurement, for example, Heartbeat.

At first, the deep learning-based method was proposed in [40] for Deepfake video detection. Two inception modules, (i) Meso-4 and (ii) MesoInception-4, were used to build their proposed network. In this technique, the mean squared error (MSE) between the actual and expected labels is used as the loss function for training. An enhancement of Meso-4 has been proposed in [41].

In a supervised scenario, the authors in [42] shows that the deep CNNs [43]–[45] outperform shallow CNNs. Some methods apply techniques for extracting the handcrafted features [46]–[47], spatiotemporal features [48]–[51], common textures [52], [53], 68 face landmarks [54]–[56] with visual artifacts (i.e., eye, teeth, lip movement, etc.) from the video frames. Such features were used as input to the these networks for detecting Deepfake manipulations. Besides data augmentation [57], super-resolution reconstruction [58], localization strategies in pixel levels [11] are formulated on the entire frame, and maximum mean discrepancy (MMD) loss [59] is applied to discover a more general feature.

Further innovations are achieved by introducing an attention mechanism [61] while promising outcomes are shown in [62]–[63] by using an architecture named capsule-network (CN). The CN needs a smaller number of parameters to train than very deep networks. An ensemble learning technique [64]–[65] is applied to increase such structures' performance, which achieves more than 99% accuracy.

We observe that many approaches were proposed to apply frame-by-frame analysis in videos or images to manipulate face and track facial movement to obtain better performance. For example, in [66]–[71], RNN based networks are proposed to extract the features at various micro and macroscopic levels for detecting Deepfake. Regardless of these exciting results in detection, it is seen that most of the methods lean towards overfitting. The optical fow based technique [72] and autoencoder-based architectures [73]–[76] are introduced to resolve such problems. A pixel-wise mask [77] is imposed on various models to get the essential depiction of the face's affected area. Fernando *et al.* [78] applied adversarial training approaches followed by attention-based mechanisms for concealed facial manipulations. In [93], researchers proposed a clustering technique by integrating a margin-based triplet embedding regularization term in their classification loss function. Finally, they converted the three-class classification problem to a two-class classification problem. The authors in [94]–[95] proposed a data pre-processing technique for detecting Deepfakes by applying CNN methods. The researchers in [96] proposed patch and pair convolutional

neural networks (PPCNN). In [97], authors performed an analysis in the frequency domain by exploiting the image latent patterns' richness. A modern approach called ID-revelation [98] was proposed to learn temporal facial features based on a person's movement during talking. A novel feature extraction method [99] had been proposed for effectively classifying Deepfake images. In [100], a multimodal approach was proposed for detecting real and Deepfake videos. This method extracts and analyzes the similarities between the audio and visual modalities within the same video. In [101], a Deepfake detection method is applied to find the discrepancies between faces and their context by combining multiple XeptionNet models.

In [101], a separable convolutional network is used for detecting such manipulations. [103] resorts to the feature extraction process's triplet loss function to better classify fake faces. A patch-based classifier was introduced in [104] to focus on local patches rather than the global structure. In [105]–[106], the authors extracted features using improved VGG networks. A hypothesis test was performed in [107].

3) STATISTICAL MEASUREMENTS BASED METHODS

Determining different statistical measures such as average normalized cross-correlation scores between original and suspected data helps to understand the originality of the data. Koopman *et al.* [108] examined the photo response non-uniformity (PRNU) for detecting Deepfakes in video frames. PRNU is a unique noise pattern in the digital images that occurred due to the defects in the camera's light-sensitive sensors. Because of its distinctiveness, it is also considered the fingerprint of digital photos. The research generates a sequence of frames from input videos and stores them in chronologically categorized directories. Each video frame is clipped with the same pixel range to preserve and clarify the portion of the PRNU sequence. These frames are then divided into eight equal groups. It then makes the standard PRNU pattern for each frame using the second-order FSTV method [147]. After that, it correlates them by measuring the normalized cross-correlation scores and calculating the differences between the correlation scores and the mean correlation score for each frame. To evaluate statistical significance between Deepfakes and original videos, the authors conduct a t-test [109] on the results.

To model a basic generating convolutional structure, the authors in [110] extracted a collection of regional features using the Expectation-Maximization (EM) algorithm. After the extraction, they apply ad-hoc validation to those architectures, such as GDWCT, STARGAN, ATTGAN, STYLEGAN, and STYLEGAN2, using preliminary experiments naive classifiers. Agarwal *et al.* [111] performed a hypothesis test by proposing a statistical framework [112] for detecting the Deepfakes. Firstly, this method defines the shortest path between distributions of original and GAN-created images. Based on the results of this hypothesis, this distance measures the detection capability. For example, Deepfakes can easily be detected when this distance is increased. Usually, the distance increases *iff* the GAN provides a lesser amount of correctness. Besides, an extremely precise GAN is mandatory to create high-resolution manipulated images that are harder to detect.

4) BLOCKCHAIN BASED METHODS

Blockchain technology provides various features that can verify the legitimacy and provenance of digital content in a highly trusted, secured, and decentralized manner. In public Blockchain technology, anyone has direct access to every transaction, log, and tamper-proof record. For Deepfake detection, public Blockchain is considered one of the most appropriate technological solutions for verifying video's or image's genuineness in a decentralized way. Users usually need to explore the origin of videos or images when they are marked as suspected.

Hasan and Salah [113] proposed a Blockchain-based generic framework to track suspected video's origin to their sources. The proposed solution can trace its transaction records, even though the material is copied several times. The basic principle says that digital content is considered authentic when convincingly traced to a reliable source. For Deepfakes, public Blockchain verifies video content's legitimacy in a decentralized way, as the technology can provide some critical features to prove its authenticity. The following are the main contributions of [113].

- Presents a generic framework based on Blockchain technology by setting up a proof of digital content's authenticity to its trusted source.
- Presents the proposed solution's architecture and design details to control and administrate the interactions and transactions among participants.
- Integrates the critical features of IPFS [114]-based decentralized storage ability to Blockchain-based Ethereum Name service.

Chan *et al.* [115] proposed a decentralized approach based on Blockchain to trace and track digital content's historical provenance (i.e., image, videos, etc.). In this proposed approach, multiple LSTM networks are being used as a deep encoder for creating discriminating features, which are then compressed and used to hash the transaction. The main contributions of this paper are as follows.

- Using multiple LSTM CNN models, image/video contents are hashed and encoded.
- High dimensional features are preserved as a binary coded structure.
- The information is stored in a permission-based Blockchain, which gives the owner control over its contents.

Based on the studies, taking together all these methods, Table 3 lists the categories of Deepfake detection strategies and displays the quantity (No.) and percentage (PCT) of related categories of studies. This table includes 91 studies, except 21 different reviews ([60], [116]–[135]) which merge various methods. Also, this table reveals that the

TABLE 3. Classification of Deepfake detection methods.

Category	Studies	Count	%
Deep learning-based methods	[11], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [86], [89], [90], [91], [92], [93], [94], [95], [96], [97], [98], [99], [100], [101], [102], [103], [104], [105], [106]	70	77%
Machine learning-based methods	[21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [85], [87], [88], [142]	16	18%
Statistical-based methods	[107], [109], [110]	3	3%
Blockchain-based methods	[112], [114]	2	2%

TABLE 4. The list of Deepfake datasets.

Database Name	#Deepfakes	#Actors
FaceForensics (FF) [11]	1000	977
FaceForensics++ (FF++) [42]	1000	977
DeepfakeDetection (DFD) [135]	3000	28
DeepFake Forensics (Celeb-A) [136]	202,599 (images)	10,177
DeepFake Forensics (Celeb-DF) [137]	795+ 590	13+59
Deepfake Detection Chal. (DFDC) [138]	5214	66
UADFV [27]	49	-
Deepfake-TIMIT (DF-TIMIT) [91]	620	64
DeeperForensics-1.0 (DF-1.0) [139]	60,000	100
WildDeepfake (WDF) [140]	707	100
MANFA [84]	8950 (images)	-
SwapMe and FaceSwap (SMFW) [30]	1005 (images)	-
Deep Fakes (DFS) [22]	142	-
Fake Faces in the Wild (FFW) [141]	150	-
FakeET (FE) [142]	811	40
FaceShifter (FS) [143]	5000 (images)	-
Deepfake (DF) [39]	175	50
Swapped Face Detection (SFD) [81]	420,053 (images)	86

deep learning-based approach is the most widely used technique, accounting for around 77% in all studies. The research relating to machine learning approaches and statistical methods is 18% and 3%, respectively. The number of studies in this analysis on the Blockchain-based approach is 2%. Overall, we divide the Deepfake detection techniques into four categories: deep learning-based methods, machine learning-based techniques, statistical-based techniques, and Blockchain-based techniques. Among them, deep learning-based methods are used broadly for detecting such Deepfakes.

C. RQ-2: WHAT IS THE WAY TO PERFORM EMPIRICAL TESTS TO DETECT DEEPFAKE USING THESE STUDIES?

To provide an answer to RQ-2, we review the different experimental methods in-depth and categorize the overall Deepfake detection process into six distinct stages (see Figure 6) that are summarized below.

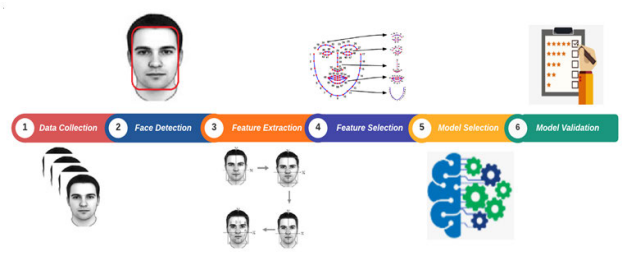

FIGURE 6. Steps of Deepfake detection.

