Mesopotamian journal of Cybersecurity
Vol.4,No.3, **pp**. 233–250
DOI: https://doi.org/10.58496/MJCS/2024/025; ISSN: 2958-6542
https://mesopotamian.press/journals/index.php/cybersecurity

Research Article

Deepfake Detection in Video and Audio Clips: A Comprehensive Survey and Analysis

Wurood A. Jbara [1],*, Noor Al-Huda K.Hussein [2], Jamila H. Soud [1],

[1] *Department of Computer Science, College of Science, Mustansiriyah University, Baghdad, IRAQ.*

[2] *Computer Technology Engineering Department, Technical College, Imam Ja'afar Al-Sadiq University, Baghdad, IRAQ.*

ARTICLEINFO

Article history

Received 28 Sep 2024
Accepted 30 Nov 2024
Published 24 Dec 2024

Keywords

Deep Learning

Machine Learning

CNN

LSTM

GAN

ANN

ABSTRACT

Deepfake (DF) technology has emerged as a major concern due to its potential for misuse, including privacy violations, misinformation, and threats to the integrity of digital media. While significant progress has been made in developing deep learning (DL) algorithms to detect DFs, effectively distinguishing between real and manipulated content remains a challenge due to the rapid evolution of DF generation techniques. This study aims to address two key issues: the need for a comprehensive review of current DF detection methods and the challenge of achieving high detection accuracy with low computational cost. We conducted a systematic literature review to evaluate various DF detection algorithms, focusing on their performance, computational efficiency, and robustness. The review covers methods such as Convolutional Neural Networks (CNNs), Long Short Term Memory (LSTM) networks, hybrid models, and specialized approaches like spectral and phonetic analysis. Our findings reveal that while some methods achieve high accuracy, up to 94% in controlled environments, they often struggle to generalize across diverse DF applications. Hybrid models that combine CNNs and LSTMs typically offer a better balance between accuracy and computational efficiency. This paper provides valuable insights into the current state of DF detection and highlights the need for adaptive models that can effectively address the evolving challenges of DF generation.

1. INTRODUCTION

In recent years, the field of aartificial iintelligence (AI) and deep learning (DL) has undergone rapid advancements, leading to significant transformations across various industries, from healthcare to media and entertainment [1]. One of the AI application that has stirred up controversies in recent is deepfake (DF) technology, which helps to create highly authentic fake audio-visual content using AI techniques. While it has beneficial implications in areas like media and film-making, the malignant misuse of this technology flags distress [2]. They could include creating believable voices, changing videos to depict a different scene, or even staging entire events. Defamation, propaganda, false information and even threats to national security are a serious issue when applied maliciously to DFs. Such abuse underscores the critical need for robust ways to detect these high-level digital deceptions [3], [4]. Being able to detect DFs is incredibly difficult as the technology behind them iterates and evolves. Present discovery techniques continue to experience difficulty with the wide exhibit of deep fakes, regardless of whether they are in sound or video structure. The newer methods could tweak facial expressions, voice intonations and even background settings so well that it got imperceptible for whether the content was altered or not [5]. In addition, many of the existing detection techniques are resource-intensive but in most real-world settings resources are scarce [6].

The goal of this study, in this context, is to add to the body of knowledge a comprehensive contribution by a systematic literature review on DF detection methods which is scoped specifically in manipulated video and audio content [7], so by testing multiple algorithmic models including: Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and hybrid methods that use both, assessing models based on accuracy, computational cost, and ability to process diverse data. This review also identifies the research gaps and describes a few of those challenges should be resolved to elevate these detection techniques for efficient usages in future.

Although some DFs are still created using conventional visual effects or graphics programming techniques, the most recent and most highly prevalent method of creating spurious deep fakes involves DL models like autoencoders and GANs which have long been in use in computer vision. One of the most important advancements in the field of manipulating multimedia

*Corresponding author. Email: wo_abdulkarim@uomustansiriyah.edu.iq content this year has been due to advancements in artificial neural networks (ANNs) [8]. For example, AI-based software applications such as FaceApp [9] and FakeApp [10] allow realistic face-swapping in images or videos. These apps include features like changing looks, gender, age etc. The models reconstruct the facial expressions and movements of a person to obtain synthetic faces showing similar expressions and gestures [11]. A ton of image and video data are needed to train models who produce photo-realistic images and videos. DFs tend to focus mainly on famous personalities, including politicians and celebrities because they have more pictures and videos available in the public domain. This vast amount of data makes it possible for the models to get the practice they need and able to produce really great looking and very convincing DFs. Moreover, DFs of top officials gather media attention could be exploited to shape public opinion or spread misleading information [12].

DF technology has also been used to superimpose the faces of celebrities or politicians onto bodies in pornographic images and videos. Despite its potential for positive applications in virtual reality and filmmaking, DF technology is still widely misused for malicious purposes, including extortion and defamation [13], [14], [15]. However, DFs also have creative and productive applications in fields such as photography, video games, virtual reality, film production, and entertainment. For example, they can be used to create convincing video dubbing for foreign films, educate audiences by reanimating historical figures, or allow consumers to virtually try on products while shopping [16]. Figure 1 illustrates the large volume of fictitious recordings shared online, most of which target celebrities and politicians [17].

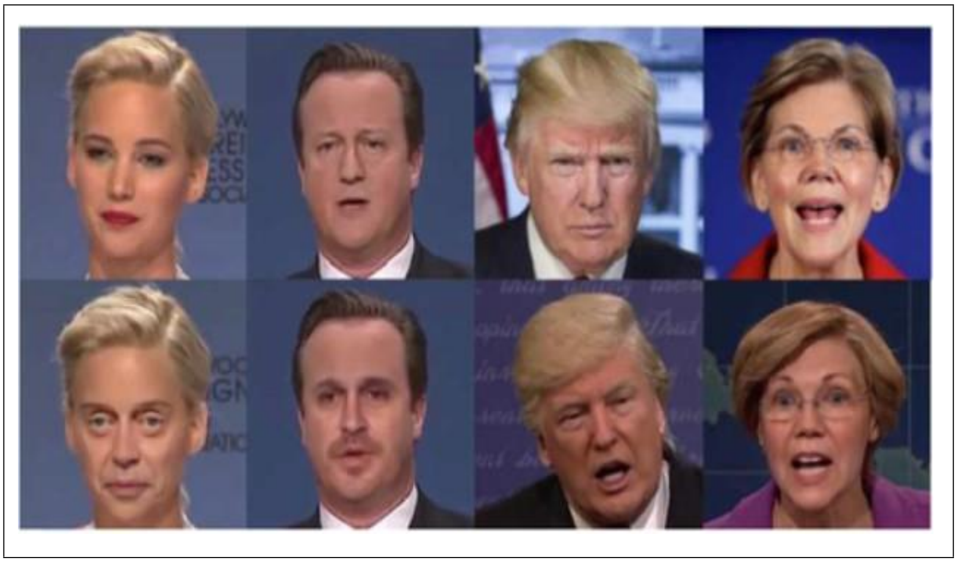

Fig.1 . Video frames generated by DF algorithms. The first line shows the original video frames and the second line shows the corresponding video frames generated by DF methods

Identifying DF material has emerged as a major concern for people, organizations, and governments globally. Even if there have been some improvements, there are still a lot of important problems with the DF detection techniques that need to be resolved. As DF techniques continue to advance, produced videos are becoming more realistic. In this instance, it’s unlikely that conventional techniques will be effective in identifying altered footage produced by novel DF algorithms [18]. It is important to evaluate and project how DF related research will evolve and to enhance relevant detection techniques. Truth finding in the digital realm has thus grown more crucial. Dealing with DFs is far more difficult because they are mostly utilized for malevolent intent and are now easily created by virtually anybody utilizing available DF tools [19]. Many techniques have been put out thus far to identify DFs

Since DL is the foundation of the majority of them, a conflict has emerged between the beneficial and harmful applications of DL techniques. The United States Defense Advanced Research Projects Agency (DARPA) launched a media forensics research program called Media Forensics, or Media Forensics++, to speed up the development of phony digital visual media detection techniques in response to the danger posed by face swapping technologies and DFs [20], [21], [22]. Recently, Facebook Inc. teaming up with Microsoft Corp and the Partnership on AI coalition have launched the DF Detection Challenge which aims to spur more research and development in identifying and preventing the use of DFs to deceive viewers [23]. At the end of 2023, data from https://app.dimensions.ai revealed a notable rise in the quantity of DF publications in the previous several years Figure.2. The number of DF articles acquired may not match the actual amount, however there is a clear upward trend in this topic study.

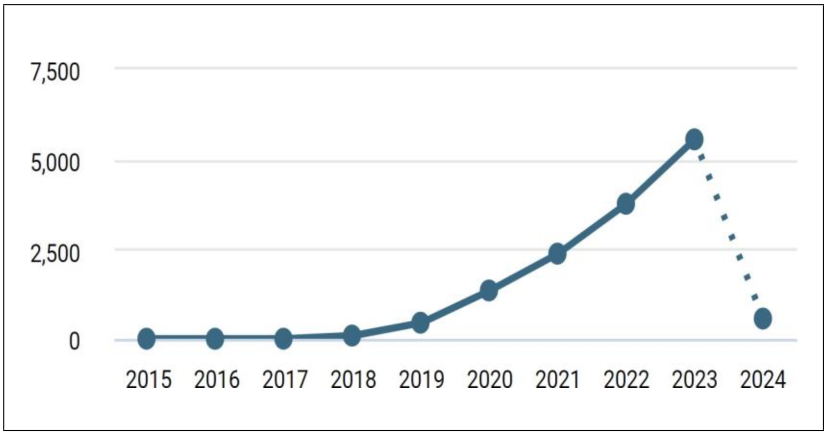

Fig.2 . The number of articles published between 2018 and 2023 about DFs was retrieved at the end of 2023 from https://app.dimensions.ai using the search term "DF" on the full text of academic publications.