- *Data collection*: Collecting and organizing unadulterated original and Deepfaked data (images or videos) is done in this initial phase.
- *Face detection*: Identifying which parts of an image or video need to be focused on to reveal characteristics like age, gender, emotions, etc., using facial expressions fall under this stage.
- *Feature extraction*: Extracting various features from the face area as candidate features for the detector.
- *Feature selection*: Select from the extracted features those that are most useful for Deepfake detection.
- *Model selection*: Finding a suitable model from a pool of available models for classification. These models include deep learning-based models, machine learning-based models, and statistical models.
- *Model evaluation*: Finally, evaluating the performance of the selected models using various measurement metrics.

The following sub-sections describe the datasets used in several experiments, the frequently utilized features, models used for detection tasks, and measurement metrics used to evaluate models' performance in detecting Deepfakes.

1) SRQ-2.1: WHAT DATASETS ARE TYPICALLY USED IN DEEPFAKE DETECTION EXPERIMENTS?

We found various Deepfake datasets used in numerous studies for training and testing purposes. In turn, these datasets have enabled incredible advances in Deepfake detection. Most of the real videos in these datasets are filmed with a few volunteer actors in limited scenes. The fake videos are crafted by researchers using a few popular Deepfake software.

Figure 7 displays various datasets that are used in different studies. From this figure, it is observed that FaceForensics++, Celeb-DF, and DFDC are quite popular and were used in plenty of studies. Table 4 describes a summary of these datasets.

2) SRQ-2.2: WHAT FEATURES ARE TYPICALLY UTILIZED IN DETECTING DEEPFAKE?

Based on the categories of Deepfake detection and analysis techniques described in RQ-1, 21 studies use

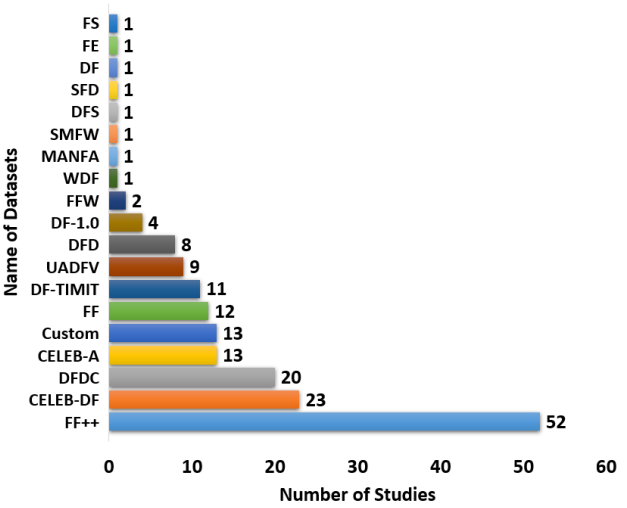

FIGURE 7. List of datasets used in Deepfake related studies.

special artifacts-based features generated by various editing processes. Among them, 20 studies use texture and Spatio-temporal consistent features, 14 studies involve facial landmarks-based features. Also, 13 research papers perform experiments using visual artifacts-based elements, for example, eye blinking, head posing, lip movement, etc. Eight pieces of work apply biological characteristics, whereas seven studies concern intra-frame inconsistencies with frequency domain analysis. In addition, six studies use GAN-based features, and four studies cover latent space-based features. Ten studies use custom features utilizing various analyses that include error level analysis, mesoscopic analysis, steganalysis, super-resolution, augmentation, maximum mean discrepancy, PRNU pattern analysis, etc. The details are described in the Result section using RQ-1. The study shows that special artifacts-based features, face landmarks, and Spatio-temporal features are used widely to detect Deepfakes.

3) SRQ-2.3: WHAT MODELS ARE USED TO DETECT DEEPFAKE MANIPULATION?

This segment describes various models that are used for detecting Deepfake. Based on this study, we divide these models into three groups: (i) deep learning model, (ii) machine learning model, and (iii) statistical model.

- **Deep Learning models**: In computer vision, deep learning models have been used widely due to their feature extraction and selection mechanism, as they can directly extract or learn features from the data. In Deepfake detection studies, we found the following deep learning-based models have been used: convolutional neural network (CNN) model (e.g., XceptionNet, GoogleNet, VGG, ResNet, EfficientNet, HRNet, InceptionResNetV2, MobileNet, InceptionV3, DenseNet, SuppressNet, StatsNet), Recurrent Neural Network (RNN) model (e.g., LSTM, FaceNet), Bidirectional RNN model, Long-term Recurrent Convolutional Neural Network (RCNN) model, Faster RCNN model, Hierarchical Memory Network (HMN) model, Multi-task Cascaded CNNs MTCNN) model and Deep Ensemble Learning (DEL).
- **Machine learning model**: This technique creates a feature vector by defining the right features using various state-of-art feature selection algorithms. It then feeds this vector as input to train a classifier to classify whether the videos or images are manipulated by Deepfake or not. Support Vector Machine (SVM), Logistic Regression (LR), Multilayer Perceptron (MLP), Adaptive Boosting (AdaBoost), eXtreme Gradient Boosting (XGBoost), and K-Means clustering (k-MN), Random Forest (RF), Decision Tree (DT), Discriminant Analysis (DA), Naive Bayes (NB) and Multiple Instance Learning (MIL) are used as machine learning-based models.
- **Statistical Model**; The statistical models are based on the use of the information-theoretic study for validation. In these models, the shortest paths are calculated between original and Deepfake videos/images distributions. For example, in [108], a significance is measured for mean normalized cross-correlation scores between the original and the Deepfake videos, classifying them as fake or real. The often-applied statistical models are Expectation-Maximization (EM), Total Variational (TV) distance, Kullback-Leibler (KL) divergence, Jensen-Shannon (JS) divergence, etc.

TABLE 5. Distribution of used models.

Category	Model	#Studies	PCT (%)
Deep Learning	CNN	71	78%
	RNN	12	13%
	RCNN	2	2%
Machine Learning	SVM	11	12%
	k-MN	4	4%
	LR	3	3%
	MLP	3	3%
	BOOST	2	2%
	RF	1	1%
	DT	1	1%
	DA	1	1%
	NB	1	1%
	MIL	1	1%
Statistical	EM	1	1%
	TV, KL, JS	1	1%

Based on these studies, we conduct a categorization in the deep learning models, machine learning models, and statistical methods, as shown in Table 5. The table outlines the number and the percentage of models used in the studies, except for 21 different reviews. Also, we observe that the DL-based studies hold the highest proportion of SLR.

Figure 8 displays the full versions of detector groups that are found from these primary studies, where CNN has the most divisions. Based on this Table 5, we further apply a subcategorization on CNN models and found that the following 3 CNN models: (i) XeptionNet and ResNet take 17% and (iii) VGG with 12%, respectively. Besides, LSTM models

TABLE 6. Summary of works towards Deepfake detection.

Reference	Focus	Methods	Models	Features	Datasets
Sharp_Multi_Instance_Learning [23]	DMF	ML	MIL	STC	CELEB-DF, FF, DFDC, FF+
Conv_Traces_on_Images [24]	DMF	ML, STAT	SVM, DA, KNN, EM	STC	CELEB-A, FF+
Dynamic_Texture_Analysis [25]	DMF	ML	SVM	TEX	FF++
Anomalous_Co-motion_Pattern [26]	DMF	ML, STAT	ADB, CRA	FL	FF++
Unmasking_DeepFakes [29]	FM	ML	SVM, LR, k-MN	FDA	CELEB-A, FF++, Other
Metric_Learning [32]	FM	DL, ML	MTCNN, RNN, MLP	SA, FL	CELEB-DF, FF+
Audio_Visual_Dissonance [35]	FM	DL	CNN	BA	DFDC, DF-TIMIT
DeepRhythm [36]	FM	DL	CNN, RNN	BA, FL	DFDC, FF++
DeepFakesON-Phys [38]	DMF	DL	CNN	BA	DFDC, CELEB-DF
A_Note_on_Deepfake [41]	FM	DL	CNN	MES	FF++
Conditional_Distribution_Modelling [45]	FM	DL	CNN	SA	FF
Spatio-temporal_Features [48]	FM	DL	CNN	STC	DFDC, FF++, DF-1.0
Time-Distributed_Approach [49]	FM	DL	CNN, RNN	TEX	DFDC
Cost_Sensitive_Optimization [50]	FM	DL	CNN, RNN	TEX	FF++, DF-TIMIT
Lips_Do_not_Lie [51]	FM	DL	CNN, MSTCN	BA	DFDC, CELEB-DF, FS, FF++, DF-1.0
3D_Decomposition [52]	FM	DL	CNN	TEX	DFDC, FF++, DFD
Auxiliary_Supervision [53]	FM	DL	CNN	STC, TEX	FF, FF++
Forensics_and_Analysis [54]	FM	DL	CNN	BA, FL	CELEB-DF, DF-TIMIT
Identity_Driven_DF_Detection [55]	DMF	DL	CNN	SA, FL	CELEB-DF, DFD, FF++, Other
Patch_Wise_Consistency [56]	FM	DL	CNN	FL, IFIC	DFDC, CELEB-DF, DFD, FF++, DF-1.0
Data_Augmentations [57]	FM	DL	CNN	IMG	DFDC, CELEB-DF, DFD, FF++
Super-resolution_Reconstruction [58]	FM	DL	CNN	SA	FF++
MMD_Discriminative_Learning [59]	FM	DL	CNN	SA	UADFV, CELEB-DF, DF-TIMIT, FF++
On_the_Detection [61]	FM	DL	CNN	GAN	FF++
Ensemble_of_CNNs [64]	FM	DL	CNN	SA, IFIC	DFDC, FF++
DeepfakeStack [65]	FM	DL	CNN	SA	CELEB-DF, FF++
Conv_LSTM_Residual_Net [69]	FM	DL	MTCNN, RNN	FL	FF++
Two-Branch_RNN [70]	FM	DL	RNN	FDA	DFDC, CELEB-DF, FF++
Recurrent_Conv_Structures [71]	DMF	DL	CNN, RNN	STC	CELEB-DF, FF+
Dynamic_Prototypes [76]	FM	DL	CNN	SA	DFDC, FF+
Face_X-ray [79]	FM	DL	CNN	FL	DFD, CELEB-DF, DFDC, FF++
Manipulated_Face_Detector [80]	FM	DL	CNN	FL	FF, CELEB-A, FF++
Subjective_Assessment [82]	FM	DL	CNN	SA	Other
Adaptive_Residuals_Extract_Net [83]	DMF	DL	CNN	SA	CELEB-A, FF++
Automatic_Face_Weighting [84]	FM	DL	CNN, RNN	STC, VA	DFDC
Real_or_Fake [86]	FM	DL	CNN	TEX	Other
Watch_Your_Up-Convolution [87]	FM	DL, ML	CNN, MLP	GAN	CELEB-A, FF++
Visual_Artifacts_and_MLP [88]	FM	ML	MLP	FL, VA	UADF, DFD
Easy_to_Spot_for_Now [90]	DMF	DL	CNN	GAN	CELEB-A, FS, FF++, Other
Adversarial_Perturbations [92]	DMF	DL	CNN	GAN	CELEB-A
Cluster_Embed_Regularization [93]	FM	DL	CNN	VA	UADF, DFD, DF-TIMIT
Face_Preprocessing_Approach [94], [95]	FM	DL	CNN	IMG, VA	CELEB-DF, DFDC, FF+
Patch_and_Pair_CNN [96]	FM	DL	CNN	IFIC	FF, DF-TIMIT, Other
Efficient-Frequency [97]	Both	DL	CNN	FDA	DFDC, UADFV, DFW, CELEB-DF, DF-TIMIT, FF++
ID-Reveal [98]	FM	DL	CNN	VA	CELEB-DF, DFD, FF++
Counterfeit_Feature_Extraction [99]	DMF	DL	CNN	VA	Other
Emotions_Do_not_Lie [100]	FM	DL	CNN	FL	DFDC, DF-TIMIT
Face_Context_Discrepancies [101]	FM	DL	CNN	STC, VA	CELEB-DF, DFDC, FF+
Deep_Detection [102]	FM	DL	CNN	CPRNU	UADFV, CELEB-DF, FF++
What_Makes_Fake_Images [103]	FM	DL	CNN	IMG, VA	CELEB-A, FF++, Other
Improved_VGG_CNN [104]	FM	DL	CNN	IMG, VA	CELEB-DF
Interpret_Residuals_Bio-Signals [105]	FM	DL	CNN	BA	CELEB-DF, FF++