This study seeks to achieve several key objectives, including: study the different strengths and weaknesses of the models which exist for detecting DF phenomena, evaluate their performances across varying contexts, critically look at methods based on DL that can be used to overcome these adversities brought by DFs and suggest some future directions to explore on. By addressing these objectives, the study aims to support scientific research and encourage further efforts toward the development of models capable of detecting digital manipulations with high accuracy and resource efficiency. The significance of this study lies in its ability to provide a deeper understanding of the mechanisms behind the generation and detection of DFs [24] . This understanding is crucial for preserving the integrity of digital content, especially in sectors that rely on trustworthy digital media, such as journalism, cybersecurity, and law enforcement. Moreover, the study contributes to the ongoing discourse on the ethics of AI use, offering important insights into the need for strict regulation and oversight to prevent the misuse of these advanced technologies.

2. DEEPFACK DETECTION

Deepfac detection refers to the process of identifying altered or created media using advanced ML techniques, specifically using DL models. The rapid advancement of DF technology has raised serious concerns about potential abuse in spreading misinformation. influence public opinion and privacy violations. In many cases, DF detection methods analyze various aspects. of media content to reveal any irregularities or patterns that indicate artificial data synthesis or transformation ML-based classification can include semantic feature analysis and signal level analysis. It depends on the type of media being examined [6][65]. Developing effective deep counterfeit detection capabilities is critical to preventing the spread of altered content and the associated risks. Continued cooperation between government agencies, industry, and academia is essential to advance the field and develop robust countermeasures against artificial media abuse [19].
DF detection is usually considered a binary classification problem. Its aim is to differentiate between real and fake videos using classifiers. This approach requires a large dataset of both real and fake videos to train the classification algorithm. Although more and more fake videos are circulating online, but it still lacks a comprehensive dataset that can be used as a benchmark to evaluate various detection techniques [20]. To address this issue, Korshunov and Marcel [25] coded the open source Faceswap-GAN DF on a large-scale dataset of 620 videos. The dataset was created using scenes from the publicly available VidTIMIT database to create high and low quality DFs [26]. These videos simulate facial expressions. lip movement and blinking in high-level reality Several DFs detection methods have been tested on this dataset, however, popular facial recognition algorithms such as VGG [27] and FaceNet [28] have shown limited success in detecting DFs Other methods such as lip synchronization techniques. Combined with support vector machines (SVM) and the quality assessment also gives a high error rate when identifying deep-level videos from this dataset [29], [30]. This highlights the urgent need for more reliable techniques to separate real videos from deep videos. In response to these challenges Researchers have turned to deep neural networks (DNN) for their powerful feature extraction capabilities. It aims at developing more accurate and reliable detection methods through implicit feature learning [31], [32]. Several studies used CNN and LSTM models to detect DF videos created by face swapping techniques [33]. They were able to do so. CNNs have also been proposed to detect facial distortion in videos [34]. Research other people have increased the detection flexibility with low-quality videos by introducing multi-head attention and fine-grained classification [35]. Liu et al. present

a semantic-based DF detection method using photoplethysmography (PPG), a biological signal that detects subtle changes in skin color due to facial blood flow [36].

Moreover, M. Suresha et al. [37] have made significant contributions to deep feature extraction from video frames using DNNs. N. Bonettini et al. [38] further explored the use of CNN ensembles for detecting face manipulations in videos. Meanwhile, H. Zhao et al. have advanced the field by applying multi-head attention and fine-grained classification to improve detection in low-quality videos. Liu et al. have also employed phase spectrum and frequency domain signal analysis to enhance the detection of DF videos.

Figure 3 provides a flowchart outlining the DF detection process, detailing the key steps from data collection to performance evaluation. It also illustrates the various analysis methods used in detecting DFs.

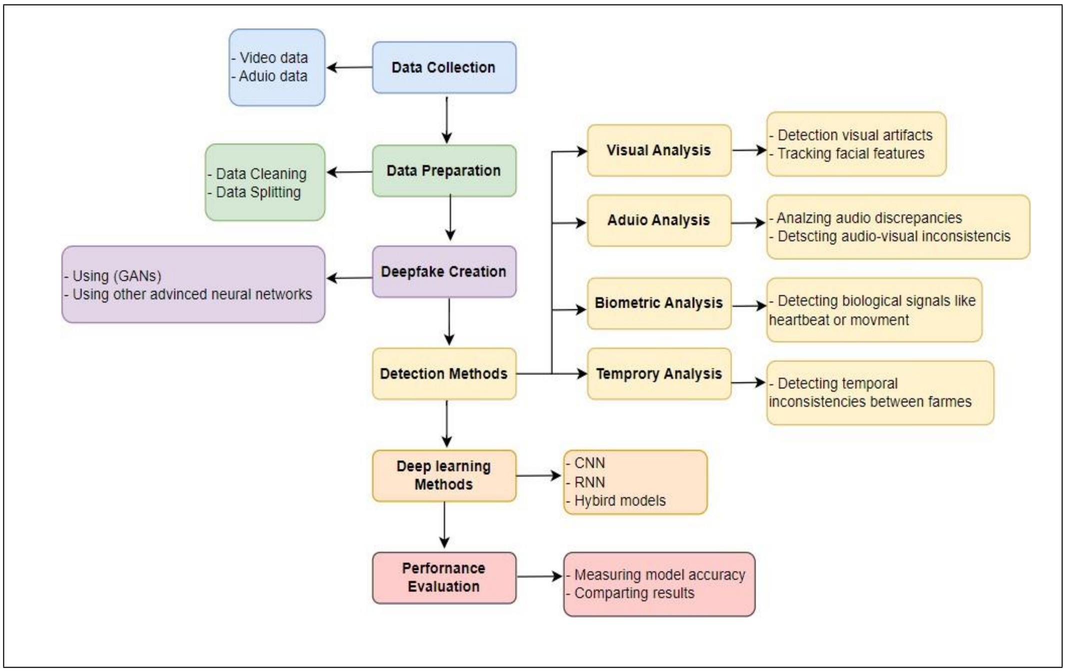
- Video data
- Aduio data
Data Collection
- Data Cleaning
- Data Splitting
Data Preparation
- Using (GANs)
- Using other advinced neural networks
Deepfake Creation
Detection Methods
Visual Analysis
- Detection visual artifacts
- Tracking facial features
Aduio Analysis
- Analzing audio discrepancies
- Detscting audio-visual inconsistencis
Biometric Analysis
- Detecting biological signals like heartbeat or movment
Temprory Analysis
- Detecting temporal inconsistencies between farmes
Deep learning Methods
- CNN
- RNN
- Hybird models
Performance Evaluation
- Measuring model accuracy
- Comparting results

Fig.3 . DFs and Methods for Detecting them

Sabir et al. [39]. used spatiotemporal properties of video streams to detect DFs, based on the finding that temporal coherence is not maintained well in the synthesis process of DFs. Because video modification is addressed frame-by-frame, it is thought that low level artifacts from face alterations will also show up as temporal artifacts that are inconsistent between frames. To take advantage of temporal differences between frames, a recurrent convolutional model (RCN) was presented, which is based on the integration of the convolutional network DenseNet [40] and the gated recurrent unit cells [41]see Figure 4 [42] Tested on the 1,000-video FaceForensics ++ dataset [33], the suggested approach yields encouraging results.

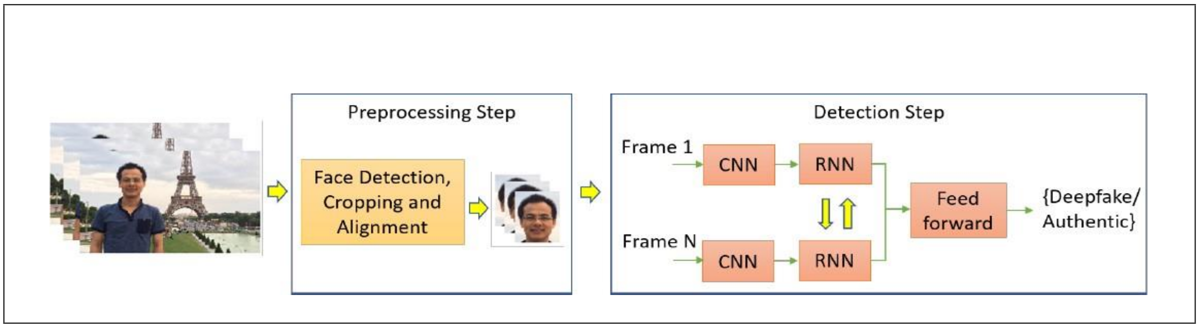

Fig.4 . a two-step procedure for detecting face manipulation that combines CNN and RNN in order to separate manipulated and real face images the preprocessing stage seeks to identify, crop, and align faces on a series of frames.

Table. 1 provides a comprehensive comparison of various methods used to detect DFs, emphasizing differences in feature extraction, performance, flexibility, computational requirements, and robustness. Modern methods leverage DL, which necessitates substantial computational resources and extensive data, whereas hybrid approaches optimize between traditional and modern techniques to achieve a balance of effectiveness and cost efficiency. This holistic perspective equips researchers and engineers with valuable insights to choose the most appropriate tools tailored to their specific requirements in combating DF technologies.

TABLE I. COMPARISON OF TRADITIONAL AND MODERN METHODS

Method	Characteristics	Performance	Data Dependency	Flexibility	Computational Cost	Robustness
Traditional Neural Networks	Relies on manually extracted features	Variable	Low	Low	Low	Moderate
Modern Neural Networks	Relies on DL and implicit feature extraction	High	High	High	High	High
Hybrid Methods	Combines traditional and modern techniques	Moderate	Moderate	Moderate	Moderate	High
Handcrafted Feature Methods	Utilizes specific predefined features	Low to Moderate	Low	Low	Low	Low to Moderate
Ensemble Methods	Combines multiple models to improve accuracy	High	High	High	High	High
Temporal Analysis Methods	Focuses on temporal Inconsistencies	High	High	Moderate	Moderate to High	High
Physiological Signal Methods	Analyzes physiological signals like eye blinking or heart rate	Moderate to High	High	Low to Moderate	Moderate	High
Audio-Visual Methods	Combines audio and visual cues for detection	High	High	High	High	High
Capsule Networks	Utilizes capsules to retain spatial hierarchies	High	High	High	High	Moderate to High
Generative Adversarial Networks	Uses adversarial training to detect DFs	High	High	High	High	High

3.1 Deepfake Video Detection

Video DF detection is the process of distinguishing between authentic and fake video footage created using sophisticated ML techniques, particularly DL models. This form of Video DF detection involves distinguishing between genuine videos and those that have been falsified or fabricated [34]. Some methods of identifying video DFs include the following:

- Analysis of visual artifacts: This involves checking for irregularities in video frames or something that does not seem right or that is inconsistent with others. Various techniques to spot the anomalies inserted throughout the compilation procedure include facial feature tracking, motion analysis, and frame analysis.
- analytical-visual: involves the simultaneous analysis of the visual and auditory components of a film that can identify DFs by detecting differences or discrepancies between them.

Several methods have been developed to recognize altered films. While current detection methods mostly focus on fundamental characteristics, previous methods mainly addressed irregular features resulting from the face synthesis process. These tactics can be classified into five categories based on the attributes they use. As Table 2 illustrates [17].

TABLE II. CLASSIFICATION FOR EXISTING DETECTION METHODS

no.	Methods	Description
1	Generalnetwork-based methods	According to this approach, CNNs complete the frame-level classification task of detection.
2	Temporalconsistency-based methods	It is discovered that the forgery algorithm’s flaws cause discrepancies between neighboring frames in DF films. RNN is therefore used to detect these discrepancies.
3	Visualartefacts-based methods	The blending operation in generation process would cause intrinsic image discrepancies in the blending boundaries these artefacts are identified using CNN-based techniques.

4	Camerafingerprints-based methods	Devices leave various imprints in the photos they take because of specific generation processes. Faces and background pictures are recognized to originate from distinct devices concurrently. Thus, by utilizing these traces, the detecting work may be finished.
5	Biologicalsignals-based methods	It is challenging for GAN to decipher the biological information concealed in faces, which makes it challenging to create synthetic human faces with rational behavior. This discovery is used to extract biological signals in order to identify DF films.

The group of researchers and their studies that they conducted in this field will be discussed, with mention of the conclusions they achieved, and then detailing each study to know the strengths, weaknesses, accuracy, and results that they obtained in their research:

Salive et al. [43] offers a way to identify DFs, which are artificial media made with DL methods. The authors suggest a multimodal strategy that makes the final prediction by examining the subject's voice and face. The study examines the effectiveness of the suggested approach by taking into account different fusion procedures between the two modalities on a number of cutting edge multimodal video DF datasets. They assessed the performance of the detectors under consideration using Receiver Operating Characteristic (ROC) curves and confusion matrices, using the Area Under the Curve (AUC) and Balanced Accuracy (BA) as evaluation measures. The authors show that a multimodal approach is equally more functional and robust than a monomodal one. The results show the effectiveness and robustness of the proposed approach, indicating high generalization capabilities on unseen data. Zhang et al. [44] To identify DF videos, they suggested using a Temporal Dropout 3Dimensional Convolutional Neural Network (3DCNN). It is fed fixed-length frame volumes collected from a video in order to extract features at various scales and classify them as authentic or fake. To specifically sample frames at random within each batch, a temporal dropout procedure is implemented. It is a straightforward but efficient method of augmenting data, and it can im- prove generalization and representation while preventing over fitting of the model and increasing detection precision. Consequently, the video level classifier is trained to accurately and efficiently identify DF videos. Extensive experiments conducted with widely recognized benchmarks clearly demonstrate the effectiveness and generalizability of this method. Xiaodan Li et al. [45] they presented a new partial face attack problem in DF films in which just video level labels are given and not all of the false videos' faces are altered. The multiple instance learning (MIL) framework was used to overcome this issue, considering input video and faces as instances and bags, respectively. Unlike standard MIL, which produces a mapping from instance embedding to instance prediction and then to bag prediction, a sharp MIL (S-MIL) is presented that builds a direct mapping from instance embedding to bag prediction. It can be demonstrated theoretically that S-MIL relieves the gradient vanishing that occurs in standard MIL. Spatial-temporal encoded instances are designed to adequately describe the intra-frame and inter-frame inconsistencies in order to build instances that may accurately contain the partially modified faces, thus enhancing the detection performance. Additionally, a new dataset called FFPMS was created for the purpose of partially attacked DF video detection. Trials conducted on FFPMS and the popular DFDC dataset confirm that S-MIL outperforms its competitors in par- tially attacked DF video identification. Furthermore, S-MIL can be tailored to conventional DF image identification assignments and get cutting-edge results on single-frame datasets. Yuezun et al. [46] They developed a cutting edge DL method that effectively distinguishes between authentic and AI generated fake videos, which they have since named DF films. This method was developed in response to the finding that the current DF algorithm can only generate images with a certain level of detail, necessitating fur- ther warping in order to match the original faces in the film. These changes result in distinct artifacts in the DF movies that are generated, proving that CNNs can effectively capture them. Nguyen et al. [47] The technique makes use of a capsule network to identify several types of spoofs, ranging from computer-generated films with deep CNN to replay attacks with printed graphics or recorded videos. It expands on the original use of capsule networks by using them to solve inverse graphics challenges. This extensive set of studies proved that developing a generic detection technique that works for a variety of faked image and video attacks is feasible. Additionally, they showed that capsule networks are not limited to computer vision applications. Most of the time, the suggested usage of random noise during the training stage worked well. Xiaodan et al. [48] propose method for identifying DFs produced by the GANs model by utilizing the DeepVision algorithm to examine a notable alteration in the pattern of blinking, which is a spontaneous and voluntary motion that doesn't involve conscious effort. It has been observed that a person overall physical state, cognitive function, biological makeup, and degree of information processing can all have a major impact on how they blink. The pattern can be influenced by various factors such as an individual age or gender, the time of day, their emotional condition, or their level of awareness. Hence, DFs can be identified through integrity verification by monitoring notable alterations in the eye blinking patterns in DFs using a heuristic approach based on the findings of research in medicine, biology, and brain engineering, in addition to ML and different algorithms based on statistical and engineering knowledge. This means that by monitoring notable alterations in a subject eye blinking pattern while they are captured on camera, they can carry out integrity verification. When eye blinks are repeatedly repeated in a short amount of time, the suggested DeepVision approach is used as a precaution to confirm an anomaly based on the duration, repeated number, and elapsed eye blink time. Abdul and Janet [49] did a study that provided an approach that distinguishes actual videos from modified or fraudulent videos using a DL-based methodology. To identify these manipulations, our approach combines CNN and LSTM. Frame-level features are extracted using CNNs and LSTM. RNN is trained on these data to distinguish between real and false videos. The results were excellent when they compared them to other approaches that were already in use. A variety of sources provided the dataset for the DF detection model training, including Face Forensics, DFDC, Celeb DF, and the DF Detection Challenge dataset, to mention a few. Where they succeeded in obtaining a competitive result of 92 percent accuracy while using a simple architecture.

X. H. Nguyen et al. [50] They presented of a 3-dimensional 3DCNN model, which has the ability to extract spatiotemporal information from a video's neighboring frame sequence. Over 99% of the proposed network's binary detection accuracy was achieved on the two largest benchmark datasets, VidTIMIT and DF of FaceForensics++. The experimental findings of the proposed method outperformed the state-of-the-art methods. priti et al. [51] they demonstrated a DL-based technique for automatically identifying DFs. When they used transfer learning to successfully develop a DF detection model, the LSTM layer was trained using the features, and the resulting confusion matrix provided the validation and testing accuracy. As a result, the model splits user video into frames, and LSTM and InceptionResNetV2 were used to further preprocess these frames. The suggested methodology is able to analyze any video using a convolutional LSTM system and also aids in detecting DF faces that have been altered, preventing people from defaming. young et al. [52] they suggested an effective vision transformer model In order to extract both local and global information for DF detection. In order to interact with all positions and define the artifact zone, they integrated patch-based positioning with vector concatenated CNN feature. The sigmoid function is used to train the logit using binary cross entropy for the distillation token. The performance of the suggested model is enhanced by adding this distillation.