TABLE 6. *(Continued.)* **Summary of works towards Deepfake detection.**

Reference	Focus	Methods	Models	Features	Datasets
Eyebrow_Recognition [106]	DMF	DL	CNN	VA	CELEB-DF
Analyze_Convolutional_Traces [109]	DMF	STAT	EM	GAN	CELEB-A
Multi-LSTM_and_Blockchain [114]	DMF	BC	RNN	TEX	DF-TIMIT
FakeET [142]	FM	DL, ML	CNN, RF, NB, LR, k-NN, DT, SVM	SA	DFDC, FE
Exploit_Visual_Artifacts [21]	DMF	ML	MLP, LR	VA	FF, CELEB-A, Other
FakeCatcher [22]	DMF	DL, ML	CNN, SVM	STC, BA	FF, Other
Inconsistent_Head_Pose [27]	FM	ML	SVM	SA, FL	UADFV
Protect_World_Leaders [28]	DMF	ML	SVM	SA	FF
Comp_Face_Forensic [31]	DMF	DL, ML	CNN, SVM	FL	FF, CELEB-A, FF++, Other
Detecting_Simulating_Artifacts [33]	FM	DL	CNN	SA, FDA	Other
Predict_Heart_Rate [37]	FM	DL	RNN	BA	DF-TIMIT
Hybrid_LSTM [39]	FM	DL	CNN, RNN	SA	Other
FaceForensics++ [42]	FM	DL	CNN	Other	FF++
Face_Warping_Artifacts [47]	FM	DL	CNN	SA	UADFV, DF-TIMIT
Capsule [62], [63]	DMF	DL	CNN	LS	FF++
Poster [67]	DMF	DL	RNN	IFIC	FF++
Recurrent_Conv_Strategies [68]	FM	DL	CNN	FL	FF++
Optical_Flow [72]	DMF	DL	CNN	VA	FF++
ForensicTransfer [73]	DMF	DL	CNN	LS	FF, Other
Multi-task_Learning [74]	DMF	DL	CNN	SA	FF, FF++
Locality-aware_Auto-Encoder [75], [77]	DMF	DL	CNN	LS	CELEB-A, FF++
Human_Social_Cognition [78]	FM	DL	HMN	VA	FF, FFW, FF++
Face_Image_Manipulation [85]	FM	DL, ML	CNN, XGB, ADB	FL	MANFA, SMFW
Pairwise_Learning [89]	FM	DL	CNN	STC	CELEB-A
Separable-CNN [101]	DMF	DL	CNN	SA	FF++
Robust_Estimation_Viewpoint [110]	DMF	STAT	Other	N/A	N/A
Blockchain_Smart_Contracts [111]	DMF	BC	RNN, ETH	N/A	N/A
FaceForensics [11]	FM	DL	CNN	Other	FF
Two-Stream_Neural_Networks [30]	FM	DL, ML	CNN, SVM	IMG	Other
Learn_Rich_Features [34]	FM	DL	RCNN	SA	Other
MesoNet [40]	FM	DL	CNN	MES	DF, FF
In_Ictu_Oculi [46]	FM	DL	RCNN	SA	UADFV
DF_Detection_by_RCNN [66]	FM	DL	CNN, RNN	STC	Other
Forensics_Face_Detection [81]	DMF	DL	CNN	GAN	CELEB-A
Face_Recognition_Threat [91]	DMF	DL	CNN	STC, VA	DF-TIMIT
Photoresponsive_pattern [107]	DMF	STAT	STAT	CPRNU	Other

take 13% of RNN. In addition to this, the most popular machine learning model is SVM with 12% and k-MN with 4%. The detail distribution in various models is presented in Figure 9 that shows the proportion of used models (e.g., DL, ML, Statistical) in various studies for detecting Deepfake. Besides, it provides the answer for SRQ-2.3. The reviewed papers show that the deep neural network (DNN) models are successful in Deepfake detection, where CNN-based models demonstrate more efficiency among all the DNN models.

At a glance. **Focus** indicates the clue for the detection (DMF: Digital Media Forensics, FM: Face Manipulation, Both: DMF and FM), **Methods** indicates method category (ML: Machine Learning, DL: Deep Learning, STAT: Statistical method, BC: Blockchain), **Models** represents types of model (DL: (CNN: Convolutional Neural Network, RNN: Recurrent Neural Network, RCNN: Regional Convolutional Neural Network, MTCNN: Multi-task Cascaded CNN, MSCNN: multi-scale Temporal CNN), ML: (SVM: Support Vector Machine, RF: Random Forest, MLP: Multilayer Perceptron Neural Network, LR: Logistic Regression, k-MN: K means clustering, XGB: XGBoost, ADB: AdaBoost, DT: Decision Tree, NB: Naive Bayes, KNN: K-Nearest Neighbour, DA: Discriminant Analysis), STAT: (EM: Expectation Maximization, CRA: Co-relation Analysis), BC: (ETH: Ethereum Blockchain)), **Features** (SA: Special Artifacts, VA: Visual Artifacts, BA: Biological Artifacts, FL: Face Landmarks, STC: Spatio-temporal Consistency, TEX: Texture, FDA: Frequency Domain Analysis, LS: Latent Feature, GAN: Generative Adversarial Network based feature, MES: Mesoscopic features, IFIC: Intra-frame inconsistency, CPRNU: Constrastive and Photoresponsive PRNU pattern, IMG: Image Metadata, Augmentation & Steganalysis, Other: Different feature not in

TABLE 7. Confusion matrix.

		Actual	
		Positive	**Negative**
Prediction	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

the common list), **Datasets** (FF: FaceForensics, FF++: FaceForensics++, DFD: Deepfake Detection, CELEB-A: DeepFake Forensics V1, CELEB-DF: DeepFake Forensics V2, DFDC: Deepfake Detection Challange, DF-TIMIT: Deepfake-TIMIT, DF-1.0: DeeperForensics-1.0, WDF: Wild Deepfake, SMFW: SwapMe and FaceSwap, DFS: Deep Fakes, FFD: Fake Faces in the Wild, FE: FakeET, FS: Face Shifter, DF: Deepfake, SFD: Swapped Face Detection, UADFV: Inconsistent Head Poses, MANFA: Tampered Face, Other: Authors' Custom datasets).

Finally, we summarize all at a glance using Table 6 that specifies the features, methods and models, datasets used throughout the studies and also focuses on specific manipulation detection techniques with having a reference to each of the primary studies.

4) SRQ-2.4: WHAT MEASUREMENT METRICS ARE USED FOR COMPUTING THE PERFORMANCE OF DEEPFAKE DETECTION METHODS?

This section briefly describes various measurement metrics applied for assessing the models' performance in detecting such Deepfakes. A confusion matrix holds info about actual and predicted classification results. The accounts of the detection capabilities of the used methods are measured and confirmed using this matrix data. Table 7 describes the confusion matrix.

Using Table 7, we can define the term, TP, which provides the number of Deepfakes that are correctly predicted as Deepfake, and TN offers the number of actual images/videos correctly predicted as *real*. Besides, FP stands for the number of *real* images/videos incorrectly predicted as Deepfake, where FN is the number of Deepfakes incorrectly predicted as the *real*. Similarly, using Table 8, we can define various measurement metrics and show how many studies are related to these metrics.