A DF identification technique by Hadi et al. [53] achieves great accuracy in differentiating real films from deep- fakes by using noise residuals rather than RGB pixels. A CNN architecture with two convolution layers and one max pooling layer serves as the foundation for the model. A training accuracy of almost 98% and a validation accuracy of almost 97% were attained by the suggested model. The model demonstrated reduced training time, more layers and parameters, and improved detection accuracy when compared to other approaches. Overall, the results of the study suggest that using noise residuals can be an effective approach to DF detection with high accuracy rates. Yuezun et al. Modern [54] technique ues make it possible to create a fake version of a single video from a social network in real time. Even though a lot of techniques have been created for identifying fake photos and videos, they are usually limited to certain domains and rapidly become outdated when new attack types emerge. This research presents a method that employs a capsule network to detect several types of spoofs, ranging from computer-generated films utilizing deep CNN to replay attacks employing printed images or recorded videos. It expands on the original use of capsule networks by using them to solve inverse graphics challenges. Joseph et al. [55] proposal for a DF detection and classification model included CNNs with five layers. After the model has extracted the face region from video frames, the CNN augmented with ReLU is utilized to extract features from these faces. For the DF-detection-influenced video, a CNN equipped with a ReLU model was utilized to ensure model correctness while keeping an appropriate weight. The suggested model's performance assessment was conducted with the Face2Face and DF datasets for first-order motion. According to experimental data, under real network diffusion conditions. Suratkarand and Kazi [56] By using a hybrid model of CNN and RNN in conjunction with transfer learning in auto-encoders, they created a novel framework to recognize false films. An analysis is conducted on a test set of unseen input data to verify the model generalizability. Additionally, the impact of residual picture input on the model's accuracy is examined. To verify its efficacy, results are shown with and without transfer learning. The generaliz ability issue, which enhances the networks resilience to various attacks, needs more research. where they discussed using a transfer learning strategy to improve the accuracy of deep fake detection models. Compared to models that are built from scratch, fine tuned models are able to offer superior accuracy. kosarkar et al. [57] They compared a modified CNN algorithm to two other methods in order to determine which was more effective at identifying DF images from a collection of movies. Data from Kaggle was used to test and train our model. Three distinct CNN models are used to train CNNs to distinguish between real and DF images. Furthermore, a customized CNN model has been built and implemented. It has many more layers, such as a dropout layer, MaxPooling, and a thick layer. This method undergoes the processes of frames extraction, face feature extraction, data preprocessing, and classification to determine whether the images in the video are real or fraudulent.

DF videos involve the use of advanced AI techniques to create realistic but fake videos, where the original video content is manipulated to change individuals appearance, actions or voices. These videos can range from simple face swaps to very complex manipulations involving multiple actors and environments. Datasets used to train and evaluate DF detection methods are essential for developing robust and reliable detection systems. See Table. 3 for more details about the types of data used in DF video.

TABLE III. COMMONLY USED DATASETS FOR DF VIDEO DETECTION AND THEIR EFFECTIVENESS

Dataset	no. of Videos	Types of DFs Covered	Resolution	Effectiveness	Remarks
FaceForensics++	1,000+	Face swaps, reenactments	720p	High	Widely used benchmark dataset with diverse manipulations.
DF Detection (DFD)	3,000+	Face swaps	1080p	High	High-quality videos, good for training DL models.
Celeb-DF	5,000+	Face swaps	480p-720p	Moderate	Contains more realistic and challenging DFs.
UADFV	100+	Face swaps	1080p	Low to Moderate	Limited number of videos, less diversity.
DF Detection Challenge DFDC	124,000+	Face swaps	Various	High	Largest dataset, highly diverse, used in competitions.
VoxCeleb2	1,000+	Audio-visual DFs	720p	Moderate to High	Good for multi-modal DF detection.
YouTube DFs	500+	Face swaps	Various	Moderate	Real-world DFs, varied quality.
DeeperForensics-1.0	60,000+	Face swaps, Manipulations	480p-1080p	High	Contains a wide variety of manipulations and conditions.
WildDF	7,000+	Face swaps, Manipulations	Various	Moderate to High	Collected from the internet, highly diverse.
DF-TIMIT	620	Face swaps	720p	Low to Moderate	Limited number of videos, mostly used for initial research.

Table 4 provides a comparison between traditional and modern methods for detecting DF videos. Traditional methods include image forensics, frequency domain analysis, and metadata analysis, which focus on identifying simple manipulations through visual inspection, frequency anomalies, and metadata discrepancies. Modern methods utilize DL and NNs, such as CNN, RNN, and GANs, to achieve higher accuracy in detecting DFs. However, these advanced techniques require substantial computational resources and extensive training data. The table offers a comprehensive comparison of these methods in terms of their descriptions, strengths, and weaknesses.

TABLE IV. COMPARISON OF TRADITIONAL AND MODERN METHODS FOR DF VIDEO DETECTION

Method	Type	Description	Strengths	Weaknesses
Image Forensics	Traditional	Analyzes inconsistencies in video frames, such as lighting, shadows, and reflections.	Effective for detecting simple manipulations.	Struggles with high-quality DFs and complex edits.
Frequency Domain Analysis	Traditional	Examines anomalies in the frequency domain (e.g., JPEG compression artifacts).	Good for identifying compression inconsistencies.	Limited by resolution and can be bypassed by high-quality forgeries.
Metadata Analysis	Traditional	Looks for discrepancies in video metadata (e.g., editing history, timestamps).	Useful for detecting straightforward tampering.	Ineffective against sophisticated DFs that alter metadata.
Convolutional Neural Networks (CNN)	Modern	Uses DL to detect patterns and features indicative of DFs in video frames.	High accuracy and can learn complex features.	Requires large datasets and significant computational resources.
Recurrent Neural Networks (RNN) / Long Short-Term Memory (LSTM)	Modern	Analyzes temporal sequences in video to detect inconsistencies over time.	Effective for capturing temporal dependencies.	Computationally expensive and may require extensive training.
Generative Adversarial Networks (GANs)	Modern	Detects DFs by learning to differentiate between real and fake videos through adversarial training.	High accuracy and can improve detection over time.	Computationally intensive and can sometimes produce false positives.
Hybrid Models (e.g., CNN + LSTM)	Modern	Combines CNN for spatial analysis and LSTM for temporal analysis to enhance detection accuracy.	Captures both spatial and temporal features effectively.	Even more computationally demanding and complex to implement.
Optical Flow Analysis	Modern	Tracks motion patterns in videos to identify unnatural movements or inconsistencies.	Good for detecting dynamic inconsistencies in motion.	May struggle with subtle manipulations.
Attention Mechanisms	Modern	Uses attention layers to focus on critical regions of video	Improves model focus on important areas, enhancing	Requires complex architectures and significant tuning.

		frames for better detection accuracy.	accuracy.	

Table. 5 summarizes the performance metrics of various methods used in detecting DFs across video clips. It provides an overview of their accuracy, precision, recall, F1-score, computational cost, and robustness to variations. Methods such as CNN and RNN show moderate to high accuracy, while hybrid methods and ensemble approaches demonstrate superior performance in most categories. These methods vary in computational cost, with ensemble methods and GANs being computationally intensive but highly robust to variations. This comparison helps identify the most suitable techniques for specific detection tasks based on performance and resource requirements.

TABLE V. PERFORMANCE ANALYSIS OF DIFFERENT METHODS FOR DF VIDEO DETECTION

Method	Accuracy	Precision	Recall	F1-Score	Computational Cost	Robustness to Variations
Convolutional Neural Networks (CNN)	85-92%	0.85-0.91	0.85-0.92	0.85-0.91	Moderate	High
Recurrent Neural Networks (RNN)	80-88%	0.80-0.86	0.80-0.88	0.80-0.87	High	Moderate to High
Hybrid Methods (CNN + LSTM)	87-94%	0.87-0.93	0.87-0.94	0.87-0.93	High	High
Temporal Analysis Methods	83-90%	0.83-0.89	0.83-0.90	0.83-0.89	Moderate	High
Generative Adversarial Networks (GANs)	89-95%	0.89-0.95	0.89-0.95	0.89-0.95	High	High
Capsule Networks	84-91%	0.84-0.90	0.84-0.91	0.84-0.90	High	Moderate to High
Physiological Signal Methods	81-89%	0.81-0.87	0.81-0.89	0.81-0.88	Moderate	High
Two-Stream Networks	88-94%	0.88-0.94	0.88-0.94	0.88-0.94	High	High
Ensemble Methods	90-97%	0.90-0.97	0.90-0.97	0.90-0.97	Very High	Very High
DeepRhythm	85-92%	0.85-0.91	0.85-0.92	0.85-0.91	Moderate	High

Table.6 offers a thorough summary of the numerous classifiers and DL algorithms used to detect video DFs. The table covers a range of techniques, including hybrid approaches, CNNs, and GANs, applied to diverse datasets such DFDC, CelebDF, FaceForensics, and CEW. It features the techniques employed, datasets, year of publication, authors, and high-performance metrics attained by these approaches. Researchers and practitioners can better com- prehend current developments in this field and anticipate future trends by referring to the table, which demonstrates the great accuracy these systems have achieved in detecting fraudulent videos.

The performance data that have been disclosed demonstrate a spectrum of accuracy, ranging from 81% to almost flawless scores, hence highlighting the efficacy and continuous improvement of detection models. Interestingly, ad- vances in accuracy rates are a reflection of iterative improvements in the robustness of the model against ever-more complex DF generating techniques. This collection aims to clarify the development of detection technologies and their critical function in protecting digital content integrity in the face of ever-changing risks from synthetic me- dia manipulation. Recent research in DF detection is shown in the table below, which includes information on the authors, publication years, methods utilized, and datasets used. It draws attention to improvements in model performance and offers insights into how video manipulation detection has changed recently.

TABLE VI. AN OVERVIEW OF VIDEO DF

Ref.	Authors	Year	Classifiers/ Techniques	Dataset	High performance	Limitations/Constraints
[54]	Yuezun Li et al.	2018	GAN + CNN	CEW dataset	High performance is 0.99 to CNN 0.98 ERO 0.79	Dataset limited to specific facial manipulation examples.
[34]	Darius et al.	2018	DF + Face2Face	DF video dataset	95% Face2Face 98% DF	Overfitting to specific dataset features; lacks diversity.
[46]	Yuezun et al.	2019	CNN	DF video dataset from defferent source	accuracy is 84.6-99.4	High variance between performance across different datasets.
[47]	Nguyen et al.	2019	CNN	capsule-Forensics-noise capsule-Forensics	accuracy rate 89.10-100.00	Ineffective for low-quality videos with high compression.
[40]	Sabir etal.	2019	GAN	FaceForensics++	accuracy is 98%	GAN-based methods vulnerable to adversarial attacks.

[58]	Sakina et al.	2019	GAN	celebDF	accurcey rate 95% - 98%%	Dataset lacks diversity, limited to celebrity videos.
[48]	Jung et al.	2020	GAN	heir DF dataset	accurcey rate 87.5%	Small dataset size leading to less generalizable results.
[45]	Xiaodan Li et al.	2020	S-MIL S-MIL-T	DFDC celebDF FaceForensics++	0.8511 DFDC 0.9944 celebDF 1.0000 FaceForensics++	Method complexity increases computational cost.
[59]	Suratkar et al.	2020	CNN	FaceForensics++ Google Deep Fake Dataset Facebook Deep Fake Dataset	accuracy is 98%	Difficulty in generalizing across different types of fakes.
[49]	Abdul and Janet	2021	CNN +LSTM	FaceForensics DFDC celebDF	accuracy is 92%	High computational resource requirements.
[44]	Zhang et al.	2021	TD-3DCNN	FaceForensics++ DFDC celebDF	81.08 celebDF 82.64 DFDC 79.09 FaceForensics++	Struggles with real-time processing of video frames.
[50]	X. H. Nguyen et al.	2021	CNN	FaceForensics VidTIMIT datasets	accurcey rate of (99.4-94.5)	Model vulnerable to adversarial noise.
[51]	priti et al.	2021	InceptionResNetV2 LSTM	FaceForensics VidTIMIT celeb DF	accurcey is 91.48	Limited to facial DF detection, struggles with audio.
[55]	Joseph et al.	2022	CNNs Face2Face	DFfluenced video	accuracy is 86%	Limited performance on low-resolution videos.
[60]	Jacob et al.	2022	GAN+CNN	celeb DFv2 DF TIMIT FaceForencsics++	accuracy rate is 81%-99%	Lower accuracy on new and complex types of DFs.
[53]	Hadi et al.	2022	CNN+SRM	FaceForensics dataset	accuracy is 98%	Limited generalizability across other datasets.
[43]	Salvi et al.	2023	CNN, LSTM and SVM	DFDC DF TIMIT FakeAVceleb	AUC is 0.83 FakeAVceleb AUC is 0.79 TIMIT AUC is 0.66 DFDC	Multi-modal data handling remains a challenge.
[52]	young et al.	2023	CNN SOTA model	celeb-DF v2 DFDC	0.993 AUC 0.978 score	SOTA models prone to overfitting on smaller datasets.
[56]	Suratkarand and Kazi	2023	CNN-RNN	DFDC DFD FaceForensics FaceForensics++	accuracy is 92.77%	Struggles with large-scale datasets and real-time detection.
[57]	kosarkar et al.	2023	CNNs	Kaggle dataset	accuracy is 95.57%	Dataset biased towards specific fake techniques.

2.2 Audio DF detection

One of the key challenges in the field of audio DFs lies in generating realistic and convincing multimedia content using methods such as generative adversarial networks (GANs) and variational autoencoders (VAEs) [61]. These technologies have diverse applications, ranging from media production and entertainment to more malicious uses, such as creating fake audio recordings for fraud or disinformation campaigns [62]. This has raised significant concerns regarding security, privacy, and ethical implications due to the potential for abuse. The widespread use of content manipulation techniques exacerbates fears surrounding the authenticity of digital content, particularly audio DFs [63].

Detecting audio DFs remains a complex task that often requires the analysis of synchronized facial and acoustic features to induce realistic speech or non-verbal behaviors [64]. Despite ongoing research, the lack of a comprehensive overview of audio DF creation and detection methods makes it challenging to build reliable detection systems [66]. However, there are significant opportunities to advance detection and classification models for distinguishing between authentic and synthetic audio. These advancements are crucial for mitigating the risks associated with DFs, including damage to credibility, reputation, and market deceptions [67].

Audio DFs are commonly generated by DL models such as CNNs and GANs. Huge collections dataset of recorded human speech are used to train these models so they can comprehend and replicate the subtle differences between voice characteristics. In addition to their incredible powers, these tools present formidable challenges, principally in the areas of detection as well as prevention [68]. To detect audio DFs and mitigate their impact, scientists and engineers are devoting a lot of effort to this endeavor. In order to limit the use and accessibility of these technologies, laws and regulations are being put in place, along with the development of complex algorithms that can distinguish between synthetic and genuine sounds [69].

This report offers an extensive summary of the present state of audio DF skill. We review the basic methods, evaluate model presentation, and consider safety and ethical effects. Additionally, we review the latest advancements in detection techniques and propose further study directions for this rapidly growing topic. Audio DFs are sound records that are created artificially to simulator the voice of a certain person, usually without that person consent. With advanced ML methods, such DL models, these modified audio clips can be made to resemble the speaking patterns, tone, and voice of a goal individual. It is possible to fake listeners into thinking that audio DFs are genuine, which raises concerns about deceit, identity theft, and attacks of privacy [43]. Advance on the detection of audio DFs is current, and several methods have been established to know soundtracks that are fake or altered. These techniques may involve analyzing low-level features of the audio signal, such as artifacts introduced during the generation process, as well as greater level semantic features that are more challenging for DF producers to replicate accurately [70][100]. Advanced ML algorithms are used to concept audio DFs, sometimes referred to as artificial or altered audio recordings, which mimic the speech and vocal patterns of a individual person. These methods frequently use DL models, like GANs and RNNs, to produce extremely lifelike audio content that imitates the voice of a target person [71]. Generally, a model is trained on a sizable dataset of audio examples from the target person in order to generate audio DFs. After learning to know the subtleties of a speaker intonation, accent, and speech patterns, the model can produce new audio clips that sound a lot like the goal voice. This skill has the potential to generate convincing audio forgeries that can be used for malicious determinations, such as dispersal misinformation or impersonating individuals [72, 74].

TABLE VII. COMMONLY USED DATASETS FOR DF AUDIO DETECTION AND THEIR EFFECTIVENESS

Dataset	Number of Samples	Types of DFs Covered	Sampling Rate	Effectiveness	Remarks
ASVspoof 2019	19,500+	Voice conversion, speech synthesis	16kHz	High	Comprehensive benchmark for spoofing detection.
WaveFake	12,000+	Waveform-based audio DFs	16kHz	High	Covers various audio synthesis techniques.
FakeAVCeleb	12,000+	Audio-visual DFs	16kHz	Moderate to High	Multi-modal dataset for joint detection.
Voice Conversion Challenge (VCC)	10,000+	Voice conversion	16kHz	Moderate	Focuses specifically on voice conversion.
LA Voice	5,000+	Synthetic speech	16kHz	Moderate	Limited number of samples, used for initial research.
LibriSpeech	1,000 hours	Various, including text-to-speech	16kHz	High	Widely used for various speech synthesis tasks.
VoxCeleb2	1,000+	Audio-visual DFs	16kHz	Moderate to High	Good for multi-modal DF detection.
Fake YouTube Audio	2,000+	Various, including voice cloning	16kHz	Moderate	Real-world DFs, varied quality.
ASVspoof 2015	7,350+	Speech synthesis, voice conversion	16kHz	Moderate	Earlier benchmark, still relevant for basic research.
FakeSpeechNet	15,000+	Speech synthesis, voice conversion	16kHz	High	Large, diverse dataset specifically for fake speech detection.

A summary of some of the greatest often used datasets in the ground of DF audio detection is given in Table 7. The quantity of examples, the variety of DFs comprised, and the sampling speeds of these datasets differ. The table also contains comments on the unique features and scenarios for use of each dataset, as well as how well it contributed to the detection attempts. Researchers can more effectively choose important datasets for their studies and develop the status of DF audio detection skill by being aware of the advantages and disadvantages of these datasets.

Analyzing different aspects of the audio signal to detection indiscretions, distortions, or designs suggestive of fake or real content can be portion of detection methods for audio DFs. These techniques can contain low-level analysis of the audio signal for example, spectral features and temporal patterns or higher-level semantic feature analysis concerning prosody, speech content, and speaker attributes[73] [75].

Some common approaches to audio DF detection include [76, 77]:

1. **Stream-level analysis:** This method looks for differences or artifacts in the audio stream that could view to altered or synthesized content. Unevenities created during the creating process can be found by means of methods like waveform analysis, frequency field analysis, and spectrogram analysis.

2. **Semantic feature-based analysis:** This method looks for changes or discrepancies that can fact to fake or altered audio records by investigative higher-level semantic aspects of the audio physical, such as prosody, expressive cues, and speaker documentation qualities.

The techniques listed in the Table. 8 refer to the use of different NNs such as CNN and RNN, in addition to other advanced techniques such as ad hoc networks, spectral feature analysis, and temporal harmony. Each has advantages and disadvantages in accuracy, efficiency, and computational cost. Variation in accuracy and effectiveness shows that hybrid methods, GANs, and sampling methods give superior results compared to other methods, but they may require higher computational costs. In summary, this table provides a comprehensive overview of how different methods perform in the field of counterfeit audio detection, helping researchers and engineers choose the right tool according to their technical needs and requirements.