Based on the Table 8, it is seen that the often-applied measurement metrics are Accuracy (AC), receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC). Recall, error rate (ER), precision (P), f1-score, and log loss occupy a similar proportion. The least used performance measure is frechet-inception-distance (FID). Based on the study, accuracy and AUC are widely used measurement metrics in detecting Deepfake.

D. RQ-3: WHAT IS THE CLASSIFICATION FRAMEWORK FOR DEEPFAKE DETECTION APPROACHES?

For better insights, we summarize our key findings in Figure 10. As demonstrated in Figure, we classify overall approaches concerning different elements such as input data, features, method categories, and type of techniques. A path between two elements denotes the related components used in the companion paper for any method. As presented in the Figure, most papers apply image or video as the input data, whereas many papers use both image and video as the input. *Special Artifacts* and *Texture and Spatio-temporal Consistency* are the commonly used features in various papers. About 75% of the methods used the DL-based techniques as the detection method category. Only a few papers used Blockchain and Statistical approaches for detecting such Deepfake.

In detecting Deepfake, various underlying techniques are available, such as Biological Signals, Phoneme-Viseme Mismatches, facial expression and movements (i.e., 2D and 3D facial landmark positions, head pose, and facial action units), etc. We combine them under two central umbrellas of the methods that include *Facial Manipulation* and *Digital Media Forensics*. As shown in Figure 10, most of the DL-based methods exploit *Facial Manipulation* for the Deepfake detection. However, Machine Learning based methods almost equally utilize both techniques. Common to both Blockchain and Statistical approaches, they apply only Digital Media Forensics as part of the detection technique.

E. RQ-4: WHAT IS THE GENERAL EFFICIENCY OF A VARIETY OF DEEPFAKE DETECTION STRATEGIES BASED ON EXPERIMENTAL PROOF?

This segment attempts to decide the efficacy of Deepfake detection methods. The output assessment values are first obtained and stored in an Excel document based on the studies. After that, we count the number of studies that use the same method and the same measurement metrics (precision, accuracy, and recall). And finally, we apply four metrics: the minimum, maximum, mean, and standard deviation on these values (see Table 9).

In Table 9, based on the mean values of accuracy and AUC, deep learning-based methods outperform other methods and achieve 89.73% and 0.917, respectively. Besides, we also compare the recall and precision values for both techniques. Based on the overall results, we found deep learning-based techniques are efficient for detecting Deepfake.

F. RQ-5: IS THE EFFICIENCY OF DEEP LEARNING MODELS BETTER THAN NON-DEEP LEARNING MODELS IN DEEPFAKE DETECTION BASED ON EXPERIMENTAL RESULTS?

We split the models into two groups: (i) deep learning-based models and (ii) non-deep learning-based models. We determine the mean accuracy, AUC, recall, and precision.

Next, we apply a comparative analysis of these two models' performance and obtain an average result. Based on the evaluation of these models using performance

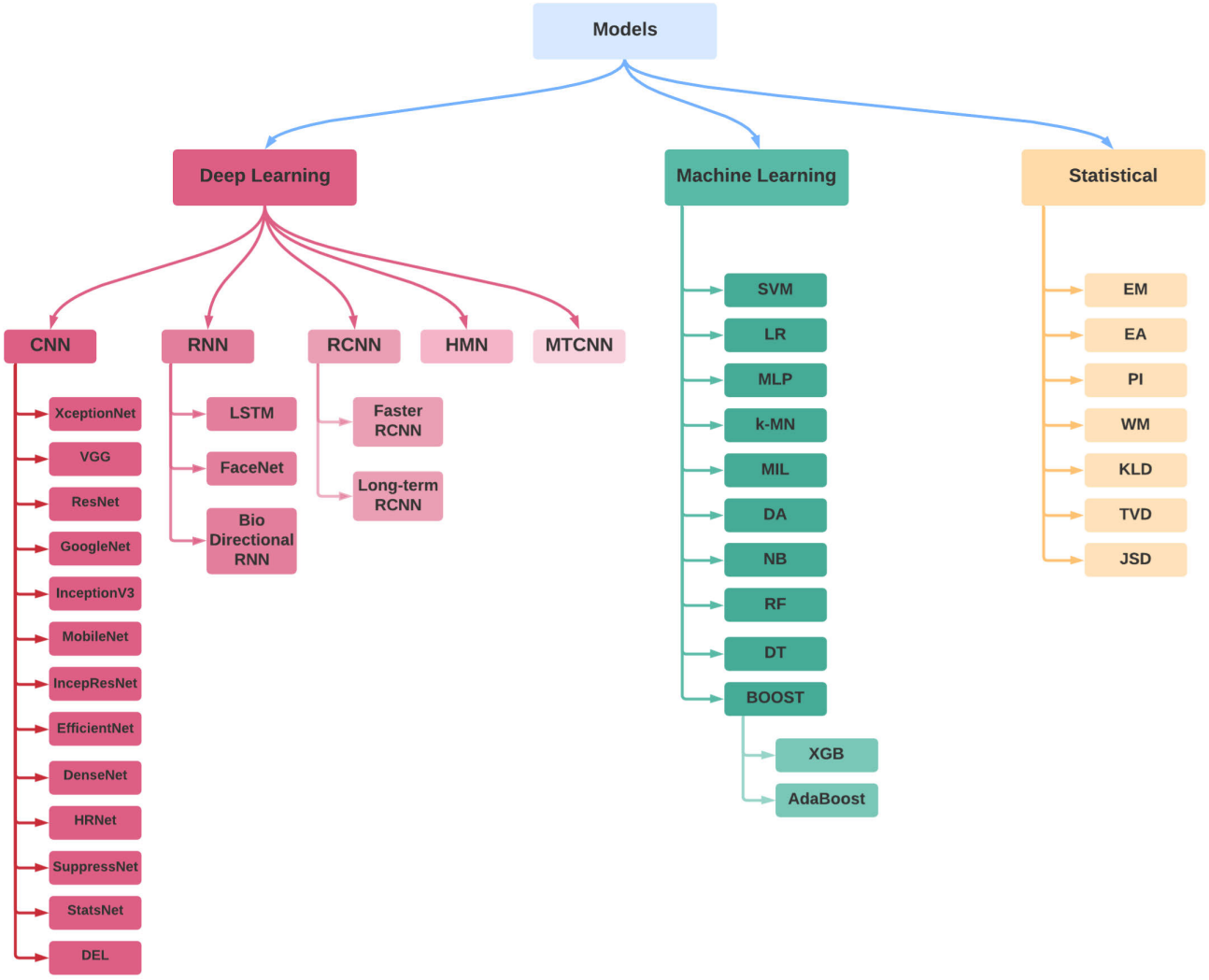

FIGURE 8. The list of Deepfake detection models.

measures (accuracy, AUC, recall, and precision), we observe, in general, deep learning-based models outperformed non-deep learning-based models. As the results are reported in Figure 11, the accuracy and precision performance in deep learning models are significantly better than non-deep learning models. However, in the case of AUC and recall, the performance is pretty similar. The overall results demonstrate the superiority of deep learning-based models over non-deep learning-based models.

IV. OBSERVATIONS

A. COMBINING DIFFERENT DEEP LEARNING METHODS IS CRITICAL FOR THE ACCURATE dEEPFAKE DETECTION

Based on the review, we see that multiple strategies are applied using numerous features. In general, primary methods used handcrafted features collected from face artifacts. Recent research applied deep learning-based approaches, especially the CNN models, to learn how to mechanically or directly learn perceptible and selective features to identify such Deepfake. For example, Ding *et al.* [82] introduced a two-phase CNN method for Deepfake detection. The first stage extracts particular features among counterfeit and actual images by incorporating various dense units, where each of them includes a list of dense blocks that are forged images. The second phase uses these features to train the proposed CNN to classify the input images, whether fake or real.

Due to the typical use of lossy compression in video compression, most detection techniques used in an image are not suitable for videos, as these methods degrade the frame data. Because videos have temporal features and vary the frames' size, it is challenging for techniques to distinguish just counterfeit images. In [68], a recurrent convolutional model (RCN) was proposed to use these spatiotemporal features [48]–[51] of videos for detecting Deepfakes. Likewise,

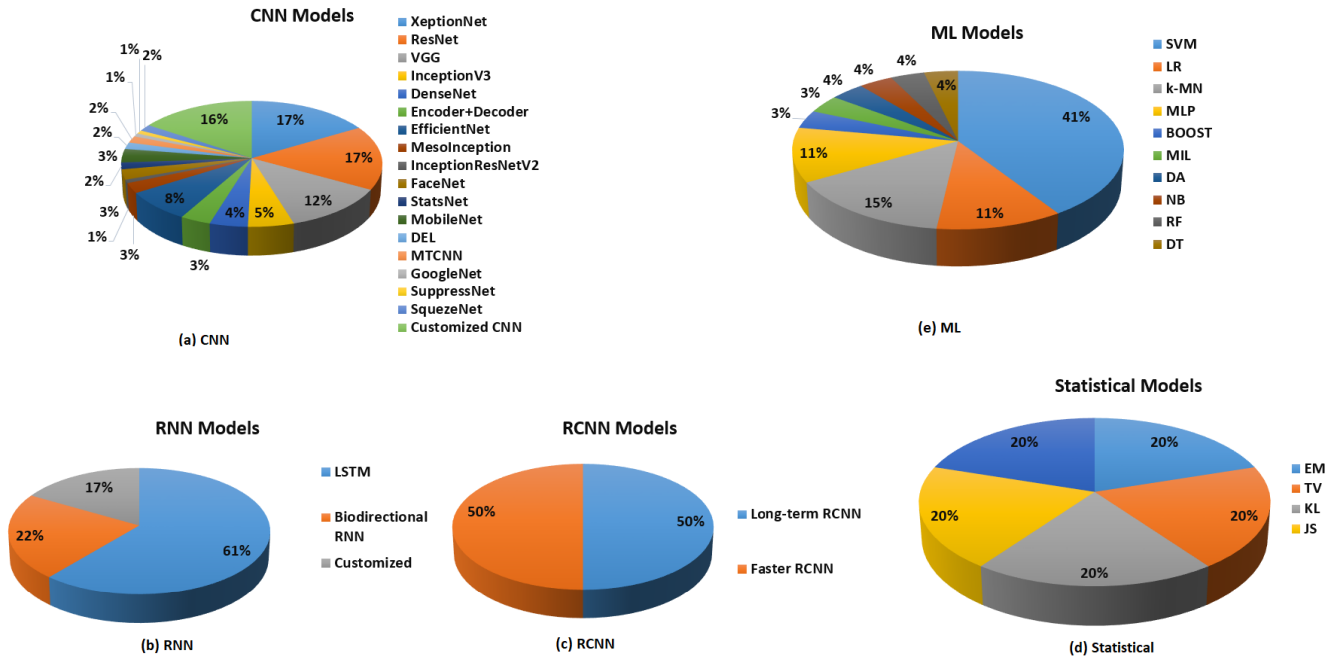

FIGURE 9. The allocation of subcategories of detection models. ML: Machine Learning; DL: Deep Learning.

TABLE 8. Measurement metrics used by various studies.

Metrics	Definition	#Studies
Accuracy (AC)	AC = (TN + TP) / (TN + FN + FP + TP)	62
Area Under Curve (AUC)	AUC provides an aggregate measure of performance across all possible classification thresholds	49
Receiver Operating Characteristic (ROC)	ROC is plotted with recall values on the y-axis and the 1-specificity values on the x-axis	17
Error Rate (ER)	Error rate = (FP + FN) / (TN + FN + FP + TP)	12
Precision (P)	Precision = TP / (TP + FP)	9
Recall/Sensibility/ True Positive Rate (TPR)	Recall = TP / (TP + FN)	8
Loss (Log)	It considers the uncertainty of the prediction based on how much it varies from the actual label	7
F-measure (f1)	F1=2 * Recall * Precision / (Recall + Precision)	6
False Positive Rate (FPR)	FPR= FP / (FP + TN)	3
Correlation coefficient	MCC, normalized cross-correlation, t-Test	2
Mean Absolute Error (MAE)	MAE measures the average magnitude of the errors in a set of predictions, without considering their direction	1
t-Distributed Stochastic Neighbor Embedding (t-SNE)	It converts similarities between data points to joint probabilities. It tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data	1
Frechet Inception Distance (FID)	A method for measuring the quality of generated image samples	1

Guera and Delp [66] discovered intra-frame and temporal inconsistencies among the Deepfake videos' frames. They proposed a network composed of CNN and LSTM to detect such discrepancies in Deepfakes. In this architecture, CNN handles extracting the frame-level features and LSTM to use these features as input to generate a descriptor accountable for analyzing the temporal sequence. Besides, to use physical indications [35]–[36], for example, eye blinking as features in detecting Deepfake, Li, *et al.* [46] proposed a long-term recurrent convolutional network (LRCN). Their method highlighted that the total eye blinking of an individual in Deepfake videos is always lower than in real videos. It can easily extract from the eye areas based on six eye landmarks and use them as features.

On the other hand, Rana and Sung [65] proposed a deep ensemble learning strategy, namely, DeepfakeStack, to detect Deepfake by analyzing multiple deep learning models. The concept behind DeepfakeStack is to train a meta-learner to top base-learners with pre-trained experience. It provides an interface for fitting the meta-learners on the base learners' prediction and demonstrates how an ensemble method executes the role of classification. The DeepfakeStack

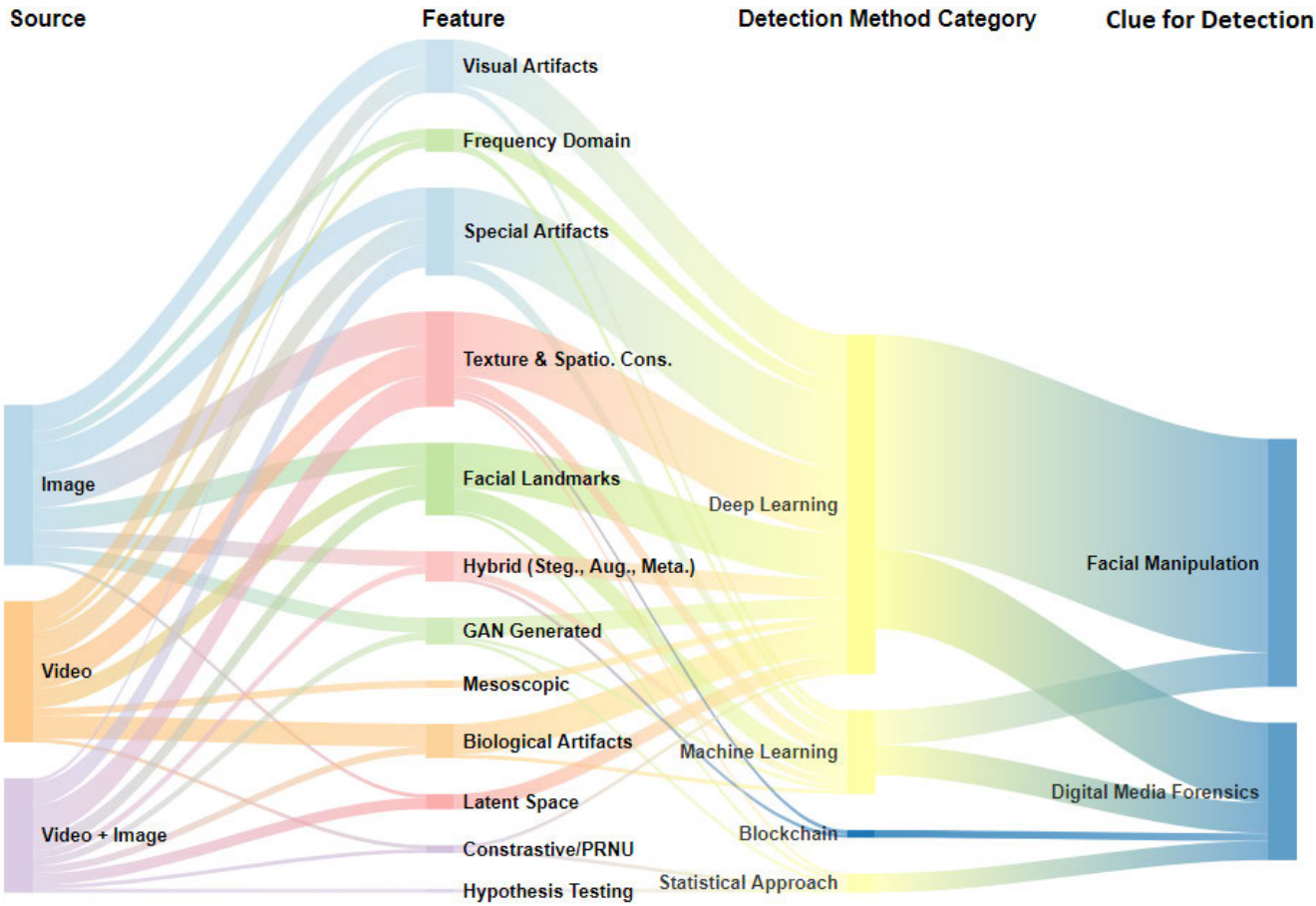

FIGURE 10. Taxonomy of Deepfake detection techniques. This taxonomy classifies the detection algorithms according to the media (image, video, or image and video), the features used (among the 12 features), the detection method (DL, ML, Blockchain, or statistical), and the clue for the detection (facial manipulation of digital media forensics, or other indications). The size of the connection line reflects the relative count of papers.

architecture includes multiple base-learners, the level-0 model, and a meta-learner, a level-1 model. The experiment reveals the DeepfakeStack achieves 99.65% accuracy and 1.0 of AUROC.

To some extent, these deep learning methods are complementary. In practice, combining multiple deep learning methods could obtain improved results compared to a single process. For example, the DeepfakeStack [65] integrates multiple state-of-the-art classification algorithms focused on deep learning and produces a sophisticated composite classifier that achieves 99.65%. Based on the RQ-1, it is seen that a maximum number of studies have applied deep learning techniques for detecting Deepfake. Therefore, it may be appropriate to explore the compatibility of deep learning methods and integrate some of them for further progress in Deepfake detection.

B. DEEP LEARNING-BASED METHODS ARE RECOMMENDED IN DEEPFAKE DETECTION

Compared with the traditional machine learning approaches, we note that applying deep learning algorithms to detect Deepfake from SRQ-2.3 has become a hot subject. We also find that most studies follow a traditional CNN approach to classify Deepfake in the deep-learning environment. Still, researchers have not yet figured out how to determine Deepfake authorship.

Based on the outcomes of RQ-4, it is observed that the deep learning-based models achieve better performance than the non-deep learning models in Deepfake detection. Therefore, deep learning-based approaches are advised when detecting Deepfake.

TABLE 9. Performance of various detection methods.