TABLE VIII. PERFORMANCE ANALYSIS OF DIFFERENT METHODS FOR DF AUDIO DETECTION

Method	Accuracy	Precision	Recall	F1-Score	Computational Cost	Robustness to Variations
Convolutional Neural Networks (CNN)	80-88%	0.80-0.86	0.80-0.88	0.80-0.87	Moderate	High
Recurrent Neural Networks (RNN)	78-86%	0.78-0.84	0.78-0.86	0.78-0.85	High	Moderate to High
Hybrid Methods (CNN + LSTM)	83-90%	0.83-0.89	0.83-0.90	0.83-0.89	High	High
Audio Forensic Analysis	75-85%	0.75-0.83	0.75-0.85	0.75-0.84	Low to Moderate	Moderate
Generative Adversarial Networks (GANs)	85-92%	0.85-0.91	0.85-0.92	0.85-0.91	High	High
Spectral Feature Analysis	80-88%	0.80-0.86	0.80-0.88	0.80-0.87	Moderate	High
Temporal Coherence Analysis	78-85%	0.78-0.83	0.78-0.85	0.78-0.84	Moderate	Moderate to High
Phoneme Pattern Analysis	82-90%	0.82-0.89	0.82-0.90	0.82-0.89	Moderate	High
Ensemble Methods	87-94%	0.87-0.93	0.87-0.94	0.87-0.93	Very High	Very High
DeepRhythm	83-90%	0.83-0.89	0.83-0.90	0.83-0.89	Moderate	High

Some studies conducted in this field have been described and the results of each study have been mentioned, along with the strengths, limitations and accuracy that were reached from them:

Mittal et. al. [78] assesses the effectiveness of single modality, ensemble-based, and multimodal detection techniques using the three different types of fakes found in the Fake Audio Video Celebrity dataset. These tests demonstrated that multimodal approaches outperformed unimodal approaches that concentrated on a single modality, but that unimodal approaches did not perform as well as ensemble-based approaches. She also talked about the shortcomings of the detection techniques used today and the necessity for greater study to create more advanced DF detectors. Alzantot et. al. [79] they introduced a brand new audio spoofing detection algorithm that works in both physical and logical access situations.they presented an analysis of the differences in performance between thier model and three distinct feature extraction strategies. The assessment dataset results show that the fusion of thier models CM scores improves the baseline algorithm’s t-DCF and EER metrics by 71% and 75%, respectively, against replay assaults. Additionally, thier fusion of models increases the t-DCF and EER metrics by about 25% each against the TTS and VC threats. Yu Hong et.al. [80] they uses dynamic acoustic features and DNN classifiers to detect spoofing in speaker verification systems. Five different types of SBCC and CQCC features were tested, and the performance was improved using DNN classifiers. The results showed that using dynamic SBCC and CQCC features works better in detecting spoofing, and DNN classifiers significantly improve the accuracy of spoofing detection. Mittal et. al.[81] suggested an approach compares perceived emotions to differentiate between real and fake videos, utilizing a Siamese Network architecture and modality embeddings. The method achieves promising results on bench- mark DF datasets DFDC and DF-TIMIT, showcasing its effectiveness in identifying fake multimedia content. Despite some limitations in misclassifying real videos, the approach outperforms existing state-of-the-art DF detection methods like Two-stream, MesoNet, HeadPose, FWA, and VA. The study suggests future work involving incorporating more modalities and context for enhanced DF detection capabilities.

Khalid et. al. [83] in an attempt to combat the growing threat of DF audio and video impersonation assaults release the FakeAVCeleb dataset. In order to demonstrate how well different models perform when identifying DFs utilizing audio and visual inputs, the study evaluates alternative DF detection strategies. The study emphasizes how advanced detection methods are required to effectively combat this ever-changing threat.In order to demonstrate how well different models perform when identifying DFs utilizing audio and visual inputs, the study evaluates alternative DF detection strategies. The research emphasizes the need for improved detection methods to successfully tackle this evolving threat. Combining Faceswap with FSGAN allowed for the creation of switched DF videos. Moreover, a real-time voice cloning tool based on transfer learning was used to create cloned audios (SV2TTS). Furthermore, Wav2Lip was applied to reenact the videos based on the generated DF videos.

Dora M. et. al. [83] suggested approach uses picture augmentation and dropout and is based on a CNN. 864 histograms were used for cross-validation after 2092 histograms from real and phony speech recordings were used to train the suggested architecture. For external validation, 476 new histograms were employed, and Precision (P) and Recall (R) were computed. For recordings based on imitation and recordings based on deep voice, the detection of false audios was achieved. There was 0.985 worldwide accuracy. The outcomes show that the suggested method is effective in identifying phony voice material. Piotr et. al. [84] they looked into how adversarial attacks affected audio DF detection systems and suggested defenses to lessen the harm they caused. Adversarial attacks were used to reduce the effectiveness of the detecting system, but countermeasures were implemented. The study assessed the transferability mechanism, which creates assaults utilizing other models without requiring access to the targeted model, and analyzed the effectiveness of white box attacks employing multiple adversarial agents. Adaptive adversarial training was used to improve the classifiers robustness to adversarial attacks, as evidenced by the experimental investigation. In the end, the research strengthened the resilience against different types of attacks in both transferability and white box situations, producing satisfactory outcomes even in the face of adversarial attacks.

Shin et.al. [85] the HM-Conformer model for identifying digital audio manipulation is introduced in the document. This is accomplished by using progressive aggregation techniques and multi-level data aggregation approaches. Based on the Conformer architecture, HM-Conformer handles both local and global aspects in audio data by fusing the ben- efits of RNN and Transformer. In comparison to contemporary systems, the HM-Conformer model has demonstrated competitive performance in the audio modification detection challenge. Through the use of progressive aggregation techniques and multi-level data aggregation methods, the Conformer structure is improved to achieve this performance. Zexin et. al. suggested [86] using a DL-based frame level detection method to pinpoint the modified portions and identify partially faked audio as a solution to the issue. they traind and assess that proposed technique using the ADD2022 Challenge data. They assessed that detection model with respect to different network setups and acoustic characteristics. Consequently, detection system attains the greatest performance among partly spofed audio detection systems that can discover manipulated clips, with an EER of 6.58% on the ADD2022 challenge test set.

Mvelo Mcuba et. al. [87] they address the issues raised by new technologies in cybercrimes and data exploitation by advancing digital forensic tools for spotting artificial voices and DF audio. To evaluate the validity of audio files, many methods are used, including CNNs, RNNs and fully convolutional neural networks (FCNNs). The models make use of features including convolutional layer kernel sizes, encoding-decoding procedures, broad blocks, and bidirectional LSTM. The architectures also emphasize the usage of activation functions such as ReLU, padding, and max-pooling. Results from the experiments show varying levels of accuracy and performance across the different models and optimizers used. The custom architecture by Malik et al demonstrates superior accuracy, especially with the SGD optimizer, while the VGG-16 model also shows promising results with the Adadelta optimizer. The study emphasizes the importance of utilizing DL models and preprocessing techniques to enhance the detection of DF audio for forensic investigations.

A thorough summary of the several classifiers and DL algorithms used to identify DF audio can be found in Table 9. The table covers a broad range of approaches, including spectral feature analysis, phoneme pattern analysis, CNNs, RNNs, and hybrid approaches like CNN + LSTM. These methods are assessed according to their computing costs and resilience to changes in data, as well as their accuracy, precision, recall, and F1-score metrics.The performance metrics depicted in the table demonstrate a spectrum of accuracy, ranging from 75% to 94%, underscoring the effectiveness and evolving sophistication of detection models in tackling the challenge of DF audio. Researchers and practitioners can utilize this resource to gain insights into the current landscape of DF detection technologies, aiding in the selection of appropriate methods for practical applications and future research endeavors. In the face of increasing dangers from audio operation techniques, this group seeks to clarify the growths in DF detection systems and their dangerous role in protecting digital content integrity. The following table highlights recent research contributions in the subject of DF audio detection, as well as authors, publication years, techniques used, datasets used, and consistent performance metrics. This highlights continuing growths and trends in the field.

TABLE IX. AN OVERVIEW OF AUDIO DF

Ref.	Authors	Year	Classifiers/ techniques	Dataset	High performance	Limitations/Constraints
[80]	Yu Hong et.al	2018	DNN-HLL	ASPspoof	ERR is 0.045	Limited dataset variety; model may not generalize well to other data types.
[79]	Alzantot et.al	2019	Residual CNN	ASVspoof2019	t-DCF 71% ERRscores 75%	Limited ability to detect audio DFs with complex noise patterns.
[77]	Hasam K. et.al	2019	CNN+RNN AVC	MesoInception-4 meso94 Xceptio EfficeientNet-B0 VGG16	0.7287 0.4593 0.4394 0.6318 0.7804	High computational complexity; limited by dataset quality.
[81]	Mittal et.al	2020	CNN Two-stream MesoNet HeadPose FWA	DFDC DF-TIMIT	accurcy is 84.4% DFDC 96.6% TIMIT	Inconsistent results across datasets; struggles with highly realistic fakes.
[82]	Khaild et.al	2021	FaceSwap FSGAN	Falce-AVceleb	AUCscores is 65%	AUC score indicates lower detection capacity; limited data diversity.
[72]	Khochare et.al	2021	TCN + STN	ASPspoof	accurcy is 92%	Only tested on a single dataset, lacks generalization for unseen attacks.
[83]	Dora M. et.al	2021	CNN	deep voice	accurcy is 0.985	Limited to speech-based DF detection, struggles with real-time detection.
[84]	Piotr K. et.al	2022	LLCNN	SpecRNet RowNet	average ERR is 0.1247	Higher error rates on low-quality audio data.
[85]	Shin et.al	2023	HM-Conformer	ASPspoof 2021	ERR is 15.71	Limited performance in multi-lingual environments or noisy conditions
[86]	Zexin et.al	2023	Wav2Vec	ADD2022	ERR is 6.58%	Model struggles with cross-dataset generalization and noisy audio samples.
[87]	Mvelo Mcuba	2023	MFCC Mel-spectrum VGG-16	Baidu Silicon Valley AI Lab dataset	acurccy is 68.63%	Lower accuracy rate; unable to generalize well across different languages.