Category	Metrics	#Studies	Min	Max	Mean	STD
Deep Learning	Accuracy	50	63.15	100.0	89.73	10.08
	AUC	37	0.572	1.000	0.917	0.114
	Recall	5	82.74	100.0	89.47	12.88
	Precision	6	90.55	100.0	88.89	4.948
Machine Learning	Accuracy	12	85.00	91.07	86.86	11.04
	AUC	12	0.531	1.000	0.909	0.127
	Recall	2	82.74	92.11	89.92	10.15
	Precision	2	90.55	96.40	93.48	4.137

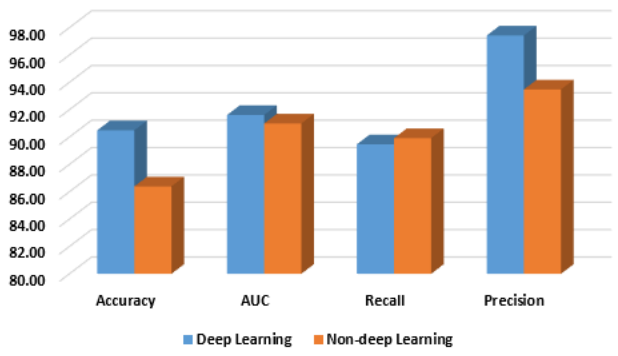

FIGURE 11. The comparison of the results among deep learning and non-deep learning based models.

C. A UNIQUE FRAMEWORK IS REQUIRED FOR THE FAIR EVALUATION OF DIFFERENT HETEROGENEOUS DEEPFAKE DETECTION METHODS

After reviewing the studies listed above, we note that several studies have used other datasets. Secondly, there are also differences with specific experiments that use the same dataset. (1) The measurement metrics used in the studies in question are not standard. For example, some experiments evaluate the performance of detection tasks using Accuracy and AUROC. Some studies use precision and recall only; (2) In these studies, it is also seen that the dataset's size is not consistent. For example, the FF++ dataset has 1000 Deepfake videos, but a few studies use the entire dataset while others use half. Some studies use only 100 videos; (3) The initial videos in these experiments are hardly available in public. The above conditions may lessen the trustworthiness of these RQ-3 and RQ-4 findings.

Based on the above circumstances, this section concludes that creating a unique framework for the fair assessment of the performance is essential.

V. LIMITATIONS AND CHALLENGES

This section will discuss some limitations and challenges that we observed during the preparation of this SLR.

A. CONSTRUCT VALIDITY

It is related to the collection of studies. We compile the associated articles from journals, seminars, conferences, workshops, and archives of many electronic libraries in this SLR. It is still possible that some of the related papers might still be missing from our collection of studies. Further, we might have a few mistakes sorting these experiments through the selection or rejection parameters we used in the process. We evaluated our catalog of studies using a double-checking approach to address such errors.

B. INTERNAL VALIDITY

The internal validity is related to data extraction and analysis. The present work involved an intense workload of data extraction and data processing. The cross-checking process is applied to the collected data, and we retrieved the final data after we agreed on the comparative results. Nevertheless, errors may still be present in how we collected and processed data. We believe the original authors could cross-check the reported results to avoid any unexpected error.

C. EXTERNAL VALIDITY

It is about the summary of the results obtained from various studies that we considered. To improve the quality of the findings in RQ-3 and RQ-4 in future studies, we recommend setting up a unique framework to reduce the inconsistencies in the results reported. Besides, more Deepfake detection experiments might be required to be obtained to produce definitive and systematic outcomes.

VI. CONCLUSION

This SLR presents various state-of-the-art methods for detecting Deepfake published in 112 studies from the beginning of 2018 to the end of 2020. We present basic techniques and discuss different detection models' efficacy in this work. We summarize the overall study as follows:

- The deep learning-based methods are widely used in detecting Deepfake.
- In the experiments, the FF++ dataset occupies the largest proportion.
- The deep learning (mainly CNN) models hold a significant percentage of all the models.
- The most widely used performance metric is detection accuracy.
- The experimental results demonstrate that deep learning techniques are effective in detecting Deepfake. Further, it can be stated that, in general, the deep learning models outperform the non-deep learning models.

With the rapid progress in underlying multimedia technology and the proliferation of tools and applications, Deepfake detection still faces many challenges. We hope this SLR provides a valuable resource for the research community in developing effective detection methods and countermeasures.

REFERENCES

[1] *FaceApp*. Accessed: Jan. 4, 2021. [Online]. Available: https://www.faceapp.com/
[2] *FakeApp*. Accessed: Jan. 4, 2021. [Online]. Available: https://www.fakeapp.org/
[3] G. Oberoi. *Exploring DeepFakes*. Accessed: Jan. 4, 2021. [Online]. Available: https://goberoi.com/exploring-deepfakes-20c9947c22d9
[4] J. Hui. *How Deep Learning Fakes Videos (Deepfake) and How to Detect it*. Accessed: Jan. 4, 2021. [Online]. Available: https://medium.com/how-deep-learning-fakes-videos-deepfakes-and-how-to-detect-it-c0b50fbf7cb9
[5] I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.
[6] G. Patrini, F. Cavalli, and H. Ajder, "The state of deepfakes: Reality under attack," Deeptrace B.V., Amsterdam, The Netherlands, Annu. Rep. v.2.3., 2018. [Online]. Available: https://s3.eu-west-2.amazonaws.com/rep2018/2018-the-state-of-deepfakes.pdf

[7] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2387–2395, doi: 10.1109/CVPR.2016.262.
[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Oct. 2017, pp. 2242–2251, doi: 10.1109/ICCV.2017.244.
[9] S. Suwajanakorn, S. M. Seitz, and I. K. Shlizerman, "Synthesizing Obama: Learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, p. 95, 2017.
[10] L. Matsakis. *Artificial Intelligence is Now Fighting Fake Porn*. Accessed: Jan. 4, 2021. [Online]. Available: https://www.wired.com/story/gfycat-artificial-intelligence-deepfakes/
[11] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: A large-scale video dataset for forgery detection in human faces," 2018, *arXiv:1803.09179*.
[12] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–14, Aug. 2018, doi: 10.1145/3197517.3201283.
[13] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," 2018, *arXiv:1808.07371*.
[14] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4396–4405, doi: 10.1109/CVPR.2019.00453.
[15] D. Budgen and P. Brereton, "Performing systematic literature reviews in software engineering," in *Proc. 28th Int. Conf. Softw. Eng.*, New York, NY, USA, May 2006, pp. 1051–1052, doi: 10.1145/1134285.1134500.
[16] Z. Stapic, E. G. Lopez, A. G. Cabot, L. M. Ortega, and V. Strahonja, "Performing systematic literature review in software engineering," in *Proc. 23rd Central Eur. Conf. Inf. Intell. Syst. (CECIIS)*, Varazdin, Croatia, Sep. 2012, pp. 441–447.
[17] B. Kitchenham, "Procedures for performing systematic reviews," Softw. Eng. Group; Nat. ICT Aust., Keele; Eversleigh, Keele Univ., Keele, U.K., Tech. Rep. TR/SE-0401; NICTA Tech. Rep. 0400011T.1, 2004.
[18] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Softw. Eng. Group; Keele Univ., Durham University Joint, Durham, U.K., Tech. Rep. EBSE-2007-01, 2007.
[19] M. A. Babar and H. Zhang, "Systematic literature reviews in software engineering: Preliminary results from interviews with researchers," in *Proc. 3rd Int. Symp. Empirical Softw. Eng. Meas.*, Lake Buena Vista, FL, USA, Oct. 2009, pp. 346–355, doi: 10.1109/ESEM.2009.5314235.
[20] H. Do, S. Elbaum, and G. Rothermel, "Supporting controlled experimentation with testing techniques: An infrastructure and its potential impact," *Empirical Softw. Eng.*, vol. 10, no. 4, pp. 405–435, 2005.
[21] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Waikoloa Village, HI, USA, Jan. 2019, pp. 83–92, doi: 10.1109/WACVW.2019.00020.
[22] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 15, 2020, doi: 10.1109/TPAMI.2020.3009287.
[23] X. Li, Y. Lang, Y. Chen, X. Mao, Y. He, S. Wang, H. Xue, and Q. Lu, "Sharp multiple instance learning for deepfake video detection," 2020, *arXiv:2008.04585*.
[24] L. Guarnera, O. Giudice, and S. Battiato, "Fighting deepfake by exposing the convolutional traces on images," 2020, *arXiv:2008.04095*.
[25] M. Bonomi, C. Pasquini, and G. Boato, "Dynamic texture analysis for detecting fake faces in video sequences," 2020, *arXiv:2007.15271*.
[26] L. Guarnera, O. Giudice, and S. Battiato, "Fighting deepfake by exposing the convolutional traces on images," 2020, *arXiv:2008.04095*.
[27] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 8261–8265, doi: 10.1109/ICASSP.2019.8683164.
[28] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Long Beach, CA, USA, Jun. 2019, pp. 1–8.
[29] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, "Unmasking DeepFakes with simple features," 2019, *arXiv:1911.00686*.
[30] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1831–1839, doi: 10.1109/CVPRW.2017.229.
[31] K. Songsri-in and S. Zafeiriou, "Complement face forensic detection and localization with faciallandmarks," 2019, *arXiv:1910.05455*.
[32] A. Kumar and A. Bhavsar, "Detecting deepfakes with metric learning," 2020, *arXiv:2003.08645*.
[33] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in GAN fake images," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2019, pp. 1–6.
[34] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1053–1061, doi: 10.1109/CVPR.2018.00116.
[35] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other- audio-visual dissonance-based deepfake detection and localization," 2020, *arXiv:2005.14405*.
[36] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, Y. Liu, and J. Zhao, "DeepRhythm: Exposing deepfakes with attentional visual heartbeat rhythms," 2020, *arXiv:2006.07634*.
[37] S. Fernandes, S. Raj, E. Ortiz, I. Vintila, M. Salter, G. Urosevic, and S. Jha, "Predicting heart rate variations of deepfake videos using neural ODE," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1721–1729.
[38] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, and A. Morales, "DeepFakesON-phys: DeepFakes detection based on heart rate estimation," 2020, *arXiv:2010.00400*.
[39] J. Bappy, C. Simons, L. Nataraj, B. Manjunath, and A. R. Chowdhury, "Hybrid LSTM and encoder–decoder architecture for detection of image forgeries," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3286–3300, Jul. 2019.
[40] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.
[41] P. Kawa and P. Syga, "A note on deepfake detection with low-resources," 2020, *arXiv:2006.05183*.
[42] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.
[43] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.
[44] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.
[45] A. Khodabakhsh and C. Busch, "A generalizable deepfake detector based on neural conditional distribution modelling," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, Sep. 2020, pp. 1–5.
[46] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.
[47] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, 2019, pp. 46–52. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2019/html/Media_Forensics/Li_Exposing_DeepFake_Videos_By_Detecting_Face_Warping_Artifacts_CVPRW_2019_paper.html
[48] I. Ganiyusufoglu, L. Minh Ngô, N. Savov, S. Karaoglu, and T. Gevers, "Spatio-temporal features for generalized detection of deepfake videos," 2020, *arXiv:2010.11844*.
[49] A. Singh, A. S. Saimbhi, N. Singh, and M. Mittal, "Deepfake video detection: A time-distributed approach," *SN Comput. Sci.*, vol. 1, p. 212, Jun. 2020, doi: 10.1007/s42979-020-00225-9.
[50] I. Kukanov, J. Karttunen, H. Sillanpää, and V. Hautamäki, "Cost sensitive optimization of deepfake detector," 2020, *arXiv:2012.04199*.
[51] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," 2020, *arXiv:2012.07657*.