3. CONCLUSIONS

The conclusion of the paper emphasizes the critical analysis of DF detection techniques, highlighting the strengths and limitations of various methodologies. The study provides a comprehensive overview of state-of-the-art approaches, including CNNs, LSTM networks, and hybrid models, which have shown promising results in terms of accuracy and efficiency. Key benefits of the research include the identification of effective detection methods that can achieve high accuracy rates, as well as the exploration of multimodal techniques that enhance robustness against diverse DF outputs. The systematic literature review conducted in this study serves as a valuable resource for researchers and practitioners in the field, guiding future developments in DF detection technologies. The implications of this study are significant, particularly in the context of preserving the integrity of digital content across various sectors such as journalism, cybersecurity, and law enforcement. It underscores the necessity for ongoing research to develop adaptive models capable of generalizing across a wide range of DF applications. Furthermore, the study calls for the establishment of strict legal and ethical frameworks to mitigate the risks associated with DF technology, ensuring that advancements in artificial intelligence are harnessed for the greater good of society. Overall, this research contributes to the ongoing discourse on the ethical use of AI and the importance of safeguarding truth and integrity in digital media.

Conflicts of interest

The author has no conflicts of interest relevant to this article.

Funding

The absence of funding details in the author's paper suggests that the research was entirely self-funded.

Acknowledgement

The author would like to thank the institution for their institutional support, which played a vital role in the implementation of this study.

References

[1] 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, 2020.

[2] L. Floridi, "What the near future of artificial intelligence could be," 2019 Yearb. Digit. Ethics Lab, pp. 127–142, 2020.

[3] M. Westerlund, "The emergence of deepfake technology: A review," Technol. Innov. Manag. Rev., vol. 9, no. 11, 2019.

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

[5] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," ACM Comput. Surv., vol. 54, no. 1, pp. 1–41, 2021.

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

[7] B. Dolhansky, "The dee pfake detection challenge (DFDC) pre view dataset," arXiv Prepr. arXiv1910.08854, 2019.

[8] S. Salman and J. H. Soud, "Deep Learning Machine using Hierarchical Cluster Features," Al-Mustansiriyah J. Sci., vol. 29, no. 3, pp. 82–93, 2018.

[9] J. Rohaniyah and S. Rijal, "Utilizing Faceapp Application as Media in Teaching Speaking (Practical Ideas and Implementation Guidelines for Speaking Class)," English Teach. J. A J. English Lit. Lang. Educ., vol. 8, no. 2, pp. 67–87, 2020.

[10] T. Sathish, T. S. Abinaya, B. Anupriya, and L. Uma, "Manual fakeapp detection using sentimental analysis through webpage," Semant. Sch., pp. 208–221, 2018.

[11] S. Feldstein, "How artificial intelligence systems could threaten democracy," Conversat., 2019.

[12] G. Al-Kateb, I. Khaleel, and M. Aljanabi, "CryptoGenSec: A Hybrid Generative AI Algorithm for Dynamic Cryptographic Cyber Defence," Mesopotamian J. CyberSecurity, vol. 4, no. 3, pp. 22–35, 2024.

[13] B. A. Jaafar, M. T. Gaata, and M. N. Jasim, "Home appliances recommendation system based on weather information using combined modified k-means and elbow algorithms," Indones. J. Electr. Eng. Comput. Sci., vol. 19, no. 3, pp. 1635–1642, 2020.

[14] H. B. Dixon Jr, "Deepfakes: More frightening than photoshop on steroids," Judges J., vol. 58, p. 35, 2019.

[15] S. A. H. Alazawi and J. H. Al-A'meri, "Face Feature Recognition System Considering Central Moments," Int. J. Comput. Eng. Res, vol. 3, no. 1, pp. 52–57, 2013.

[16] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," Calif. L. Rev., vol. 107, p. 1753, 2019.

[17] M. T. Jafar, M. Ababneh, M. Al-Zoube, and A. Elhassan, "Forensics and analysis of deepfake videos," in 2020 11th international conference on information and communication systems (ICICS), 2020, pp. 53–58.

[18] M. Subhi, O. F. Rashid, S. A. Abdulsahib, M. K. Hussein, and S. M. Mohammed, "Anomaly Intrusion Detection Method based on RNA Encoding and ResNet50 Model," Mesopotamian J. CyberSecurity, vol. 4, no. 2, pp. 120–128, 2024.

[19] K. Yamamoto et al., "2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)," in International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE), 2020, pp. 1–6.

[20] A. Saihood, M. A. Al-Shaher, and M. A. Fadhel, "A New Tiger Beetle Algorithm for Cybersecurity, Medical Image Segmentation and Other Global Problems Optimization," Mesopotamian J. CyberSecurity, vol. 4, no. 1, pp. 17–46, 2024.

[21] L. Trinh, M. Tsang, S. Rambhatla, and Y. Liu, "Interpretable and trustworthy deepfake detection via dynamic prototypes," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp.

1973–1983.

[22] M. A. Younus and T. M. Hasan, “Effective and fast deepfake detection method based on haar wavelet transform,” in 2020 International Conference on Computer Science and Software Engineering (CSASE), 2020, pp. 186–190.

[23] O. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in BMVC 2015-Proceedings of the British Machine Vision Conference 2015, 2015.

[24] P. Korshunov et al., “Tampered speaker inconsistency detection with phonetically aware audio-visual features,” in International conference on machine learning, 2019.

[25] B.-S. Lin, D.-W. Hsu, C.-H. Shen, and H.-F. Hsiao, “Using fully connected and convolutional net for GAN-based face swapping,” in 2020 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), 2020, pp. 185–188.

[26] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, “Synthesizing obama: learning lip sync from audio,” ACM Trans. Graph., vol. 36, no. 4, pp. 1–13, 2017.

[27] J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6447–6456.

[28] P. Korshunov and S. Marcel, “Speaker inconsistency detection in tampered video,” in 2018 26th European signal processing conference (EUSIPCO), 2018, pp. 2375–2379.

[29] J. Galbally and S. Marcel, “Face anti-spoofing based on general image quality assessment,” in 2014 22nd international conference on pattern recognition, 2014, pp. 1173–1178.

[30] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1–11.

[31] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “Mesonet: a compact facial video forgery detection network,” in 2018 IEEE international workshop on information forensics and security (WIFS), 2018, pp. 1–7.

[32] D. Güera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), 2018, pp. 1–6.

[33] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[34] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, “Video face manipulation detection through ensemble of cnns,” in 2020 25th international conference on pattern recognition (ICPR), 2021, pp. 5012–5019.

[35] M. Suresha, S. Kuppa, and D. S. Raghukumar, “A study on deep learning spatiotemporal models and feature extraction techniques for video understanding,” Int. J. Multimed. Inf. Retr., vol. 9, no. 2, pp. 81–101, 2020.

[36] U. A. Ciftci, I. Demir, and L. Yin, “Fakecatcher: Detection of synthetic portrait videos using biological signals,” IEEE Trans. Pattern Anal. Mach. Intell., 2020.

[37] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, “Multi-attentional deepfake detection,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2185–2194.

[38] U. A. Ciftci, I. Demir, and L. Yin, “Fakecatcher: detection of synthetic portrait videos using biological signals,” 2023, Google Patents.

[39] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, “Recurrent convolutional strategies for face manipulation detection in videos,” Interfaces (GUI), vol. 3, no. 1, pp. 80–87, 2019.

[40] D. Salvi et al., “A robust approach to multimodal deepfake detection,” J. Imaging, vol. 9, no. 6, p. 122, 2023.

[41] S. Kumar, “Intelligent Bearing Fault Diagnosis and Classification based on Support Vector Machine,” in 2021 2nd Global Conference for Advancement in Technology (GCAT), 2021, pp. 1–6.

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[43] S. Kumar, “Intelligent bearing fault diagnosis and classification based on support vector machine,” in 2021 2nd Global Conference for Ad- vancement in Technology (GCAT) (IEEE, 2021) pp. 1–6.

[44] D. Salvi, H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang, and S. Tubaro, “A robust approach to multimodal DF detection,” Journal of Imaging 9, 122 (2023).

[45] D. Zhang, C. Li, F. Lin, D. Zeng, and S. Ge, “Detecting DF videos with temporal dropout 3dcnn.” in IJCAI (2021) pp. 1288–1294.

[46] T. Jung, S. Kim, and K. Kim, “Deepvision: DFs detection using human eye blinking pattern,” IEEE Access 8, 83144–83154 (2020).

[47] V. Abdul Jamsheed and B. Janet, “Deep fake video detection using recurrent neural networks,” International Journal of Scientific Research in Computer Science and Engineering 9, 22–26 (2021).

[48] X. Li, Y. Lang, Y. Chen, X. Mao, Y. He, S. Wang, H. Xue, and Q. Lu, “Sharp multiple instance learning for DF video detection,” in
Proceedings of the 28th ACM international conference on multimedia (2020) pp. 1864–1872.

[49] Y. Li and S. Lyu, “Exposing DF videos by detecting face warping artifacts,” arXiv preprint arXiv:1811.00656

(2018).

[50] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-forensics: Using capsule networks to detect forged images and videos,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, 2019) pp. 2307–2311

[51] T. Jung, S. Kim, and K. Kim, “Deepvision: DFs detection using human eye blinking pattern,” IEEE Access 8, 83144–83154 (2020).

[52] V. Abdul Jamsheed and B. Janet, “Deep fake video detection using recurrent neural networks,” International Journal of Scientific Research in Computer Science and Engineering 9, 22–26 (2021).

[53] X. H. Nguyen, T. S. Tran, K. D. Nguyen, D.-T. Truong, et al., “Learning spatio-temporal features to detect manipulated facial videos created by the DF techniques,” Forensic Science International: Digital Investigation 36, 301108 (2021).

[54] P. Yadav, I. Jaswal, J. Maravi, V. Choudhary, and G. Khanna, “DF detection using inceptionresnetv2 and lstm,” in International Conference on Emerging Technologies: AI, IoT, and CPS for Science Technology Applications (2021).

[55] Y.-J. Heo, W.-H. Yeo, and B.-G. Kim, “DF detection algorithm based on improved vision transformer,” Applied Intelligence 53, 7512– 7527 (2023).

[56] W. J. Hadi, S. M. Kadhem, and A. R. Abbas, “Unmasking DFs based on deep learning and noise residuals,” IRAQI JOURNAL OF COMPUTERS, COMMUNICATIONS, CONTROL AND SYSTEMS ENGINEERING 22 (2022).

[57] Y. Li, M.-C. Chang, and S. Lyu, “In ictu oculi: Exposing ai created fake videos by detecting eye blinking,” in 2018 IEEE International workshop on information forensics and security (WIFS) (IEEE, 2018) pp. 1–7.

[58] J. B. Awotunde, R. G. Jimoh, A. L. Imoize, A. T. Abdulrazaq, C.-T. Li, and C.-C. Lee, “An enhanced deep learning-based DF video detection and classification system,” Electronics 12, 87 (2022).

[59] S. Suratkar and F. Kazi, “Deep fake video detection using transfer learning approach,” Arabian Journal for Science and Engineering 48, 9727–9737 (2023).

[60] U. Kosarkar, G. Sarkarkar, and S. Gedam, “Revealing and classification of DFs video’s images using a customize convolution neural network model,” Procedia Computer Science 218, 2636–2652 (2023).

[61] D. Yadav and S. Salmani, “DF: A survey on facial forgery technique using generative adversarial network,” in 2019 International conference on intelligent computing and control systems (ICCS) (IEEE, 2019) pp. 852–857.

[62] S. Suratkar, E. Johnson, K. Variyambat, M. Panchal, and F. Kazi, “Employing transfer-learning based cnn architectures to enhance the generalizability of DF detection,” in 2020 11th international conference on computing, communication and networking technologies (ICCCNT) (IEEE, 2020) pp. 1–9.

[63] J. Mallet, R. Dave, N. Seliya, and M. Vanamala, “Using deep learning to detecting DFs,” in 2022 9th International Conference on Soft Computing & Machine Intelligence (ISCMI) (IEEE, 2022) pp. 1–5.

[64] S. Lyu, “DF detection: Current challenges and next steps,” in 2020 IEEE international conference on multimedia & expo workshops (ICMEW) (IEEE, 2020) pp. 1–6.

[65] T. Sakirin and S. Kusuma, “A Survey of Generative Artificial Intelligence Techniques”, Babylonian Journal of Artificial Intelligence, vol. 2023, pp. 10–14, Mar. 2023.

[66] Z. Khanjani, G. Watson, and V. P. Janeja, “Audio DFs: A survey,” Frontiers in Big Data 5, 1001063 (2023).

[67] K. T. Mai, S. Bray, T. Davies, and L. D. Griffin, “Warning: humans cannot reliably detect speech DFs,” Plos one 18, e0285333 (2023).

[68] M. Pawelec, “DFs and democracy (theory): how synthetic audio-visual media for disinformation and hate speech threaten core demo- cratic functions,” Digital society 1, 19 (2022).

[69] N. C. Köbis, B. Doležalová, and I. Soraperra, “Fooled twice: People cannot detect DFs but think they can,” Iscience 24 (2021).

[70] A. Eberl, J. Kühn, and T. Wolbring, “Using DFs for experiments in the social sciences-a pilot study,” Frontiers in Sociology 7, 907199 (2022).

[71] R. A. M. Reimao, “Synthetic speech detection using deep neural networks,” (2019).

[72] A. Godulla, C. P. Hoffmann, and D. Seibert, “Dealing with DFs–an interdisciplinary examination of the state of research and implications for communication studies,” SCM Studies in Communication and Media 10, 72–96 (2021).

[73] O. M. Hammad, I. Smaoui, A. Fakhfakh, and M. M. Hashim, “Recent advances in digital image masking techniques Future challenges and trends: a review”, SHIFRA, vol. 2024, pp. 67–73, May 2024, doi: 10.70470/SHIFRA/2024/008.

[74] N. Diakopoulos and D. Johnson, “Anticipating and addressing the ethical implications of DFs in the context of elections,” New Media & Society 23, 2072–2098 (2021).

[75] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” arXiv preprint arXiv:2006.04558 (2020).

[76] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, “A deep learning framework for audio DF detection,”

Arabian Journal for Science and Engineering , 1–12 (2021).

[77] M. Lataifeh, A. Elnagar, I. Shahin, and A. B. Nassif, "Arabic audio clips: Identification and discrimination of authentic cantillations from imitations," Neurocomputing 418, 162–177 (2020).

[78] Z. Almutairi and H. Elgibreen, "A review of modern audio DF detection methods: challenges and future directions," Algorithms 15, 155 (2022).

[79] Y. Zhou and S.-N. Lim, "Joint audio-visual DF detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision (2021) pp. 14800–14809.

[80] J. Frank and L. Schönherr, "Wavefake: A data set to facilitate audio DF detection," arXiv preprint arXiv:2111.02813 (2021).

[81] H. Khalid, M. Kim, S. Tariq, and S. S. Woo, "Evaluation of an audio-video multimodal DF dataset using unimodal and multimodal detectors," in Proceedings of the 1st workshop on synthetic multimedia-audiovisual DF generation and detection (2021) pp. 7–15.

[82] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," arXiv preprint arXiv:1907.00501 (2019).

[83] H. Yu, Z.-H. Tan, Z. Ma, R. Martin, and J. Guo, "Spoofing detection in automatic speaker verification systems using dnn classifiers and dynamic acoustic features," IEEE transactions on neural networks and learning systems 29, 4633–4644 (2017).

[84] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual DF detection method using affective cues," in Proceedings of the 28th ACM international conference on multimedia (2020) pp. 2823–2832.

[85] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "Fakeavceleb: A novel audio-video multimodal DF dataset," arXiv preprint arXiv:2108.05080 (2021).

[86] D. M. Ballesteros, Y. Rodriguez-Ortega, D. Renza, and G. Arce, "Deep4snet: deep learning for fake speech classification," Expert Systems with Applications 184, 115465 (2021).

[87] P. Kawa, M. Plata, and P. Syga, "Defense against adversarial attacks on audio DF detection," arXiv preprint arXiv:2212.14597 (2022).

[88] H.-s. Shin, J. Heo, J.-h. Kim, C.-y. Lim, W. Kim, and H.-J. Yu, "Hm-conformer: A conformer-based audio DF detection system with hierarchical pooling and multi-level classification token aggregation methods," arXiv preprint arXiv:2309.08208 (2023).

[89] Z. Cai, W. Wang, and M. Li, "Waveform boundary detection for partially spoofed audio," in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, 2023) pp. 1–5.

[90] M. Mcuba, A. Singh, R. A. Ikuesan, and H. Venter, "The effect of deep learning methods on DF audio detection for digital investigation," Procedia Computer Science 219, 211–219 (2023).

[91] S.-Y. Lim, D.-K. Chae, and S.-C. Lee, "Detecting DF voice using explainable deep learning techniques," Applied Sciences 12, 3926 (2022).

[92] L. Yan, S. Yin-He, Y. Qian, S. Zhi-Yu, W. Chun-Zi, and L. Zi-Yun, "Method of reaching consensus on probability of food safety based on the integration of finite credible data on block chain," IEEE access 9, 123764–123776 (2021).

[93] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "DFs generation and detection: State-of-the-art, open chal- lenges, countermeasures, and way forward," Applied intelligence 53, 3974–4026 (2023).

[94] H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang, and L. Nie, "Voice-face homogeneity tells DF," ACM Transactions on Multimedia Computing, Communications and Applications 20, 1–22 (2023).

[95] V. M. M. G. M. G. Yash Doke, Prajwalita Dongare, "Deep fake video detection using deep learning," Journal homepage 3, 540–544, (2022).

[96] S. Ansari, "Practical example: Face recognition," in Building Computer Vision Applications Using Artificial Neural Networks: With Examples in OpenCV and TensorFlow with Python (Springer, 2023) pp. 401–428.

[97] K. K. Babu and S. R. Dubey, "Csgan: Cyclic-synthesized generative adversarial networks for image-to-image transformation," Expert Systems with Applications 169, 114431 (2021).

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

[99] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021) pp. 772–781.

[100] L. Hussain, "Fortifying AI Against Cyber Threats Advancing Resilient Systems to Combat Adversarial Attacks", EDRAAK, vol. 2024, pp. 26–31, Mar. 2024, doi: 10.70470/EDRAAK/2024/004.