[52] X. Zhu, H. Wang, H. Fei, Z. Lei, and S. Z. Li, "Face forgery detection by 3D decomposition," 2020, *arXiv:2011.09737*.

[53] X. Wang, T. Yao, S. Ding, and L. Ma, "Face manipulation detection via auxiliary supervision," in *Neural Information Processing (ICONIP)* (Lecture Notes in Computer Science), vol. 12532, H. Yang, K. Pasupa, A. C. Leung, J. T. Kwok, J. H. Chan, I. King, Eds. Cham, Switzerland: Springer, 2020, pp. 313–324, doi: 10.1007/978-3-030-63830-6_27.

[54] M. T. Jafar, M. Ababneh, M. Al-Zoube, and A. Elhassan, "Forensics and analysis of deepfake videos," in *Proc. 11th Int. Conf. Inf. Commun. Syst. (ICICS)*, Irbid, Jordan, Apr. 2020, pp. 053–058, doi: 10.1109/ICICS49469.2020.239493.

[55] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo, "Identity-driven deepfake detection," 2020, *arXiv:2012.03930*.

[56] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," 2020, *arXiv:2012.09311*.

[57] L. Bondi, E. Daniele Cannas, P. Bestagini, and S. Tubaro, "Training strategies and data augmentations in CNN-based deepfake video detection," 2020, *arXiv:2011.07792*.

[58] Z. Hongmeng, Z. Zhiqiang, S. Lei, M. Xiuqing, and W. Yuehan, "A detection method for deepfake hard compressed videos based on super-resolution reconstruction using CNN," in *Proc. 4th High Perform. Comput. Cluster Technol. Conf. 3rd Int. Conf. Big Data Artif. Intell.*, New York, NY, USA, Jul. 2020, pp. 98–103, doi: 10.1145/3409501.3409542.

[59] J. Han and T. Gevers, "MMD based discriminative learning for face forgery detection," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2020, pp. 121–136.

[60] L. Verdoliva, "Media forensics and DeepFakes: An overview," 2020, *arXiv:2001.06564*.

[61] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 5780–5789, doi: 10.1109/CVPR42600.2020.00582.

[62] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 2307–2311, doi: 10.1109/ICASSP.2019.8682602.

[63] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a capsule network to detect fake images and videos," 2019, *arXiv:1910.12467*.

[64] N. Bonettini, E. Daniele Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," 2020, *arXiv:2004.07676*.

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

[66] D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.

[67] S. Sohrawardi, A. Chintha, B. Thai, S. Seng, A. Hickerson, R. Ptucha, and M. Wright, "Poster: Towards robust open-world detection of deepfakes," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2019, pp. 2613–2615.

[68] E. Sabir, J. Cheng, A. Jaiswal, W. Abd-Almageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," in *Proc. CVPR Workshops*, 2019, pp. 80–87.

[69] S. Tariq, S. Lee, and S. S. Woo, "A convolutional LSTM based residual network for deepfake video detection," 2020, *arXiv:2009.07480*.

[70] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. Abd-Almageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Proc. 16th Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 667–684.

[71] A. Chintha, B. Thai, S. J. Sohrawardi, K. Bhatt, A. Hickerson, M. Wright, and R. Ptucha, "Recurrent convolutional structures for audio spoof and video deepfake detection," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 1024–1037, Aug. 2020, doi: 10.1109/JSTSP.2020.2999185.

[72] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1205–1207.

[73] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "ForensicTransfer: Weakly-supervised domain adaptation for forgery detection," 2018, *arXiv:1812.02510*.

[74] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *Proc. IEEE 10th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2019, pp. 1–8.

[75] M. Du, S. Pentyala, Y. Li, and X. Hu, "Towards generalizable deepfake detection with locality-aware autoencoder," 2019, *arXiv:1909.05999*.

[76] L. Trinh, M. Tsang, S. Rambhatla, and Y. Liu, "Interpretable and trustworthy deepfake detection via dynamic prototypes," 2020, *arXiv:2006.15473*.

[77] M. Du, S. Pentyala, Y. Li, and X. Hu, "Towards generalizable deepfake detection with locality-aware autoencoder," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, New York, NY, USA, Oct. 2020, doi: 10.1145/3340531.3411892.

[78] T. Fernando, C. Fookes, S. Denman, and S. Sridharan, "Exploiting human social cognition for the detection of fake and fraudulent faces via memory networks," 2019, *arXiv:1911.07844*.

[79] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," 2019, *arXiv:1912.13458*.

[80] Z. Chen and H. Yang, "Attentive semantic exploring for manipulated face detection," 2020, *arXiv:2005.02958*.

[81] T. D. Nhu, I. S. Na, H. J. Yang, G. S. Lee, and S. H. Kim, "Forensics face detection from GANs using convolutional neural network," in *Proc. Int. Symp. Inf. Technol. Converg. (ISITC)*, 2018, pp. 1–5.

[82] 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, Dec. 2020, doi: 10.1186/s13635-020-00109-8.

[83] Z. Guo, G. Yang, J. Chen, and X. Sun, "Fake face detection via adaptive manipulation traces extraction network," 2020, *arXiv:2005.04945*.

[84] D. Mas Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horváth, E. Bartusiak, J. Yang, D. Güera, F. Zhu, and E. J. Delp, "Deepfakes detection with automatic face weighting," 2020, *arXiv:2004.12027*.

[85] L. M. Dang, S. I. Hassan, S. Im, and H. Moon, "Face image manipulation detection based on a convolutional neural network," *Expert Syst. Appl.*, vol. 129, pp. 156–168, Sep. 2019.

[86] Z. Liu, X. Qi, J. Jia, and P. H. S. Torr, "Real or fake: An empirical study and improved model for fake face detection," in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, Apr. 2020, pp. 1–12.

[87] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 7887–7896, doi: 10.1109/CVPR42600.2020.00791.

[88] M. A. S. Habeeba, A. Lijiya, and A. M. Chacko, "Detection of deepfakes using visual artifacts and neural network classifier," in *Innovations in Electrical and Electronic Engineering* (Lecture Notes in Electrical Engineering), vol. 661, M. Favorskaya, S. Mekhilef, R. Pandey, and N. Singh, Eds. Singapore: Springer, 2020, pp. 411–422, doi: 10.1007/978-981-15-4692-1_31.

[89] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, "Deep fake image detection based on pairwise learning," *Appl. Sci.*, vol. 10, no. 1, p. 370, Jan. 2020, doi: 10.3390/app10010370.

[90] S. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot…for now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 8692–8701, doi: 10.1109/CVPR42600.2020.00872.

[91] P. Korshunov and S. Marcel, "DeepFakes: A new threat to face recognition? Assessment and detection," 2018, *arXiv:1812.08685*.

[92] A. Gandhi and S. Jain, "Adversarial perturbations fool deepfake detectors," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[93] K. Zhu, B. Wu, and B. Wang, "Deepfake detection with clustering-based embedding regularization," in *Proc. IEEE 5th Int. Conf. Data Sci. Cyberspace (DSC)*, Hong Kong, Jul. 2020, pp. 257–264, doi: 10.1109/DSC50466.2020.00046.

[94] P. Charitidis, G. Kordopatis-Zilos, S. Papadopoulos, and I. Kompatsiaris, "Investigating the impact of pre-processing and prediction aggregation on the DeepFake detection task," 2020, *arXiv:2006.07084*.

[95] P. Charitidis, G. Kordopatis-Zilos, S. Papadopoulos, and I. Kompatsiaris, "Investigating the impact of pre-processing and prediction aggregation on the deepfake detection task," 2020, *arXiv:2006.07084*.

[96] X. Li, K. Yu, S. Ji, Y. Wang, C. Wu, and H. Xue, "Fighting against deepfake: Patch&pair convolutional neural networks (PPCNN)," in *Proc. Companion Web Conf.*, New York, NY, USA, 2020, pp. 88–89, doi: 10.1145/3366424.3382711.
[97] 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 *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Canberra, ACT, Australia, Dec. 2020, pp. 707–712.
[98] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, "ID-reveal: Identity-aware DeepFake video detection," 2020, *arXiv:2012.02512*.
[99] 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, 2020, doi: 10.3390/e22020249.
[100] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, Oct. 2020, doi: 10.1145/3394171.3413570.
[101] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "DeepFake detection based on discrepancies between faces and their context," 2020, *arXiv:2008.12262*.
[102] C. M. Yu, C. T. Chang, and Y. W. Ti, "Detecting deepfake-forged contents with separable convolutional neural network and image segmentation," 2019, *arXiv:1912.12184*.
[103] D. Feng, X. Lu, and X. Lin, "Deep detection for face manipulation," 2020, *arXiv:2009.05934*.
[104] L. Chai, D. Bau, S. Lim, and P. Isola, "What makes fake images detectable? Understanding properties that generalize," in *Proc. 16th Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 103–120.
[105] X. Chang, J. Wu, T. Yang, and G. Feng, "DeepFake face image detection based on improved VGG convolutional neural network," in *Proc. 39th Chin. Control Conf. (CCC)*, Shenyang, China, Jul. 2020, pp. 7252–7256, doi: 10.23919/CCC50068.2020.9189596.
[106] U. Aybars Ciftci, I. Demir, and L. Yin, "How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals," 2020, *arXiv:2008.11363*.
[107] H. M. Nguyen and R. Derakhshani, "Eyebrow recognition for identifying deepfake videos," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, Sep. 2020, pp. 1–5.
[108] M. Koopman, A. M. Rodriguez, and Z. Geradts, "Detection of deepfake video manipulation," in *Proc. 20th Irish Mach. Vis. Image Process. Conf. (IMVIP)*, London, U.K., 2018, pp. 1–4.
[109] B. L. Welch, "The generalization of students' problem when several different population variances are involved," *Biometrika*, vol. 34, nos. 1–2, pp. 28–35, 1947.
[110] L. Guarnera, O. Giudice, and S. Battiato, "DeepFake detection by analyzing convolutional traces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2020, pp. 2841–2850, doi: 10.1109/CVPRW50498.2020.00341.
[111] S. Agarwal and L. R. Varshney, "Limits of deepfake detection: A robust estimation viewpoint," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, 2019.
[112] U. M. Maurer, "Authentication theory and hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1350–1356, Jul. 2000, doi: 10.1109/18.850674.
[113] H. Hasan and K. Salah, "Combating deepfake videos using blockchain and smart contracts," *IEEE Access*, vol. 7, pp. 41596–41606, 2019, doi: 10.1109/ACCESS.2019.2905689.
[114] *IPFS Powers the Distributed Web*. Accessed: Jun. 5, 2020. [Online]. Available: https://ipfs.io/
[115] C. C. Ki Chan, V. Kumar, S. Delaney, and M. Gochoo, "Combating deepfakes: Multi-LSTM and blockchain as proof of authenticity for digital media," in *Proc. IEEE/ITU Int. Conf. Artif. Intell. Good (AI4G)*, Sep. 2020, pp. 55–62.
[116] J. Li, T. Shen, W. Zhang, H. Ren, D. Zeng, and T. Mei, "Zooming into face forensics: A pixel-level analysis," 2019, *arXiv:1912.05790*.
[117] T. Thi Nguyen, Q. Viet Hung Nguyen, D. Tien Nguyen, D. Thanh Nguyen, T. Huynh-The, S. Nahavandi, T. Tam Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," 2019, *arXiv:1909.11573*.
[118] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020, doi: 10.1016/j.inffus.2020.06.014.
[119] R. Tolosana, S. Romero-Tapiador, J. Fierrez, and R. Vera-Rodriguez, "DeepFakes evolution: Analysis of facial regions and fake detection performance," 2020, *arXiv:2004.07532*.
[120] S. Lyu, "Deepfake detection: Current challenges and next steps," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, London, U.K., Jul. 2020, pp. 1–6, doi: 10.1109/ICMEW46912.2020.9105991.
[121] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," 2020, *arXiv:2004.11138*.
[122] L. Guarnera, O. Giudice, C. Nastasi, and S. Battiato, "Preliminary forensics analysis of deepfake images," 2020, *arXiv:2004.12626*.
[123] A. O. J. Kwok and S. G. M. Koh, "Deepfake: A social construction of technology perspective," *Current Issues Tourism*, vol. 24, no. 13, pp. 1798–1802, 2020, doi: 10.1080/13683500.2020.1738357.
[124] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, "Deepfakes: Trick or treat?" *Bus. Horizons*, vol. 63, no. 2, pp. 135–146, Mar. 2020, doi: 10.1016/j.bushor.2019.11.006.
[125] J. Frank, T. Eisenhofer, L. Schonherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, Jul. 2020, pp. 3247–3258.
[126] M.-H. Maras and A. Alexandrou, "Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos," *Int. J. Evidence Proof*, vol. 23, no. 3, pp. 255–262, Jul. 2019, doi: 10.1177/1365712718807226.
[127] M. Westerlund, "The emergence of deepfake technology: A review," *Technol. Innov. Manage. Rev.*, vol. 9, no. 11, pp. 40–53, 2019, doi: 10.22215/timreview/1282.
[128] C. Öhman, "Introducing the pervert's dilemma: A contribution to the critique of deepfake pornography," *Ethics and Inf. Technol.*, vol. 22, pp. 133–140, Nov. 2020, doi: 10.1007/s10676-019-09522-1.
[129] R. Thakur and R. Rohilla, "Recent advances in digital image manipulation detection techniques: A brief review," *Forensic Sci. Int.*, vol. 312, Jul. 2020, Art. no. 110311, doi: 10.1016/j.forsciint.2020.110311.
[130] N. Carlini and H. Farid, "Evading deepfake-image detectors with white- and black-box attacks," 2020, *arXiv:2004.00622*.
[131] A. Pishori, B. Rollins, N. van Houten, N. Chatwani, and O. Uraimov, "Detecting deepfake videos: An analysis of three techniques," 2020, *arXiv:2007.08517*.
[132] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, "Deepfake detection using spatiotemporal convolutional networks," 2020, *arXiv:2006.14749*.
[133] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley, "Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples," 2020, *arXiv:2002.12749*.
[134] P. Korshunov and S. Marcel, "Deepfake detection: Humans vs. machines," 2020, *arXiv:2009.03155*.
[135] H. U. U. Chi Maduakor and R. E. Alo Williams, "Integrating deepfake detection into cybersecurity curriculum," in *Proc. Future Technol. Conf. (FTC)* (Advances in Intelligent Systems and Computing), vol. 1288, K. Arai, S. Kapoor, and R. Bhatia, Eds. Cham, Switzerland: Springer, 2020, pp. 588–598, doi: 10.1007/978-3-030-63128-4_45.
[136] *Contributing Data to Deepfake Detection Research*. Accessed: Jan. 4, 2021. [Online]. Available: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html
[137] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 3730–3738, doi: 10.1109/ICCV.2015.425.
[138] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," 2019, *arXiv:1909.12962*.
[139] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.
[140] L. Jiang, R. Li, W. Wu, C. Qian, and C. Change Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," 2020, *arXiv:2001.03024*.
[141] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," in *Proc. 28th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2020, pp. 2382–2390, doi: 10.1145/3394171.3413769.
[142] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch, "Fake face detection methods: Can they be generalized?" in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, Sep. 2018, pp. 1–6, doi: 10.23919/BIOSIG.2018.8553251.
[143] P. Gupta, K. Chugh, A. Dhall, and R. Subramanian, "The eyes know it: FakeET—An eye-tracking database to understand deepfake perception," 2020, *arXiv:2006.06961*.

[144] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5074–5083.
[145] Y. Pan, X. Ge, C. Fang, and Y. Fan, "A systematic literature review of Android malware detection using static analysis," *IEEE Access*, vol. 8, pp. 116363–116379, 2020, doi: 10.1109/ACCESS.2020.3002842.
[146] L. Li, T. F. Bissyandé, M. Papadakis, S. Rasthofer, A. Bartel, D. Octeau, J. Klein, and L. Traon, "Static analysis of Android apps: A systematic literature review," *Inf. Softw. Technol.*, vol. 88, pp. 67–95, Aug. 2017, doi: 10.1016/j.infsof.2017.04.001.
[147] T. Baar, W. van Houten, and Z. Geradts, "Camera identification by grouping images from database, based on shared noise patterns," 2012, *arXiv:1207.2641*.

MD SHOHEL RANA (Member, IEEE) received the bachelor's and master's degrees in computer science and engineering from Mawlana Bhashani Science and Technology University, Bangladesh, and the Ph.D. degree in computational science from The University of Southern Mississippi, MS, USA, in 2021. He is currently a Visiting Assistant Professor of computer science at Northern Kentucky University, KY, USA. His research interests include digital image processing and computer vision, data mining and pattern recognition, machine learning, deep learning, cybersecurity, e-learning, distributed database, and blockchain.

MOHAMMAD NUR NOBI (Member, IEEE) received the bachelor's and master's degrees in computer science and engineering from Mawlana Bhashani Science and Technology University, Bangladesh. He is currently pursuing the Ph.D. degree with the Department of Computer Science, The University of Texas at San Antonio (UTSA). His research interests include cybersecurity, machine learning, computer vision, and medical image processing.

BEDDHU MURALI received the Ph.D. degree in aerospace engineering from Mississippi State University, in 1992. He is currently an Associate Professor of computing sciences and computer engineering (CSCE) at The University of Southern Mississippi, USA. His research interests include scientific computational algorithms, high-performance computing, image and video processing, robotics, and machine learning.

ANDREW H. SUNG (Member, IEEE) received the Ph.D. degree in computer science from the State University of New York at Stony Brook, USA, in 1984. He is currently a Professor of computing sciences and computer engineering (CSCE) at The University of Southern Mississippi, USA. His research interests include computational intelligence and its applications, information security and multimedia forensics, social network analysis, data mining and pattern recognition, and petroleum reservoir modeling.